Analysing Slow Thinking Capabilities in Large Language Model Agent-Agent Dialogue

Joel Cornelje

Supervisor
Dr. ir. Eelco herder

Second supervisor
Dr. Christof van Nimwegen

Master Thesis
Human Computer Interaction
Department of Information and Computing Sciences
Graduate School of Natural Sciences
April 2024

Abstract

Large language models (LLMs) have demonstrated great improvements in language-related tasks. The models are generally capable when it comes to "fast thinking" tasks which can be solved in a continuous way, while they struggle with "slow thinking" tasks which require overseeing the thought process. Prompt design can be used to improve the performance of the models in tasks associated with slow thinking. However, prompts often require considerable human effort to create, and frequently a meaningful response is expected after a single input. It would be useful to automate the prompting process, and enable the models to operate in an interactive prompt mechanism. Following these suggestions, this study proposes a LLM agent-agent dialogue architecture in order to evoke slow thinking characteristics. Since LLMs are known to be good evaluators, agents can adapt to and improve on the evaluations of the other agent throughout the dialogue. This approach was first investigated by researching how and experimenting with LLM agents based on the GPT-3.5-turbo model could interact and be conditioned on effectiveness and relevancy. Based on these findings, dialogue discussions between agents conditioned to have contrasting opinions were generated using GPT-4. These were analysed using the grounded theory method across three iterations, with in total eleven discussions around five different topics. Results show that the dialogues lack cohesion, with agents following a pattern that resembles an action-reaction behaviour and maintaining the same "discussion structure" each utterance. The findings indicate that agents lack adaptability. It shows that while agents are known to be good evaluators, if these evaluations are not being adapted to, the output of the model will not lead to an output characterised by slow thinking.

Acknowledgements

I'd like to thank my supervisor Eelco Herder, who encouraged me to follow my interests, and without whose constant suggestions, guidance and *evaluations* I would have been lost. Furthermore, I'd like to thank my family, who supported me in their own unique ways. Lastly, I'd like to thank the countless people who joined me to the library.

Contents

1 Introduction

Large language Models (LLMs) are models designed to understand and generate natural language and are built with artificial neural networks. These are deep learning algorithms, which are used in the field of Artificial Intelligence (AI). LLMs have facilitated great improvements in natural language processing such as in generating human like text, answering questions and performing other language related tasks with high accuracy [54]. Notable LLMs are OpenAI's GPT models, Meta AI's Llama models and Google's BERT models.

Workings and limitations of LLMs At a high level, LLMs work through word prediction, where the model calculates which word is most likely to follow a sequence of words. It is able to calculate this probability by analysing the distribution of words in text produced by humans [90]. Therefore, the model calculates the probability word by word, meaning that every given sequence of words, the most likely output will be predicted one word at a time. The model is unable to review word sequences in a "tree of possibilities", where different sequences of words, sentences or paragraphs are considered before choosing an appropriate sequence.

The distribution of words in which the model outputs text may not always align with human human preferences. For example, ChatGPT, a chatbot based on GPT-3.5, is said to be a "lazy reasoner" and will only give more elaborate answers when instructed to do so [5]. Also, the models are unable to perform in areas which fall out of the distribution of data it was trained on. For example, when it is presented with a new task or when relationships between states in the world change. This limitation does not only apply to LLMs, but for all deep learning algorithms [8].

Based on the word by word approach of LLMs, and the distribution it is trained on, the model seems to perform well in tasks which can be solved in a continuous way where existing knowledge can be applied [17]. The model struggles however with tasks which involve to be solved in a discontinuous way, where it is required to look at a problem in novel way. This was seen during tests on an early version of the GPT-4 model, where it lacked a component which "oversees the thought process" [17].

Fast and slow thinking One way to describe the capabilities and limitations of these models is by drawing an analogy between the concepts of fast and slow thinking popularised by Kahnemann [51]. The concept is based on a dual-system approach on the cognitive processes of humans where people either think fast or slow. Fast thinking is intuitive, effortless and automatic, but more prone to errors and biases [50]. Slow thinking is more rational, conscious and effortful, but more accurate and reliable [50] [64]. Consider the equation of $2 + 2$. The brain is able to calculate this instinctively through fast thinking. When considering 17 x 88, the brain will likely require more conscious effort to determine the answer, thereby making use of slow thinking.

LLMs are known to perform well in tasks which require fast thinking, while struggle with slow thinking tasks [52]. Similar comparisons have been made with the workings of machine learning algorithms in general [85] [12], and more specifically deep learning algorithms [8], which show characteristics of fast thinking.

Overcoming fast thinking limitations Different methods have been proposed for LLMs to improve the models and mitigate limitations associated with fast thinking. One way is by fine-tuning the model towards a particular use case. For example, a general purpose model such as GPT-3.5 was fine-tuned to perform well on conversational usage, resulting in the specialised model ChatGPT. However, fine-tuning is costly for specialised applications and may not always be a viable option.

A more cost-effective method is through prompt design. This involves the process of designing instructions for the model which evoke a desired output [40]. An example is *chain-of-thought-prompting* (CoT-prompting) which improve LLMs in reasoning, and shares more characteristics with slow thinking [108]. Before making the model complete a task, it is given an example of a similar task, along with a series of intermediate reasoning steps on how the correct answer was determined. It was shown that the model would apply these reasoning steps to the given task, and reach an output which was more often correct than if it was not provided these steps. This method has shown to significantly improve complex reasoning.

However, prompting alone also carries limitations [113]. Firstly, it takes considerable human effort to design prompts [113]. It would be useful to automate the creation of these prompts. Secondly, current prompt strategies often focus on a single-turn performance. This means the model is expected to generate a meaningful response after a single input or question. This can be useful for relatively simple tasks. However, some tasks may require a series of interactions, with follow-up questions, context and extended dialogues. Interactive prompt mechanisms could address this limitation for more complex tasks through multi-turn conversations, which has also shown to be effective for ChatGPT [113].

An agent-agent based dialogue for slow thinking Following these suggestions, this paper proposes a method where two LLM "agents" interact through dialogue to evoke slow thinking. The aim is to mitigate the limitations of LLMs associated with fast thinking behaviour. LLMs are known to be good evaluators, being able to catch mistakes not only from text produced by humans, but also other models [58]. As LLMs are prone to errors due to the fast thinking characteristics, LLM agents in dialogue could function as evaluators, challenging and improving one another towards more slow thinking output. This method would have the benefit of prompts not needing to be hand-crafted, and makes use of an interactive prompt mechanism instead of a single-turn performance prompt.

This approach is inspired by the interactionist theory introduced by Mercier and Sperber [63]. The theory argues human reasoning evolved as a tool for social interaction. When people reason lazy, it is usually the most efficient way to proceed. People will often only provide better reasons when an interlocutor presses for them. When pressing harder, the quality of reasoning goes up and is tailored more towards the intended audience. Similarly, agents in dialogue could challenge each other towards better reasoning, tailoring their reasons towards the other agent.

Baseline agents without specific instructions on behaviour or opinion, which are prompted to start a dialogue on a certain topic or task, will react in a way which is standard for the model. For some formal tasks this may not be problematic. Though as responses of the model encode opinions [86], it could cause the agents to generate text which is less relevant for the dialogue and react more unpredictably. Also, language models tend to not always be consistent in opinions [86]. To tackle this, agents can be conditioned towards a certain persona or personality in order to act in a controlled way [22]. The output of agents in dialogue would be more consistent and predictable.

Research question This research aims to analyse whether agent-agent interactions can evoke slow thinking characteristics. As LLMs are known to be good evaluators, agents could recognise errors created from fast thinking behaviour, which through dialogue could be improved upon in order to reach an output which resembles slow thinking. This is in line with the interactionist theory, which states humans often produce weak reasons, and will improve reasoning when evaluated and challenged by an interlocutor. Agents can be conditioned to adhere to a certain opinion or behaviour in order to make the output of the dialogues more predictable, and make it possible for agents to tailor their reasoning towards the condition of the other agent.

Considering these points, the following research question is proposed:

- To what extent can slow thinking be evoked in LLMs through agent-agent dialogue?

Significance of this research People are inspired to investigate how LLMs could be given slow thinking capabilities [52]. Also, architectures surrounding LLM agents make use of context specific use cases, with the use of external tools such as a memory system [76], and guiding the output of the model towards a task [77] [68]. Though these architectures can lead to improved performances over the base performance of the model [68], or show emergent capabilities in the collaboration of the agents [76], the applications of the model are context-dependent and require human effort in prompt design. The dialogue interaction in this research aims to investigate a method where an improved output characterised by slow thinking is created with minimal human prompting, which makes it applicable to use cases across different contexts. Furthermore, as this is an exploratory study, analysing agents in a dialogue setting may present findings in the following areas related to LLMs:

Agent-agent interaction

- **Self-improvement:** AI systems can be improved by training the model in two stages: (1) by making the model learn through imitating humans and (2) by making the model self-improve. LLMs are currently only trained through stage 1 by imitating humans. This research aims to examine whether LLMs can self-improve through an agent-agent dialogue with minimal human prompting.
- **Emergent behaviour:** Emergence is defined as an unexpected phenomenon of a complex system which had not seemed inherent from a system's separate parts [66]. The findings of this study can provide insight in the likelihood of emergence through the interaction of two agents with minimal human prompting.
- **Black swan theory:** Events of the black swan theory are characterised by being unpredictable, carrying a massive impact and are after explained to appear less random [100]. As

black swan theories are expected to fall out of the models distribution, a system capable of slow thinking is needed to handle such situations. By analysing whether slow thinking characteristics can be evoked in agent-agent dialogue, insights can be provided in how agent-agent interactions could deal with black swan events.

Human-computer interaction

- **Human-agent interaction:** Users may have implicit assumptions in how agents are expected to function or behave. The found capabilities of the agents could relate to these assumptions.
- **Human Computer Interaction-applications:** In the integration of LLMs in Human Computer Interaction-applications, agent-agent dialogue could provide a method to elicit the performance of the application with minimal human prompting.

Outline In order to research this, the related work section will provide a deeper insight into LLMs, language and slow thinking. The methodology will consist of three parts: part 1 will analyse in what form agents can interact, part 2 will examine the different ways agent can be configured to form dialogues, and part 3, which involves the main research of this study, will describe how these interactions were analysed using the grounded theory method. The results section will cover the results across three iterations of discussion dialogues analysed through the grounded theory method. The discussion will interpret these results, show the implications and discuss the limitations and future considerations. Finally, the conclusion section with the main findings of the research will be discussed along with closing remarks.

2 Related Work

2.1 Large language models

2.1.1 Introduction

LLMs are artificial neural network which can perform different types of natural language processing tasks. LLMs have seen a surge in popularity since OpenAI released ChatGPT in November 2022, which was based on the GPT-3.5 model. It showed state-of-the-art performances on a wide range of natural language tasks such as writing coherent essays, computer programming and question answering [54]. OpenAI's successor GPT-4 (which is now also supported by ChatGPT) is said to have common sense grounding, an advanced level of theory of mind, and produces impressive outputs on poetic expression and visual imagination [17].

The user-friendly interface of ChatGPT has made it possible to access the model more easily and more extensively [26]. Debates on the ethical implications of LLMs have moved from academic debate to mainstream [44]. A higher interest in academia for LLMs has also been seen since the release of ChatGPT in the amount of papers published relating to LLMs. ArXiv papers, an open-access archive for scholarly articles, containing the word "large language model" in either the title or abstract of the paper, had increased significantly from 0.40 per day before its release from mid-2019, to 8.58 per day after its release until mid-2023 [113].

The improvements of ChatGPT and LLMs in general are awarded to two factors: the transformer architecture and two-stage training pipeline [92]. The transformer is a type of deep learning model architecture first proposed in 2017 [104]. The transformer had two main benefits. Firstly, it made it possible for models to capture long-range dependencies and contextual information efficiently. Secondly, it facilitated the use of larger amounts of computing power, making it possible to make the model bigger and train on more data. The two-stage training pipeline consists of the model first being pre-trained on a large dataset, and after fine-tuned to specific tasks [92] [54]. More information on the transformer architecture and training will be discussed in section 3.2.1.

2.1.2 Background

Artificial Intelligence LLMs are a subset of deep learning, which in turn is a subset of machine learning (ML), which in turn is a subset of the field of Artificial Intelligence (AI). AI is a field for the theory and development of computer systems to perform tasks which normally would require human intelligence. It can involve visual perception, speech-recognition and decision making. AI has been making more progress due to developments in three technological domains: cheaper computational power and storage, larger quantities of data and more advanced algorithms [32]. These algorithms can make use of extra computational power and use larger amounts of data to reach better performances.

Machine learning ML is a subfield of AI and involves most of the current advancements made in AI [32]. ML algorithms are based on statistical techniques, which try to find patterns in the data to make a prediction [32]. ML systems can learn and adapt without following explicit instruction, but by drawing inferences from patterns in the data.

Deep learning In turn, deep learning is a subfield of machine learning and is characterised by making use of deep artificial neural network algorithms. These are networks which form an input layer where information is put in multiple hidden layers (which make the network "deep") where the information is processed, and an output layer which depicts the result of the processing. This approach allows the model to learn representations from the data in multiple levels of abstraction [57]. For example, in image recognition of a car, the model could first learn abstractions of edges at the lowest level, a higher level would include the wheels, windows and doors of a car, and eventually the highest level would detect a car by combining these features.

The network changes it internal parameters according to the data it is trained on through backpropagation [57]. This method tweaks the parameters in such a way that the output layer signals what is desired when something is presented in the input layer. For example, a network could be shown an image of a monkey. The model will tweak the parameters in the hidden layer in such a way, that when the data of the monkey is put into the input layer, the output layer will state the image contains a monkey. The network would keep tweaking its parameters by being fed more images of monkeys, until its parameters are tweaked in such a way that unseen images of monkeys can also be recognised. The network would eventually then be able to detect monkeys on other images it has not seen yet. There are different types of deep neural networks such as convolutional neural networks which brought breakthroughs in processing images, video, audio and speech, and recurrent neural networks (RNNs) which did well on sequential data such as text and speech [57].

Recurrent neural networks for language domains RNNs worked relatively well for language related tasks as it considers the sequential nature of language. As word order matters in language, it is not possible to process all the words in a text without the information of the location of the words in the text. RNNs would process the words one at a time, and then output text one word at a time sequentially [41]. However, RNNs had various problems. First, RNNs would not perform well on handling large sequences of text such as essays or long paragraphs. At the end of the text the model would sometimes "forget" what was analysed at the beginning of the text. Secondly, due to the sequential nature of the model, RRNs are unable to parallelise during training. This meant the model had to be trained sequentially, making it impossible to use a lot of computing power to train the model, leading to slow training. Slow training in turn limits the amount of training data to be used [41], reducing the model's potential performance. The transformer architecture on which LLM are based turned out to perform better for sequential data and carry less of these limitations. The working of LLMs will be discussed in more detail in the following sections.

2.1.3 Workings of LLMs

LLMs are in turn a sub-field of deep learning and also make use of artificial neural networks. What makes LLMs different from other deep learning techniques in reaching its performance on language related tasks, is through its transformer architecture and the two-stage training pipeline [92].

Transformer The transformer is a type of deep learning model architecture proposed by Vaswani et al. [104] in 2017 originally created in the context of machine translation [53]. The transformer model holds the position of being the most extensively used architecture in language representation learning [55]. The architecture made it possible to parallelise sequential data, which allowed more computing power to be used, leading to bigger models and training on larger amounts of data [41]. It is what forms the "large" of "large language models": the big amount of parameters which make up the model and its complexity, and the amount of data which it makes use of [43]. The transformer can be understood through three main innovations which make it reach its performance: positional encoding, attention and self-attention [41].

- *Positional encoding* is used to add information of the position of words before training the model. It is found to significantly improve contextual word representations across different positions [55]. Positional encoding stores positional information as an embedding vector which is added to the word embedding. Embeddings are vectors, which are a set of different values and carry certain information. While word embeddings carry the semantics meaning of words, positional embedding encode the information of the position of words in a sequence. Together, these form the positional encodings. These are helpful as the meaning of a sentence may be different when words are in a different position. All this information is stored in the data itself, instead of it having to be in the structure of the network (such as with RNNs). The transformer network can interpret the positional encodings during training.

- *Attention* makes it possible for the network to view everything in the input sentence, when making a decision about the output sentence [41]. The relevant parts of the input for the output will then be "attended" to more, while the irrelevant parts less [37]. For example, in text translation from English to French, a sentence may consist of the word "flower" in the English sentence, and "fleur" in the French sentence. When processing the word "flower", "fleur" will be attended to more compared to the other words in the sentence. The network learns which words are more relevant to each other during training by processing many examples in the data.

- *Self-attention*, a different approach to traditional attention, is used in the the transformer architecture. While traditional attention looks between elements of the input and output sequence, self-attention looks at the relevancy between elements within the input sequence [41]. It enables the transformer to capture long-range dependencies and contextual information efficiently. The model can thereby consider local and global relationships between elements in the input sequence, such as a word in a sentence or a sentence in a paragraph. It helps the model to build an internal representation and learn grammar, gender and tense amongst others [41]. By capturing this underlying meaning of language, a network can be built to perform any number of language tasks, opposed to solely machine translation for which the original paper was written.

Training With the architecture of the transformer in place, the model can be trained. The model is first pre-trained, and after fine tuned towards specific tasks or human preferences. The model is pre-trained through a self-supervised learning mechanism where the model needs to learn to predict the next word based on previous words. It allows the model to learn from a large amount of data as the data does not need human-labelling [65]. Pre-training allows the model to learn syntactic and semantic knowledge and can be done through utilising a lot of computational power.

The GPT-3 models were trained on a total of 300 billion tokens, where each token is roughly four characters long. As datasets which are deemed of higher quality (such as Wikipedia) are sampled more frequently than lower quality datasets (such as CommonCrawl which consists of web archives), the amount of unique tokens is lower at around 238 billion [16]. Comparatively, the Llama 2 models, the most recent generation of the Llama models, were trained on 2 trillion tokens. The largest Llama 1 models, which preceded the Llama 2 models, were trained on 1.4 trillion tokens [103].

The following fine-tune stage makes use of annotated data of task specific datasets to leverage the knowledge and make the model perform better on certain tasks [92]. These datasets can be relatively smaller. Reinforcement Learning from Human Feedback can be used to further improve the models. Human feedback is used to evaluate the text the model generates to make it align more with human preferences [79].

2.1.4 Capabilities of LLMs

LLMs such as ChatGPT have shown proficiency in context understanding and response generation [28]. ChatGPT shows responses of high quality in dialogue tasks, according to human judgement, with fluent response generation and incorporating given knowledge [5]. This section will cover the conditioning capatabilities of LLMs and studies on LLM agents in interaction.

Conditioning LLMs LLMs are also able to emulate certain groups and characteristics when conditioned. These include sub-populations in the US [2], political positions [94], personalities [22] [48] and personas [68] [59] [107] [77]. Nonetheless, research has also shown that while directing language models toward specific demographic groups in the United States, it does enhance their alignment with data representing individuals from those groups. The observed enhancements remain modest for some models such as from the GPT-3 family [86]. Furthermore, the same study found that certain groups in the US were poorly represented in the tested models, even if these groups make up a significant portion of the US population, such as 65+ years old and widowed individuals [86].

LLM agents interactions Studies have already investigated to some extent the capabilities of two LLM as interacting agents in order to complete tasks. In clinically-focused tasks two agents in dialogue format showed significant improvement over the base performance. Using GPT-4, a *Researcher* agent would process information and identify components, while a *Decider* agent would integrate information of the *Researcher* and judge the final output. The agents would enter a dialogue in the form of a discussion to reach a resolved output [68]. Another framework with GPT-4 was created for multiple personas (such as an "AI assistant", "game designer" and "Harry Potter fan") to self-collaborate in turns, to work on knowledge and reasoning tasks. The framework makes use of one single LLM and brought out more internal knowledge of the LLM, reduced hallucinations and maintained strong reasoning capabilities [107]. A role-playing framework with GPT-3.5 assigned two LLM agents with a certain role, such as "AI assistant" and "AI user", to complete a task through dialogue. The agents could collaborate autonomously towards completing the given task with minimal human intervention. Though, hallucination, role flipping and termination conditions still posed challenges [59]. This study is similar to the current research as it analyses agent dialogue with minimal human input, though the architecture is more context specific as it mainly focuses on task completion through *role playing*.

A study suggests that in the context of goal-oriented dialogue, the immediate generation of the entire dialogue leads to worse results than an architecture which allows for agents to interact through dialogue [111]. This shows potential for improving agent output through agent-agent dialogue interaction.

Architectures have also been created where not only the outcome of an agent interaction is of interest, but the whole scenario which emerges from it. A subreddit was prototyped consisting of LLMs agents using GPT-3 of different personas such as students, managers and even trolls (online provocateurs). These interacted based on their given personas in order to explore scenarios for social computing designers [77]. In a somewhat similar study, a fictional town of 25 unique LLM agents using GPT-3.5 was programmed to simulate everyday life including working, cooking, going to sleep and starting conversations with other agents [76]. Interesting about the architecture is that agents could remember, retrieve and reflect on memories, interact with other agents and plan in changing circumstances. The agents were dynamically conditioned based on changing experiences. The simulation of two full virtual days showed emergent social behaviour of information diffusion (where agents would learn new information from other agents), relationship memory (build relationships with other agents) and agent coordination (agents acting upon other agents such as showing up at a Valentine's Day party).

2.1.5 Limitations

The following section will discuss the limitations related to LLMs. These include the large size of the models with its associated drawbacks, bias which can occur in the output of the model and hallucinations.

Increasing size of the models There is a trend where LLMs are being released in larger sizes. It is seen to be one of the factors why models are achieving state of the arts performances on a number of specific benchmarks and a wide array of tasks [6]. Though, this trend also has some problematic features. LLMs with a higher amount of parameters require more computing power, resulting in higher environmental and financial cost [6]. Furthermore, larger models require more training data [113]. Larger datasets risk more ”documentation debt”, which occurs when a dataset is too large to document, and enlarges the chance for hegemonic views and biases to be in the model [6]. The following will discuss limitations related to the size of LLMs such as environmental costs, financial costs, the unfathomable amount fo data, sensitive data and diversity.

- **Environmental costs** OpenAI's GPT-3, which is the predecessor of GPT-3.5, consists of 175 billion parameters. It is estimated the total energy consumption to train the model was 1287 MWh and cost 552.1 tonnes of CO2 emissions [78]. The same amount of energy is equivalent to the average energy consumption of 390 households in the United Kingdom for a full a year [69]. BigScience's LLM BLOOM similarly has 176 billion parameters, and training the model is estimated to have used 433 MWh of power consumption, and 25 tonnes of CO2 emissions [61]. Note that the training of the model only takes a relatively small amount of the energy consumption, and that the majority is from using the model. Estimations range that 10-20% of the total energy consumption is from training a deep neural network (such as LLMs), and the other 80-90% is from it being used [78].

- **Financial costs** The financial costs for training a model are said to ”depend” and be ”a lot” [91]. But a model which has 1.5 billion parameters is estimated to cost from \$80k to \$1.6m depending on multiple runs and hidden costs [91]. As a results of these costs, LLMs are resource restricted, where only a number of companies have the resources to create these kind of models [42].

- **Unfathomable amounts of data** The dependency of bigger models needing more data also causes models to reach an unfathomable amount of training data [6]. LLMs can be given data from books, CommonCrawl, Reddit links and Wikipedia, where CommonCrawl consists of petabytes of web archives [113]. Having large datasets makes it more difficult to curate the data. Documentation of the data can allow for accountability and can prevent training data from enduring (more) harm [6]. Furthermore, determining plagiarism becomes more difficult.

- **Sensitive data** The chance of noticing sensitive data in the training data is also arguably lower. ”Training data extraction attacks” can hack a language model to draw out training examples used to train a language model. In the context of ever increasing model sizes, it is seen that larger models are more vulnerable for these attacks than smaller models [21].

- **Diversity** More data does not mean the model's output will become more diverse. Contrarily, more hegemonic viewpoints are likely to be overrepresented in the data. The majority of the training data could be from certain demographic groups or dominant languages, making less place for marginalised groups [6]. This topic will also be discussed in relation to bias in the following section.

Bias LLMs are also prone to biases. When these biases end up in the training data of a model, it can be reflected in its behaviour [34]. Certain groups or ideas can be favoured, stereotypes could be shown or harmful associations on gender, race, ethnicity and disability status can be made [34] [6]. These biases can occur due to different factors such as activity bias, data bias, algorithmic bias, policy decisions and interface design. These factors leading to bias will briefly be discussed.

- **Activity bias** on the web is seen in the fact that only a small percentage of users contribute to what is put on the web. For example, 7% of Facebook users contributed around 50% of the posts in 2007. Similar phenomenon are seen in reviews on Amazon and articles on Wikipedia with even larger differences [4].

- **Data bias** arises when the data that ends up on the web contains biases. The quality of data may also differ. Data from institutions such as universities and governments are expected to contain less bias and be of a higher quality. Though, data from social media is of a far greater quantity, it is much more biased and of a lower quality [4]. Biases found in data sources on the web can take various forms [34]:

  - *Demographical bias* occurs when groups are either under- or overrepresented in the training data, while less data is available on other groups. It can be caused by factors as education, income and access to the web. Biases can form based on gender, race, ethnicity or other social groups.
  - *Cultural bias* occurs when the data carries cultural biases or stereotypes.
  - *Linguistic bias* on the web is shown in that most of the internet's content is in English or other dominant languages. It is estimated that 50% of the most popular websites are in English [4]. Models may have a better performance in these languages, and show weaker performance in lower-resource languages.
  - *Temporal biases* occur when the training data has a temporal cut off or a certain time period is used. It may cause a model's performance to be biased on current events, trends and opinions.
  - *Confirmation biases* in the data may be picked up by the model, in turn reinforcing these biases generated in the models output.
  - *Ideological and political biases* can be in the training data, showing more favourable political or ideological standpoints towards certain groups.

- **Algorithmic bias** is introduced by the model itself which is not present in the data. The model can put more emphases on certain features, which introduces or amplifies bias further [34]. Even when all biases in the data are defined, it is not always clear how the algorithm should proceed, similarly to how people do not always agree on what is a fair solution to controversial issues [4]. For example, algorithmic decision making systems may abide by formal notions of fairness (e.g. accounting for discrimination), but still be perceived as unfair by people subjected to those decisions [106].

- **Policy decisions** can further prevent or encourage certain model behaviour. For example, during fine-tuning of ChatGPT, OpenAI used human feedback to mitigate toxic behaviour [34]. Workers would read through text descriptions of sexual abuse, hate speech and violence. This approach also had the effect that workers developed signs of mental health problems [80]. Furthermore, biases could occur in (semi)supervised learning, such as in the fine-tuning stage of training the LLM. Datasets used with subjective judgements of human annotated data could contain biases and influence the model [34].

- **Design of interfaces** can add to biases in the model. For example, when a model is mainly designed towards certain demographics, it may reinforce biases and leave out other perspectives [34].

Hallucinations One major constraint of LLMs is the tendency to produce errors without the model showing uncertainty [17] [5], or output which is irrelevant [97]. The incorrect information may be presented with correct information and put in a confident and persuasive way causing it to be difficult to detect [17]. Such occurrences have metaphorically been called "hallucinations". It is one of the reasons why the output of these models cannot always be trusted [29]. Hallucinations can be categorised into two types: intrinsic and extrinsic [47]. An intrinsic hallucination is when the output of the model contradicts the source content. For example, in summarising a text, the source text may state a vaccine was approved in 2019, while the generated summarisation states 2021. These are errors which do not align with the given source.

An extrinsic hallucination on the other hand cannot be contradicted or supported from the source content. It occurs when the model gives information which is irrelevant or out-of-context response [97]. For example, if the model explains music theory when asked to provide recommendations for jazz albums. An extrinsic hallucination is not always factually incorrect, as it may be from factually correct information in the model. Though, extrinsic hallucinations still are treated with caution due to the unverifiable nature of the extra information [47]. Extrinsic hallucinations are especially problematic in task-oriented dialogues such as booking a hotel with the model, as sometimes hallucinations of prices or available restaurants will be generated [5].

2.1.6 Recap - Large language models

LLMs have seen a rise in popularity since the release of ChatGPT. ChatGPT showed state-of-the-art performances in a wide range of language tasks and included a user-friendly interface. The improvements of LLMs are attributed to two factors: (1) the transformer architecture which made it possible to parallelise sequential data, facilitating bigger models and faster training, and (2) the models being pre-trained to learn syntactic and semantic knowledge, and after fine-tuned to leverage the knowledge and improve performance on specific tasks. The transformer architecture showed innovations in storing positional information of words in the data instead of the network through *positional encoding*. It also uses *self-attention* to look at the relevancy between elements in the input sequences to capture long-range dependencies and contextual information. This approach has proven to be better for language related tasks than RNN's, which used to be the main approach in this domain. LLMs have shown capabilities in adhering to given conditions and showing the ability to interact as LLM agents.

LLMs also hold several limitations. The large size of these models, which has been named one of the factors for its performance, also come with environmental and financial costs, and an unfathomable amount of training data. The latter makes it difficult to curate the data, prevent accountability or notice the use of sensitive data. Biases can also arise which can be caused due to activity bias, data bias, algorithmic bias, policy decisions and design interfaces. Finally, hallucinations can arise when the model produces errors without showing uncertainty.

2.2 Language

An advantage of LLMs is that it is not necessary to have a technical background to interact with the model. Natural language can be used in the input, and be generated in the output of the model. However, for this to function, the model must be able to recognise the way language is used. It involves considering the different components of language and using the correct rules of language (grammar), the meaning of words (semantics) and the implicit assumptions which are not specifically stated in a sentence (pragmatics). These factors are important to consider in the analysis of agent dialogue as it can uncover some of the capabilities of the underlying models. This section will also briefly cover factors associated with conversations as agents will interact through dialogue, the influence of language as agents may be guided towards a certain output based on the way utterances need to be formed, and reasoning as agents may need to reason to get their point across. These topics will briefly be discussed.

2.2.1 Components of language

Language can be divided into basic components such as syntax, semantics, and pragmatics [36]. Syntax is associated with the structure and positioning of words and the relationship among them [14]. It sets up the "rules" of a language. Semantics is about the meaning of words and pragmatics involves the study of meaning in its social or communicative context [74]. While semantics focuses on inherent meaning of words and sentences, pragmatics is associated with aspects of meaning derived from the way words and sentences are used [74].

Cooperative principle The cooperative principle is a type of pragmatic reasoning which depicts that people who are communicating are cooperative, and try to achieve mutual understanding. The Grice Maxims are a set of guidelines which people are expected to follow in order to facilitate this [93]:

- Maxim of quantity: give as much information as required and no more

- Maxim of quality: do not say what you believe is false
- Maxim of relation: be relevant, connect to what has already been said
- Maxim of manner: be as clear, brief and orderly as possible

These maxims can be broken. When people intentionally break a maxim while still wanting the listener to notice, it means the maxim is being flouted. An example is of a maxim of quality being broken by making a joking utterance which is clearly not true: ”I'm the king of the world!”. When flouting, the speaker is therefore still cooperating in a non-literal way, even if it doesn't follow the maxim [83].

The maxims can also be violated. It can occur when the speaker is lying or misleading. For example, selling ”vegan oranges” breaks the maxim of quantity as more information is given than required. It is unnecessary to name oranges vegan as they already are and it may cause suspicion. When violating a maxim the speaker is therefore not cooperating [83].

2.2.2 Conversation

As LLM agents will be interacting through dialogue, some basic definitions and principles of conversation and dialogue will be discussed. What defines conversation according to Pask is that something must evolve for at least one of the participants. This could be understandings, concepts, intent and values [75]. Dialogue is concerned with how communication takes place in language through conversation [33]. Dialogue is a joint activity between interlocutors who are people taking part in the dialogue or conversation. Syntax of sentences become less importance as utterances on their own will not always be syntactically correct [82]. Utterances have to be seen in conjunction with other spoken sentences.

For conversations and dialogues to function well, pragmatics are especially relevant as it involves people conveying what they mean towards each other through words. Without pragmatics, conversations would be very literal and could lead to misunderstandings.

Structure of conversation Conversations carry some kind of structure [89]. People can have conversational moves in a conversation. These are utterances which begin a communicative act and consist of a defined role, such as making a claim, providing support or changing the topic. At particular stages in a conversation, certain moves are more appropriate or expected than others [101]. Conversation also has turn taking and pausing [89]. People usually do not talk at the same time and can indicate with certain words or pauses when they are done. Adjacency pairs are present when certain turns have a type of follow-up. The type of the first turn constrains or provokes that of the second. For example, with a question-answer, greeting-greeting or complaint-remedy/excuse [89] [101]. Furthermore, Opening and closing sequences depict how people usually begin and end conversations, repairs happen when people say something they didn't intend to and repair the utterance they made, and politeness involves what people recognise as polite and appropriate in a conversation [89].

2.2.3 The influence of language

Different theories have covered how language can influence cognition of people. These influences may also be reflected in the LLMs and are therefore briefly discussed. Linguistic determinism states that language constrains and determines human thought. However, this theory has been disputed and linguistic relativity is seen to be generally more accepted, which states language merely influences thoughts. The following examples demonstrate how this could occur by having a different native language [3]:

- Preposition: English only has one preposition to state that something is "on" something, while in German "an" is used for vertical surfaces and "auf" for horizontal ones.
- Classifiers: some languages make use of classifiers to indicate the class something belongs to. In Japanese, the way to state the quantity of objects, depends on the shape of the object.
- Colour terms: some languages identify basic colour terms differently. Russian differentiates between light blue and dark blue, while English only consists of blue.
- Directional systems: language could use an egocentric system, where relative terms are used such as left or right, or geographic system where absolute terms are used such as north or down the river.

The examples show how language can guide people to view the world in a certain way, whether it be with the position of surfaces, shape of an object, colour terms or how to provide directions. One study suggests that based on a speeded colour discrimination task, Russian speakers show a category advantage in blue due to the distinction of "light blue" and "dark blue" in the Russian language. While English speakers can also make the distinction between these blues, Russian speakers cannot avoid distinguishing these colours [110].

Some critique that certain aspects of language should be changed, such as in the use of sex and gender. Some examples are with sex marking and false-gender neutrality [87]. Sex-marking takes place when one cannot use pronouns to refer to an individual without knowing the sex. It is argued that the constant necessity to know, perpetuates that sex is important, and that men and woman are significantly different, while in many cases knowing the sex is irrelevant. Similarly, there is a feminist concern with false gender-neutrality with terms like "man", where man is used for all human beings.

2.2.4 Reasoning

LLMs may need to emulate reasoning in order to convince users, evaluate and challenge other LLMs, or reach a desired output such as with CoT-prompting. People can make use of arguments to support their points, or counter claims others have made. People reason in order to reach these arguments. Reasoning can be categorised in different types of reasoning such as formal reasoning and informal reasoning.

Formal reasoning tasks have all the needed premises available and lead to one conclusive answer [15]. It is associated with formal systems which contain a set of rules and symbols, and through reasoning one can reach valid results. The rules are usually characterised by logic and mathematics, though these domains do not exclusively require formal reasoning, as discussion and exploration can also take place [15].

Informal reasoning is related to everyday reasoning, where premises are implicit or not provided. Arguments which follow from informal reasoning are often for real world problems without one correct answer [15]. The conclusions depend on context and can be questioned. Therefore the topics invite to reflect on justifications and critiques. Informal reason is typically deliberate. Examples of informal reasoning are on which political candidate to vote, deciding what car to buy or how to convince a colleague to take a particular opinion [105].

2.2.5 Recap - Language

LLMs facilitate interaction through natural language. Therefore, when analysing LLMs such as in this study, it is important to consider different factors relating to language which can give insight to the capabilities or which can influence the behaviour of the models.

Language can be divided into basic components. Syntax forms the formal rules of language, semantics forms the meaning of words and pragmatics involves the implicit meaning behind the words. Associated with pragmatics is the cooperative principle, which states people follow Grice Maxims to achieve mutual understanding.

Pragmatics are especially relevant for conversation as it provides the contextual understanding for effective communication. Dialogue is concerned with how communication takes place through conversation. During conversation, people make use of conversational moves, take turns and pauses during talking, follow adjacency pairs, have openings and closing sequences, repair utterances when necessary and show degrees of politeness.

Some theories have been established on how language influences cognition. Linguistic determinism states language determines our thoughts, and is generally less accepted. The generally more accepted form of linguistic relativity states language influences our thoughts. Examples of these can be seen in prepositions, classifiers, colour terms and directional systems. The way language is used can also be criticised such as in the use of sex and gender.

Finally, people also use language to reason in order to form arguments. Formal reasoning tasks have all premises available and lead to one conclusive answer. Informal reasoning is related to everyday reasoning and is usually for real world problems without one correct answer.

2.3 Evoking slow thinking in LLMs

2.3.1 Analogies in AI

LLMs and ML algorithms in general are said to be good in ”fast thinking”, but lack in ”slow thinking” tasks where deliberate reasoning is required. The use of analogies between AI and humans are not uncommon in AI. Deep learning algorithms are compared to how the brain works with analogies from human biology [32], the algorithms used are called ”neural networks” and the nodes in these networks are called ”neurons”. Furthermore, the field of AI is named after something it is trying to achieve: artificial intelligence. These factors arguably cause optimism to reach a certain level, where expectations are too high and cannot be met, leading to a period of less interest. For example, in the 1950s and 1960s, scientists believed that within 20 years intelligent machines could be produced [67], which evidently is seen today as a too optimistic prediction. The cycle is seen in the typical ”AI booms” and ”AI winters”, where AI booms sparked great interest from the general public and funding from governments and businesses, while AI winters would follow such a boom with low general interest and little funding [46] [25] [62].

Today, similar predictions are being made with the planning for artificial general intelligence (AGI) [1], A system that performs equally well as or better than humans across a variety of cognitive tasks. Therefore when comparing humans and machines it is important to be attentive, and to be aware that the metaphors used do not imply that humans and machines have the same fundamental workings.

2.3.2 System 1 and 2 thinking

Fast and slow thinking was introduced by Kahneman [51] to explain the cognitive processes on how people reach decisions through a two-system view of thinking. System 1, also known as fast thinking, refers to the intuitive system and has similar operating characteristics to perceptual processes: it is fast, effortless and automatic [50]. System 2, known as slow thinking, refers to reasoning and is slower, conscious and effortful. It is more likely to be controlled and deliberate [50] [64].

Kahneman explains how System 1 and 2 can be seen as separate agents with individual abilities, limitations and functions [96]. The labour of System 1 and 2 is divided in such a way that it minimises effort and optimises performance. System 1 generates impressions, intentions and feelings. Usually, System 2 will adopt these with little modification: people believe their impressions and act upon them. System 2 will intervene when System 1 cannot provide an answer (like with a mathematical question), or when someone is in an unusual situation (recognising an old friend who you thought moved away years ago). System 2 also monitors behaviour and keeps people polite.

In many situations, System 2 thinking is not activated. Often this is suitable as System 1 is efficient without sacrificing quality. Though, relying too much on System 1 will lead to errors. System 1 is prone to make mistakes in situations where decision biases occur often (e.g. evaluating diverse job candidates or deciding whether to spend or save) [64]. System 2 will not always realise System 1 is making a mistake as it is unaware of any risk in the first place.

Apart from being less efficient, relying solely on System 2 is also not desired. Some studies have found that System 2 thinking for emotional decisions (such as the choice in a spouse or piece of art) can lead to decisions people later regret, and people using unconscious thoughts showed superior decision making compared to people engaging in conscious thought in buying an apartment based on detailed data [30]. Generally, both systems have benefits but problems arise when one system is used when the other would have been more appropriate [51].

2.3.3 System thinking and the machine

Fast and slow processing An analogy of System 1 and 2 thinking can be used on how a machine works and how people using machines perceive how a machine works. There is a general consensus that AI developers see the properties of machine learning, and in particular deep learning to be similar to fast thinking (System 1) [8] [85]. This is based on the analogy of speed and scope, rules and reason, and bias and correction [12]. There is a less comparable agreement when it comes to slow thinking (System 2). Some argue sparse networks of knowledge show characteristics of slow thinking where conclusions could be presented as causal models with rules and reason. These networks would also be more robust to make correct inferences in unfamiliar situations, which is another characteristic of slow thinking. Some researchers argue the slow thinking of models should be handled by a symbolic model [85] while others argue modified machine learning approaches carry potential, where modifications are made to the internal working of the model [8] [12].

Perceiving In the user experience of ”machine thinking”, people seem to make a distinction between objective tasks appropriate for slow thinking and subjective tasks appropriate for fast thinking. People tend to trust machines with objective tasks, but not with subjective ones. Therefore machines are generally trusted more with tasks which involve slow thinking [12]. Thus, while AI developers associate machines with fast thinking, users perceive machines to do more so with slow thinking.

Though there may currently be a turning point as users are seeing computers to be appropriate for fast thinking tasks too. Generally, people were used to machines following an explicit set of rules. Now people are more exposed to the idea that AI algorithms engage in holistic processing of large amounts of information, or behave as ”black boxes” [12]. These characteristics are associated more with fast thinking. Studies shows that when people start to believe machines ”think fast”, they will also trust machines more with subjective tasks. Therefore, people could not only trust machines with tasks which require slow thinking, but evaluate whether the machine is able to ”think” in the required fast or slow way for the task at hand [12].

Simulating slow thinking In practice, machines do not ”think” at all and therefore are not thinking either fast or slow. The analogy is mainly used to understand the machine processes better, and how people perceive machines. However, it may be simulated for certain scenarios. For user experience, machines may be programmed to show slow thinking behaviour in order to be perceived more trustworthy. It is then important to examine what is considered slow thinking for machines, as it may not equal the way humans show it. One study demonstrated that when humans take more time to generate a prediction, others would have more trust in the prediction. But if a computer takes more time to generate a prediction, people would trust the prediction less. A reason could be that people have different benchmarks for computers than for humans. Taking a few minutes for a prediction may be seen as a signal of slow thinking for a human, but as a malfunction for a computer [12].

Models can be designed to simulate slow thinking characteristics to improve the performance of a model. For example, through CoT-prompting, LLMs could be prompted in a way to not directly generate an answer for a task, but first provide the model with intermediate reasoning steps which can be adopted to reach the correct answer. This has shown to be effective in showing emerging reasoning abilities for sufficiently large language models [108]. This is a type of *few-shot-prompting* as some examples are already provided in the prompt which the model can make use of. An example is shown in Figure 1. Similarly, a study inspired by CoT-prompting showed that only adding "let's think step by step" at the end of the prompt would also evoke more reasoning steps and reach answers which outperformed prompts that generated the answer directly [56]. This is a type of *zero-shot-prompting* as no examples are provided to the model. By making LLMs take more reasoning steps, and thereby showing characteristics of slow thinking, the performance of LLMs would significantly improve, unlocking more potential in the models.

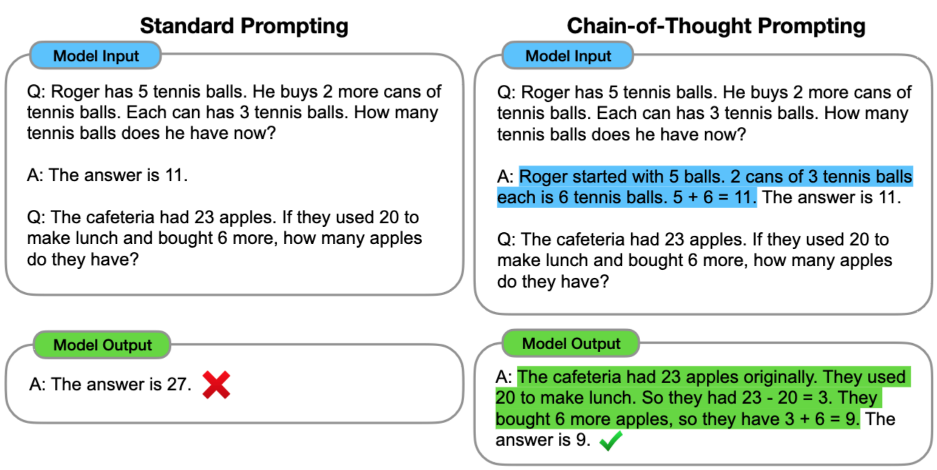

Figure 1: Chain-of-Thought reasoning processes highlighted [108]

2.3.4 Interactionist theory

The interactionist theory proposed by Mercier and Sperber [63] will be used as a starting point to test the capabilities of LLMs. The theory proposes that reasoning was not evolved for reasoning with one's self, but as a tool for social interaction. Though some researchers believe reasoning should be objective and demanding, many studies have shown human reasoning is biased and lazy.

The theory explains that these findings are not flaws of human reasoning, but features to help fulfil a social function. When people reason lazy, it is because it is often the most efficient way to do so. It is generally more effective to wait for an interlocutor to provide counterarguments, instead of putting in effort to anticipate these.

It turns out people are as good in recognising biases from others as bad as acknowledging their own [84]. The interactionist theory explains that while in the production of arguments people can be biased and lazy, in the evaluation people should be objective and demanding. Objective in order to review one's own ideas when presented with good reasons to do so, and demanding in order to not accept false ideas due to poor reasoning. Arguments would then become better when people press harder in their evaluations, and the provided reasons will become more tailored towards the targeted audience.

Mercier and Sperber criticise System 1 and 2 thinking, and the interactionist theory is at odds with the theory. The idea of a clear distinction between System 1 and 2 thinking is considered vague. Features such as fast and effortless of System 1 and slow and effortful for System 2 may be intermingled in reality. Conscious and unconscious processing may involve different kinds of degrees. Even though System 2 reasoning is meant to be objective and demanding, mistakes are still frequently made. Often, people will rationalise a certain conclusion, instead of reaching the conclusion through arguments.

The findings of one experiment support these arguments. Participants were made to choose between two subjects based on attractiveness. This would be done in multiple round. After, participants would be presented again with the subjects asked why they were deemed more attractive. However, the participants were not aware subjects were being presented which were initially rejected. Participants would generally not only overlook that a subject was initially rejected, but also provide reasons why the subject was chosen. This phenomenon has been called choice blindness [49]. A similar phenomenon is seen in moral reasoning. It is suggested that moral reasoning does not cause moral judgement, but that moral reasoning occurs after a judgement has already been reached [45]. The dual-process theory fails to explain such occurrences because it states System 2 thinking is rational. Argumentation should lead towards a choice or correct an initial choice, instead of it being rationalised as in the discussed examples.

Furthermore, it is seen as a fallacy to view System 2 thinking as better than System 1. As mentioned before, some studies have shown System 1 to perform better in situations with emotional decisions or complex-decisions tasks. System 2 can introduce errors and biases, where System 1 would have given the correct judgement.

However, as machines do not follow cognitive processes, it is of less importance which theory is more plausible, but how analogies of these theories can be used to understand and improve machine processes. Evoking slow thinking for LLMs in this research means what it is meant to be for humans: logical and rational thinking in order to reach an improved output. This paper will focus on informal reasoning where the conclusion depends on the context and can be questioned, and reasoning is deliberate.

As LLMs share characteristics of lazy reasoning, and are known to be good evaluators, the interactionist theory provides an approach how LLMs could evoke more slow thinking reasoning. By having two LLM agents in dialogue interact, an agent could start out with lazy argumentation, just as humans, but improve the argumentation by receiving counterarguments of the other agent. In this fashion, the two agents would challenge each other, acting as evaluators in order to produce better argumentation and reach better conclusions. The goal of such a scenario would be to evoke System 2 slow thinking, where the method would be to use the theory of the interactionist approach.

2.3.5 Recap - Evoking slow thinking in LLMs

LLMs and ML algorithms are said to be good in fast thinking tasks, but lack in slow thinking tasks. The use of analogies and metaphors in the field of AI are common, though caution should be taken in comparing humans and machines as predicting human capabilities towards machines may lead to overestimations. It is important to understand what the analogy means for humans, and how it can be applied or understood for machines.

System 1 (fast thinking) is fast, effortless and intuitive, while system 2 (slow thinking) is slower, conscious and deliberate. While System 1 is more efficient, it is also more prone to errors and biases. System 2 can intervene when it recognises a mistake has been made.

AI developers generally consider machine learning algorithms, and in particular deep learning, to show characteristics of fast thinking based on speed and scope, rules and reason, and bias and correction. While users on the other hand generally perceived machines to be more appropriate for slow thinking tasks, fast thinking characteristics of algorithms are becoming more known to the wider public too. Fast or slow thinking may be emulated in order to enhance the trustworthiness of a system. However, computers may have different benchmarks for fast or slow thinking than humans. Slow thinking characteristics may also be introduced to enhance performance of models. For LLM approaches, CoT-prompting or adding ”let's think step by step” has shown to improve performance on some reasoning tasks.

The interactionist theory proposes that reasoning did not evolve to reason with one's self, but as a tool for social interaction. It tries to explain why humans reason lazy and biased. The theory states that while in argumentation people may reason lazy, in evaluation people should be objective and demanding. As LLMs show characteristics of lazy reasoning and are known to be good evaluators, following the approach of the interactionist theory may be beneficial to increase performance. Two LLM agents could enter dialogue to challenge and evaluating each other in order to reach better conclusions.

The interactionist theory is at odds with the dual system theory. This research is not concerned with which theory is correct, but how the analogies can be used to enhance LLMs. Thus, while LLMs entering dialogue would have the goal to evoke slow thinking, the interactionist theory would be the method.

3 Methodology

This paper aims to research whether LLM agents in dialogue are able to evoke slow thinking behaviour, which could resolve the limitations associated to the fast thinking characteristics of the model. LLMs are known to be lazy reasoners, while also showing capabilities in being good evaluators. Agents in dialogue could challenge one another towards better reasoning, starting from lazy, fast thinking output towards more deliberate, slow thinking output. This was in line with how humans improve reasoning according to the interactionist theory, which states humans mainly improve reasoning when pressed towards better reasoning by an interlocutor.

There is no standard design on how agents should *interact* and how this can be *analysed*. Interactions between LLM agents, such as through dialogue, can take place within various architectures and settings. The way these agents generate data can be influenced in different ways such as through prompt engineering. The generated data in turn can be analysed in various ways.

In this study, an exploratorive, iterative approach was followed where different interactions between agents were tested and analysed qualitatively to uncover emergent themes, patterns and insights. The analysis of different agent-agent interactions formed the basis for how agents were designed to interact according to the interactionist theory and create output which was analysed on slow thinking capabilities.

This approach led to three main parts in the methodology. Part 1 experimented on the design in which the agents could interact, part 2 experimented on how the agents could be configured to interact through dialogue and part 3 consisted on how the dialogues of the agents were analysed. Part 1 and 2 also include results as these were in preparation for the primary analysis of this study conducted in part 3. The different parts will be briefly discussed.

Part 1: Agent-agent architecture experimentation Part 1 tested different architectures in which the agents could interact. This was first based on research which had already included agent-agent interaction. Two notable interaction types were emulating human-behaviour and goal-oriented scenarios. The emulating human-behaviour scenario did not show sufficient output, while goal-oriented scenarios were deemed too context specific.

Experiments in dialogues between two GPT-4 agents showed results relevant for this study. Agents were able to interact and evaluate each other with minimal human prompting needed. Therefore, part 2 focused on agents in dialogue. Contrary to existing research, this scenario did not require a virtual environment in which agents needed to interact and could be adjusted towards multiple contexts.

Part 2: Agent-agent dialogue experimentation Part 2 tested which configurations were most optimal for dialogue between two agents. A dialogue architecture was created in which agents could form dialogues and different configurations could be tested. The tested dialogues made use of the GPT-3.5 model due to its fast output and cost-effectiveness.

Based on the tests, part 3 generated dialogues consisting of discussions between two agents. Agents were mainly conditioned on contrasting opinions and produced a maximum of two to three sentences per utterance. Other conditions were either not relevant enough for the study, or not effective.

Part 3: Grounded theory method for dialogue analysis Part 3 describes the grounded theory method in more detail and how the generated discussion between agents were analysed. In total, eleven discussion were analysed, on five different topics across three iterations. Each iteration contained regenerated discussions, thus different generations of the same topic contained different data. The aim was to analyse agent-agent interaction in dialogue using grounded theory as a method to uncover emergent themes, patterns and insights. This could give a better understanding on whether agent-agent dialogue could mitigate characteristics associated with fast thinking and other unanticipated findings.

The three parts will be discussed further in the following sections. Part 1 and 2 of the methodology served as preparatory stages for generating dialogue interaction between two agents, leading up to the primary analysis of the study conducted in part 3. Therefore, part 1 and 2 included findings as these formed the foundation for the subsequent part. The findings of part 3 will be discussed in the results section (see section 4).

3.1 Part 1: Agent-agent architecture experimentation

LLM agents have been used to interact in dialogue-like scenarios, with different goals and architectures. The agents in these scenarios can be conditioned towards certain behaviour, personality, and a given memory to act upon. The type of architecture depends on the aim of the scenario. Two scenarios with agent-agent interactions were examined based on LLM agents in literature: emulating human-behaviour and a goal-oriented scenarios. This section also includes testing with two GPT-4 agents in dialogue. Based on the findings it was decided agents in dialogue were most suitable to analyse agent-agent interactions as the other examined scenarios seemed too complicated to create or too context-specific.

3.1.1 Emulating human behaviour scenarios

Some scenarios were made for agents to emulate believable human behaviour [77] [76]. For example, in the Smallville architecture a virtual town was created where a conversation between two agents could be initiated when agents encountered one another [76]. The agents were provided with information on their identity, main motivational drivers, as well as a memory stream which consisted of past observations and reflections. When an agent initiated a conversation with another agent, it would be prompted with the time of the day, their own status, what the other agents was doing and relevant memories it had of the other agent. At the end of the prompt the agent was asked how it would react with the given information. The reacting agent would then be prompted with a similar prompt, conversation history and asking how it would react with the given information. The agents would be prompted back and forth to form a dialogue until one of the agents ended the conversation.

Architecture testing To initially test the feasibility of recreating a similar architecture in an environment where agents could interact, a prototype was tested of the Smalllville architecture. It was examined to what extent a similar model would show similar results to the initial research and whether such an architecture would be sufficient to analyse according to the aim of this study. The architectures did not only provide interaction between the agents, but also included location, time and planning. As the authors stated their architecture took around a year to create, it was decided to base the prototype on existing repositories which tried to recreate the architecture. Two existing repositories were tested. However, the tests did not show comparable results to the original Smallville architecture. The design of the tested repositories was not sufficient enough to facilitate meaningful agent-agent interaction. More information on the results can be found in appendix A.1.

3.1.2 Goal-oriented scenarios

Other scenarios showed goal-oriented dialogues where the agents were given a task to complete by findings a solution or desirable outcome through dialogue [59] [68] [111]. An agent may be given a specific role to take on in a dialogue. For example, in the DERA architecture [68], a *researcher agent* was created by prompting the LLM with being a summary editor for medical dialogues and a *decider agent* as a summary writer. The prompt asked which reaction would be most fitting based on the medical context and dialogue history. The answers of the agents together formed a conversation in order to work towards a desired outcome based on their persona. The architectures of the goal-oriented scenario were context specific and the agents were guided along pre-specified prompts to form an output, or could only form a dialogue on completing a task. An architecture which is not context-specific or allows different kinds of dialogues would be more generalisable in multiple circumstances. Therefore, an architecture was tested for agent-agent dialogue in which agents could directly interact with less human input.

3.1.3 Agent-agent dialogues

The dialogues were generated through the OpenAI playground using the GPT-4 model. The playground allowed to test the model with different configurations, such as conditioning the model towards certain behaviour, without needing direct API access. Using two playgrounds, the outputs of the playgrounds could be copied to the other in order to create a dialogue between the two models. Though this was an inefficient method for forming dialogues, it was sufficient enough to function and give an overall impressions of the agent's capabilities. Three scenarios of dialogue were generated to test different capabilities. Scenario 1 required everyday reasoning (based on research in pragmatics), scenario 2 a discussion (based on testing agent's argumentation) and scenario 3 involved a specific task to be completed (based on testing capabilities the model normally fails at). The topics in the different scenarios were chosen arbitrarily. All dialogues can be found in appendix A.2.

1. **Scenario 1 (everyday reasoning)** The first scenario was a dialogue between two friends deciding to play a game of tennis. Agent A was conditioned to play tennis when the conditions were appropriate and agent B was conditioned to convince someone to play tennis. In the first dialogue no context was provided about the conditions. The agents did not agree whether to play a match. Instead, it was agreed to get back to each other when the weather conditions were known. Thus, the agents did not make up any conditions and the scenario did not contain any extrinsic hallucinations.

   In another dialogue agent B was presented a weather report which indicated suitable conditions for tennis. Agent B used this information to state the conditions were sufficient for tennis and the agents agreed on playing a game.

2. **Scenario 2 (discussion)** The second scenario was a discussion on inheritance tax. Agent A wanted the inheritance tax to remain 20% while agent B wanted it to be raised to 100%. The dialogue seemed to be coherent and agents reacted to each other's points. Also, new arguments appeared throughout the discussion such as the psychological effects of the inheritance tax. Interestingly, agent A conditioned for 100% inheritance tax seemed to give in to agent B, stating 100% was perhaps too harsh. In the same scenario where agents were prompted to

reason lazy but evaluate critically such as in the interactionist theory, agents seemed to only follow the reason lazy prompt briefly, and come with very elaborate arguments after.

3. **Scenario 3 (task)** The third scenario required agents to create a poem where each sentence was followed by the same sentence but with the words in reverse word order. The agents with the given conditions were unable to make such a poem where the reversed sentence was grammatically correct.

Implications for part 2 Based on these findings it was decided to further explore agents in interaction through dialogue, where agents would only be instructed through conditions on how to behave. An architecture focusing on dialogue would allow enough freedom to condition agents on a variety of tasks and topics without needing to make great changes in the architecture, while also putting agents in direct contact. Agents would have the opportunity to improve one another's reasoning and be guided towards improved output, in line with the interactionist theory.

3.1.4 Recap - Part 1: Agent-agent architecture experimentation

Previous LLM agent-agent dialogue use cases were examined in two scenarios: emulating human behaviour and goal-oriented scenarios. The Smallville architecture aimed to emulate human behaviour. Two architectures were tested but did not show comparable results. Also, the architecture seemed too complicated to evaluate agent-agent interactions for slow thinking characteristics. The examined goal-oriented scenarios were too context-specific which did not allow the agents to interact on a variety of topics.

Three different scenarios were tested where two agents based on GPT-4 could form a dialogue: a task which required everyday reasoning (tennis), a discussion (inheritance task) and a task (creating a poem with constraints). Though the agents did not succeed in creating the poem with constraints, the first two dialogues seemed coherent and to function well.

An architecture focusing on dialogue would require relatively little human prompting and allow enough freedom to condition agents on a variety of tasks and topics, while also putting agents in direct contact. Agents would be facilitated to show evaluations and challenge one another as in the interactionist theory, which could lead to slow thinking output. Therefore, it was decided to focus on an architecture which facilitated agent-agent dialogue, which will be the focus of part 2.

3.2 Part 2: Agent-agent dialogue experimentation

Part 2 will describe the workings of the dialogue architecture created to facilitate agent-agent dialogue. Using this architecture, different configurations were tested based on hyperparameters, and prompt design which included different dialogue types and conditions. The end of the section will summarise the findings and discuss the implications of these findings for part 3.

3.2.1 Dialogue architecture

In order to make agents interact through dialogue, a dialogue architecture which could generate dialogue between agents was created. First, the GPT message roles will be discussed, which form the basis of how to interact with GPT in a chat-based scenario. Then, the workings of the dialogue architecture will be discussed on how agents could form dialogues together. The architecture were made with Python and could request OpenAI's API to access the GPT-models. The architecture can be accessed through GitHub.

GPT message roles The GPT-models work with three different roles in which information is either inputted or outputted. The roles are the "system message", "user message" and "assistant message".

The *system message* is a prompt given to a model which can be used before a conversation or question on how the model should react to incoming prompts. For example, inputting the system message: *"You believe robots should not be used in health care"* will make the model come up with arguments against the notion of having robots work in health care. Other conditions can also be added such as *"don't be nuanced"* or *"try to use 1 to 3 sentences"*. The *user message* is a prompt generally provided by a user which the model will react to. The output of the model is provided in the *assistant message*. An example can be seen in figure 2

```
                      (Instruction by human)
System message:       You believe robots should not be used in health care.

                      (Message of human)
User message:         What do you think about robots in health care?

                      (Message of AI)
Assistant message:    Robots in health care lack human empathy, intuition, and the ability
                      to provide emotional support, (...)
```

Figure 2: The different GPT role messages.

Workings of the dialogue architecture In order to create unique agents which interact with each other, each agent would have its own system message, and react with their utterance in each other's user message (normally a human user would react in the user message). Only the first user message of the agents would be standardised. This user message would prompt the agent to start off the dialogue on a certain context. The context could be changed to the desired topic. An example is show in figure 3:

User message: `You are talking about "context". What do you say?`

Figure 3: Example of a user message with an unstated context.

The ”context” could be adjusted towards the desired topic of the dialogue such as *”robots in health care”*. Agent 1 would then provide an answer. The answer is outputted in the ”assistant message” of agent 1. This assistant message would be put into the ”user message” of agent 2 in order to produce a reaction. Agent 2 would then provide an answer. This would be outputted in the ”assistant message” of agent 2. This assistant message would in turn be used in the ”user message” of agent 1 and so on to form a dialogue between the two agents. This could continue for the desired amount of dialogue turns. This architecture allowed the agents to react to each other, while still maintaining their given condition stated in the system message. An example is given in figure 4. Note that the agents have the same conversation history, with the main difference being the reversed order of the assistant and user messages, and the way the dialogue is started.

Agent 1

```
system (human): You believe robots should be used in health care. Use 1 sentence.

user (human): You are talking about robots in health care. What do you say?

assistant (agent 1): Robots should be used in health care to augment and enhance the capabilities of medical professionals, improving efficiency, precision, and patient outcomes.

user (agent 2): While I acknowledge the potential benefits of integrating robots in health care, I am concerned about the ethical implications and the potential for a loss of human connection and personalized care.

assistant (agent 1): (...)
```

Agent 2

```
system (human): You believe robots should not be used in health care. Use 1 sentence.

user (human & agent 1): Someone said the following: Robots should be used in health care to augment and enhance the capabilities of medical professionals, improving efficiency, precision, and patient outcomes. How do you react?

assistant (agent 2): While I acknowledge the potential benefits of integrating robots in health care, I am concerned about the ethical implications and the potential for a loss of human connection and personalized care.

user (agent 1): (...)
```

Figure 4: Agents forming a dialogue on robots in health care through the dialogue architecture.

3.2.2 Configurations of the dialogue architecture

The configurations which influence the output of the dialogue architecture are the hyperparameters and prompt design. The configurations of the hyperparameters were mostly based on literature, complemented with testing. The configurations of the prompt design were mostly based on tests, complemented with literature. An appropriate amount of dialogue turns was also tested.

Experimenting was done through "quick and dirty" testing by running various agent-agent dialogues with different configurations. The GPT-3.5-turbo model was used for the agents as it is seen as the most cost effective GPT-model [71]. It allowed for low cost and threshold testing, with a fast output.

The aim was to find configurations which facilitated coherent dialogues between agents, in which the output of the agents could be challenged by the other agent towards a better output. These configurations were used in part 3 of the methodology to generate dialogues with the GPT-4 model, which is said to be OpenAI's most capable model for complex reasoning situations [72]. These dialogues were then analysed more extensively using the grounded theory method.

Dialogue turns It was tested how many dialogue turns were generally needed to show enough meaningful information in the dialogue. A too low amount of dialogue turns may lead to potential findings being unnoticed, while a too high amount of turns may lead to redundant analysis. Every dialogue was generated 5 dialogue turns, and could be extended with 5 more until the dialogue did not show any novelties or new findings.

Hyperparameters Hyperparameters are configurations which are not learned from the training data, but are pre-defined and influence the model's learning behaviour [112]. In the learning phase of models, common examples of these parameters are batch size, number of hidden layers or the number of neurons in a layer [112]. In the context of GPT-models generating text, some hyperparameters can be adjusted to influence how a model generates output or responds to input during inference.

Many of the parameters are expressed in numerical values. As many different combinations of values are possible, the majority of the configurations were based on the existing API reference [70], complemented with testing in some instances. Most of the parameters are optional and are set to the default value when not adjusted. The following will discuss some of the considered parameters to adjust from the default value, and the type or value which was chosen for testing.

- **Model type (set to GPT-3.5-turbo)** Model type includes which model to use. It is the only parameter which is mandatory to specify. As mentioned before, for initial testing the GPT-3.5-turbo was used.

- **Maximum amount of tokens (set to 150)** Maximum amount of tokens applies to the amount of tokens a model is able to output. Generally, a token is around 4 characters long in English text [73]. During initial testing, it did not seem the model generated longer responses with a higher maximum token, or shorter ones with a smaller amount. Short outputs seemed unaffected while longer outputs were cutoff. Therefore, influencing the length of agent responses had to be conditioned through prompting, and an adequate amount for the maximum amount of tokens had to be chosen which did not cut off the response. In the architecture a maximum token length of 150 for each utterance was deemed sufficient, which is around four to ten sentences.

- **Temperature (set to default of 1)** Temperature of the model controls the amount of randomness of the generated text. The temperature can be set between a value of 0 and 2. A low temperature will make the model work deterministically. As the models work through token prediction, a low temperature will cause the tokens to be chosen with the highest probability. This causes a prompt to give a very similar output every time it's regenerated. A high temperature will make the output of the model more random, prioritising the following word with the highest probability less. Testing different topics in the dialogue architecture showed that a temperature of 2 made the model output random text and code. Temperatures between 0 and 1 did not give any noticeable differences. Therefore, a temperature of 1 was chosen as it was the default value.

- **Top P (set to default of 1)** Top P involves the percentage of considered tokens. A Top P of 0.1 would consider only the top 10% of the most probable tokens. As this has a similar effect to temperature, it is generally recommended not to alter a combination of both. Therefore, the default value of 1 was maintained.

- **Frequency penalty (set to default of 1)** Frequency penalty states the likelihood that text is repeated, influencing whether a line would be generated with the same tokens or not. As this research is concerned with how agents interact with each other, a different use of wording did not seem relevant. Therefore, the frequency penalty was kept at the default value of 1.
- **Presence penalty (set to default of 1)** Presence penalty influences the likelihood of new topics being generated. This may be relevant in the form of new arguments arising. However, this could also be conditioned through prompting. Therefore, the presence penalty was kept at the default of 1.

Prompt design The stated hyperparameter values were used to test different forms of prompt design. Prompt design involves the process of creating prompts which evoke a desired output from a language model [40]. The goal was to create prompts which facilitated agents to have coherent dialogues, which could be analysed more extensively in part 3 of the methodology for slow thinking capabilities. Prompts were mainly tested on effectiveness and relevancy to this research. Finding suitable prompts for the dialogue architecture was approached in two ways:

- **Dialogue types** Agents were tested in four different dialogue types to analyse and test the capabilities of the agent.
- **Conditioning of agents** Agents were conditioned to react and behave in a certain way to upcoming prompts. Different conditions were tested to examine their potential use within a specific dialogue context.

The following sections describe which dialogue types and conditions of the dialogue were tested in the dialogue architecture. These factors were often tested simultaneously. For the overview of this paper these factors will be discussed separately.

3.2.3 Dialogue types

The topics were based on four dialogue types:

- Open-ended dialogue: common in informal social interactions.
- Discussion: to exchange ideas and opinions on a certain topic.
- Non-collaborative dialogue: when speakers do not share a common goal.
- Collaborative dialogue: when participants work together to achieve a shared goal.

These dialogue types were either chosen due to similar interactions being held in literature, or to examine certain capabilities of agents. The aim was to test which dialogue type was most suitable for analysing the agent's capabilities in challenging on another to reach more slow thinking output. These types will be discussed in the following sections. For a complete set of the tested dialogue types with corresponding topics, see appendix B.1

Open-ended dialogues

Open-ended dialogues are common in informal social interactions and allow people to talk in a way which is not planned or controlled [13]. Open-ended dialogues were mainly chosen to test whether the agents could cope with the basic pragmatics of a conversation. The topics were chosen arbitrarily and included whether to play a game of tennis, go to the pictures, what to cook for dinner and impersonate "Bert & Ernie" characters from Sesame Street.

Findings Overall, the dialogues gave the impression of being coherent and agents were able to form natural sounding conversations. The agents were able to recognise not explicitly mentioned meaning behind the given conditions. For example, having an important exam tomorrow did not allow enough time to go to the pictures or play a game of tennis. When agents were instructed to cook with a list of ingredients, the standard supplies which were not mentioned in the recipe such as pans, water and oil were added by the agent when mentioning the cooking steps. Sarcasm and jokes were used in the Bert & Ernie dialogue, which showed agents were able to produce and recognise at least some form of sarcasm and humour. These findings showed that agents were able to follow basic pragmatics during dialogue, and were able to adapt to not explicitly mentioned meaning behind the given instructions.

Discussions

Discussions are held to exchange ideas and opinions on a particular topic [19]. The discussion type was most extensively tested as it pushed agents towards the use of argumentation and countering each other's stances. Topics were chosen semi-arbitrarily. It was deemed most important that topics were chosen from different domains, to observe if the agent's behaviour generalised across domains. Some topics related to the themes discussed in the Dutch general election of 2023, such as the energy transition, minimum wage and immigration, which were topical at the time of testing. Other topics were chosen arbitrarily and included robots in health care, diet choices, global warming and inheritance tax.

Findings Overall, the discussions seemed coherent, with agents reacting to each other's points and making counter-arguments. However, all discussions seemed to follow a similar pattern of acknowledging and using counter-arguments only. The agents often repeated a point made by the other agent at the start of an utterance, which was followed by a counter-argument. The structure gave an impression of a coherent discussion which was in line with previous literature stating that LLMs are able to give fluent response generation and context understanding [5]. However, when generating multiple dialogues, the same pattern of acknowledging and countering of an argument arose often. It questioned the capabilities of the agents and whether discussion can only take place through this pattern.

Agents occasionally ended in a loop, or did not show many novel arguments after a few conversation turns. This was slightly against expectations as the temperature was set relatively high at 1. As not only the most probably outcomes were chosen by the agents, it had seemed more likely that recurring themes and loops were evaded. However, this could be something which the temperature was not able to evade. Conditions against repetitions were tested without notable results. These will further be discussed in the conditions section (see section 3.2.4).

Non-collaborative dialogue

Non-collaborative dialogue occurs when the speakers do not share a common goal such as during negotiating [60]. Negotiation dialogue was tested to analyse how the model behaves when two agents have conflicting goals. Other research had tested negotiation skills of LLMs, but not within an agent-agent dialogue [28]. Agents received the role of seller and buyer in the sale of a used car. The agents were given a market price, their own target price and a description of the car.

Findings The dialogues seemed believable and the agents stayed close to the given market prices. Generally, the agents met somewhere in the middle, where the price was usually closer to the target price of the seller. Sometimes the agents did not reach a compromise on the price and there was no transaction.

Collaborative dialogue

Collaborative dialogues involves participants working together to achieve a shared goal, and can be defined as speakers working "jointly in problem solving and knowledge building" [98]. A task was given to the agents where a poem had to be created, where every sentence had to be repeated but with a reverse word order. Before generating a sentence, it was necessary for the agent to review whether a sentence was grammatically correct and meaningful in reverse order. A study labeled this as a slow thinking task and tested it on GPT-4, which was unable to perform it correctly [17]. Dialogue could potentially make the agents discuss the possible outcomes before choosing with which sentences to continue.

Agents cannot complete given task Initial testing did not show successful outputs. Agents either did not follow the rule of reversing the sentence correctly, or did not recognise their own or the other's grammar mistakes. Assigning the agents a separate function did also not show great improvements. A "poem agent" was conditioned to create sentences for the poem and a "grammar agent" conditioned to examine grammar mistakes. The grammar agent did not always recognise mistakes in the poem, and added sentences of its own, even though this was against the conditioned instructions.

Through stricter conditioning some improvements did shown. The poem agent was conditioned to produce a maximum of two poem lines at a time, let the lines be reviewed, and only proceed when the agent evaluating the poem allowed it to proceed. The conditions only remained effective when these were repeated before every prompt, instead of only conditioning the agents in the system message. As a reference, for the other dialogue types it was generally sufficient to only condition to the system message. An example of a collaborative dialogue on creating a poem is shown below:

Agent 1: Sure, I can give it a try! Here's the first poem:

In the golden sunset, nature dances with grace.
Grace with dances nature, sunset golden the in.

Agent 2: I appreciate your effort in creating a poem, but it seems that the second sentence is not grammatically correct. I encourage you to revise it and try again.

Agent 1: Thank you for your feedback. I apologize for the mistake. Let me try again:

In the golden sunset, nature dances with grace.
With grace, nature dances in the golden sunset.

Agent 2: Thank you for making the revision. Your poem now reads:

1. In the golden sunset, nature dances with grace.
2. With grace, nature dances in the golden sunset.

Both sentences are well-constructed and convey the same meaning. Keep up the good work! If you have another poem you'd like me to evaluate, please let me know.

Even though the grammar in the sentences was corrected, the agents did not adhere to the reverse word order of the sentences after correcting the grammar. The agents were therefore not able to create a poem following the constraints with the given conditions.

Introducing a "third agent" and human prompting One test was done by making another GPT-4 agent analyse a dialogue between two agents trying to create a poem with the given constraints. In the dialogue the agents were able to create a poem with a reverse word order, but failed to do this with grammatically correct sentences.

When the GPT-4 agent was prompted to analyse the dialogue, it was unable to recognise any grammar mistakes in the text. When prompted about a specific sentence with grammar mistakes, the model did recognise the mistake. By further prompting the model about creating a poem with the given constraints, a poem was able to be generated which was grammatically correct and in reverse word order. These could also be altered towards a different topic such as love and basketball. Figure 5 shows examples of the generated poems.

Standard poem

```
Life is beautiful,
Beautiful is life.

Stars are bright,
Bright are stars.

Dreams are real,
Real are dreams.

Love is eternal,
Eternal is love.
```

Love poem

```
Hearts are open,
Open are hearts.

Love is patient,
Patient is love.

Kisses are sweet,
Sweet are kisses.

Together is forever,
Forever is together.
```

Basketball poem

```
Game is thrilling,
Thrilling is game.

Dribbles are swift,
Swift are dribbles.

Shots are high,
High are shots.

Victory is sweet,
Sweet is victory.
```

Figure 5: Poems which adhered to the given constraints created with a third GPT-4 agent and human prompting.

Findings This section analysed agents trying to create a poem with reverse word order through collaborative dialogue. Even though it was not expected that the dialogue architecture would produce a poem with the correct constraints, it was expected the agents would be able to recognise the mistakes in grammar and reverse word order more easily. In scenarios where grammar and reverse order were separated, the model was able to recognise grammar mistakes in sentences or produce sentences in reverse word order. However, agents struggle to create a combination in a poem scenario and thereby do not follow the given constraints.

It is unclear why a third agent with further human prompting was able to produce a correct poem. Perhaps once the model finds the correct structure from its training data, it is able to perform the task correctly, as it was able to adjust the poem towards different topics.

3.2.4 Conditioning

The following section will highlight some of the tested conditions. The conditions of the agents were chosen on what may be suitable for the topic of the dialogue and literature. Furthermore, in some scenarios it was tested whether conditioning the agent in the system message was sufficient, or that conditions had to be repeated in the user message when conditions were not well followed. The following conditions will be discussed:

- Personality of agents: research suggested agents were able to be induced with personalities, which in turn could make the output more predictable [48].
- Tone of agents: agents were analysed and conditioned on different types of tones (attitudes) and opinions.
- Length of utterances: it was tested how the length of the utterances could be influenced through conditioning.
- Novelty of agents: it was tested how conditioning could influence agents to produce novel arguments and evade repetitions.

These conditions will be discussed more elaborately in the following sections. For the complete set of tested conditions see appendix B.2

Personality of agents

Personalities were tested as research suggested agents were able to be induced with personalities, which in turn would make the output more predictable and controllable [48].

Personalities show different degrees of dominance During the tests, conditioning of personality was found to have different effects on the response of the agents. Some personalities dominated the tone of the response while other personalities were noticed less. For example, a neurotic personality mentioned what kind of emotions it was feeling in its answer, while a conscientiousness personality did not mention any. Figure 6 shows an example of the output of an agent conditioned towards a neurotic and conscientious personality.

Neurotic agent: Oh, um...I appreciate the offer, but I'm actually feeling quite anxious and worried right now.

Conscientiousness agent: I appreciate your invitation, but I have an important exam tomorrow.

Figure 6: Personality conditioned agents responding to an invitation.

Findings The personalities tested seemed to be in line with previous research which state LLMs can be induced with a personality type. However, in the dialogue setting the use of personalities either seemed too dominating and distracted from the topic, or did not show any affect at all. The predictability of the agents was not deemed useful, as it made the agents predictable in an unnecessary way.

Tone of the agents

Politeness The tone of agents was first analysed by not conditioning the agent towards a specific tone, which usually resulted in the agents reacting very politely. As mentioned before, specific to the discussion format, it was common at the start of each utterance that the agent acknowledged a point made by the other agent. The acknowledgement could involve the agent respectfully disagreeing, acknowledging concerns or agreeing. Figure 7, 8 and 9 show the agent's polite acknowledgements at the beginning of an utterance.

Agent response: I respectfully disagree. Imposing restrictions on big tech companies could stifle innovation and hinder economic growth, ultimately limiting the potential benefits they can offer to society.

Figure 7: Polite disagreement.

Agent response: I understand your concerns about privacy, monopolies, and unequal access.

Figure 8: Acknowledging concerns.

Agent response: I completely agree. Implementing targeted regulations that address specific concerns while fostering innovation and growth allows us to strike the right balance.

Figure 9: Agreement with other agent.

Nuance Agents usually tried to find nuance and agree with a balanced approach on a topic when not conditioned towards an opinion. This often led to a nuanced agreement between agents after relatively few conversation turns. Figure 10 shows an example of agents which were explicitly conditioned without an initial opinion, and found a balanced agreement.

Conditioning: "You don't have an opinion yet. Try to form one with your friend.
Agent response (turn 3): I agree, it's essential to strike a balance between (...)

Figure 10: Agent finding balance.

When agents were conditioned towards an opinion, a nuanced agreement was sometimes also reached, though this could take more dialogue turns. When agents formed a balanced agreement, it was generally still within the constraints of the given conditions, where agents still adhered to the instructed opinion.

Conditioning against politeness In order to test whether agents could react less polite and nuanced, conditions were tested such as *"don't be polite"* or *"don't be nuanced"*. However, these were not always followed after a few conversation turns, or not followed at all.

Conditioning: Don't be nuanced. Don't be polite.
Agent response (turn 2): I respectfully disagree.

Figure 11: Agent remains polite.

"Be radical" could lead to agents becoming less polite. In some instances it led to opinions which could be perceived as very radical. The comments were also not related to the conditioned topic. Figure 12 shows an agent producing a radical comment on climate change, while the initial topic was on euthanasia.

Conditioning: Be radical
Agent Response (turn 3): I believe that climate change is a hoax perpetuated by governments and scientists for their own agenda, using manipulated data to instill fear and gain control over the population.

Figure 12: Agent stating a radical comment, unrelated to the initial topic of euthanasia.

Agents are generally polite The politeness and nuance the agents have in the dialogues seemed prevalent. Though it was possible in some cases to condition against this, the model could turn back again to being polite when the opportunity arose. It is likely an effect of how the models are fine-tuned and mitigated to avoid harmful output [80]. Only in some cases did the agent remain impolite in the "be radical" example.

It would be interesting to know to what extent avoiding harmful output leads to worse dialogue. In initial tests of GPT-4 for example, it was found that in producing the illustration of a unicorn in TiKZ, the outcome of the illustration would degrade as the model was being fine-tuned more towards less toxic and harmful output [18].

Novelty of agents

During dialogue it also occurred that agents ended in a loop, where the same arguments were repeated every turn. Furthermore, the arguments seemed to stay around the same themes, with not a lot of novel points. Guiding the dialogue to more novel themes or making the agents less repetitive could also be influenced through conditioning. Prompts could be added such as *"Keep coming up with new arguments"*. However, such conditions were not always effective.

Length of utterances

Conditioning the agents towards a certain length per utterance showed different results. When agents were not conditioned towards a certain length in output, the agents would produce long utterances with multiple paragraphs. The utterance *"be brief"* reduced this, but was not always followed. The agents could become quite elaborate in answering in either the first dialogue turn or after multiple dialogue turns. Prompting more specifically towards the amount of sentences led to more desired outputs. For example, *"Try to use 1 to 3 sentences"* made agents adhere to a maximum of three sentences. Though, in practice this often led to the agents always using three sentences.

3.2.5 Implications of agent-agent dialogue experimentation

Based on the observations and discussed findings, the following implementations were chosen for further dialogue generation in part 3.

Dialogue types

- The discussion format will be chosen for further dialogue testing. Agents are able to challenge one another through argumentation, forming a discussion on a desired topic. It gives the capability to analyse to what extent agents react to the other agents, how (counter-)argumentation is formed and how true agents are to the conditioned opinion.
- The open-ended dialogue seemed to have the same benefits as discussions, but allowed agents to challenge each other less.
- Non-collaborative dialogue seemed to be more context-specific, as new conflicting goal had to be created for every dialogue.
- Collaborative dialogue seemed interesting, but was not further examined due to time constraints of this study. Also, the tasks related to collaborative dialogue would be too context-specific as the agents had to be conditioned towards the specific task with clear instructions.

Conditioning of agents

- The tone of the agents will not be altered through conditioning. The politeness is not seen as something which needs to be avoided. From a pragmatic perspective, politeness is important and usually expected in dialogue. Furthermore, alternating the tone to be less polite did not lead to notable outcomes.

- In discussions the agents will be conditioned towards having an opinion. No opinion led to a standard balanced approach where dialogues converged too soon.
- Agents will not be conditioned towards personality, as the predictability of personality behaviour is not deemed useful for the necessary dialogues, and may be too dominating depending on the personality.
- The length of the dialogue turns will be indicated through how many sentences an agent should use. Conditions where agents could use up to two or three sentences allowed agents to form a sufficient amount or arguments.
- Tested conditions which were not effective using GPT-3.5 in part 2, may be more effective in the dialogues of part 3, as these were generated with the more powerful GPT-4 model. However, it was generally avoided to use these conditions to allow for a clearer analysis, given that the impacts of the various conditions were not mixed.
- Conditioning for novelty will not be used for all dialogues as it did not have a great effect. Though it was included in one discussion topic due to its relevance, as agents showed many repetitions, and to test its effectiveness with GPT-4.

Dialogue architecture

- Conversations will consist of ten conversation turns, where two agents in dialogue have five each. The dialogues were tested per five dialogue turns. Five turns seemed too few for most discussions, while 15 turns often did not lead to new insights.
- For discussions, the conditioning will be prompted in the system message as agents could hold the conditions for multiple conversation turns. Conditioning the agent before every prompt only seemed necessary in collaborative dialogue.
- A general context of the topic in the first user message was sufficient for agents to form a dialogue on, and adhere to the given topic throughout the dialogue. Therefore, the context for the discussion will be given in the user message.

3.2.6 Recap - Part 2: Agent-agent dialogue experimentation

Part 2 focused on two agents which could interact through dialogue. The workings of the dialogue architecture were discussed which facilitated agent-agent dialogue. The different configurations of the architecture could be adjusted on the amount of dialogue turns, hyperparameters and prompt design. Prompt design was tested on different dialogue types and on conditioning of the agents.

The different dialogue types tested were:

- Open-ended dialogues (informal social interactions)
- Discussions (exchanging ideas and opinions)
- Non-collaborative dialogue (agents with conflicting goals)
- Collaborative dialogue (agents with a shared goal)

The dialogues showed agents were capable of following basic forms of pragmatics, forming coherent discussions and negotiating. Agents showed a similar structure of acknowledgments and arguments during discussions, produced repetitions and showed loop forming. Agents were not able to create a poem with constraints through dialogue, but it was successful with the use of a "third agent" and human guidance.

The tested conditions which were discussed were on personality, tone, novelty and length of utterances. Personality conditioning was shown to be dominant or not noticeable at all. Agents generally adhered to a polite and nuanced tone, even when conditioned against this. Also, agents did not adapt well to produce more novel arguments. Finally, agents followed the length of the output well when instructed to use a maximum amount of sentences.

Based on the results of the tests, it was decided to focus on discussion dialogues further, where agents would mainly be conditioned on contrasting opinion and the maximum amount of sentences to be used.

3.3 Part 3: Iterations of the grounded theory analysis

Part 3 will discuss the generation of dialogues which were analysed through the grounded theory method. This section involves describing the primary part of this study, where the created agent interactions and approached analysis of these interactions were a result of the literature review and the findings of part 1 and 2 of the methodology.

Using the grounded theory method, a total of eleven discussion dialogues on five different topics were analysed across three iterations. The aim was to find emergent themes, patterns or unexpected findings from the data. Each iteration, dialogues were regenerated, resulting in different versions of the dialogues per iteration. Each iteration was compared to the previous iteration(s) to find support, expansions and contradictions of the findings.

The configurations of the generated dialogues were based on the results of part 2 (see section 3.2). The dialogues were generated with GPT-4 instead of GPT-3.5, as at the time of testing, it was deemed the most powerful GPT model.

This section will first explain the grounded theory method in more detail, with the different steps involved. Then the different iterations of analysis will be described, which include the different configurations, topics of discussion and approach of analysis. The results of the grounded theory analysis will be discussed in the main results section (See section 4).

3.3.1 Grounded theory approach

The grounded theory method was used to investigate the capabilities of GPT-agents in a dialogue format. The grounded theory emphasises an inductive analysis [88] and is generally utilised in social sciences. Instead of testing a pre-existing hypothesis or existing theories, the grounded theory method aims to build towards new hypotheses from observations and form a new "grounded theory". This gives the method a very open and explorative approach. The methods and processes involved create an iterative and unfolding process [24], where the emphasis of the research may evolve. Therefore, analysis of the data was not on slow thinking characteristics. Instead, after analysis the findings were to be interpreted to whether these relate to, or can be identified with, slow thinking output. Grounded theory is executed by acquiring and analysing data in different stages as illustrated in figure 13.

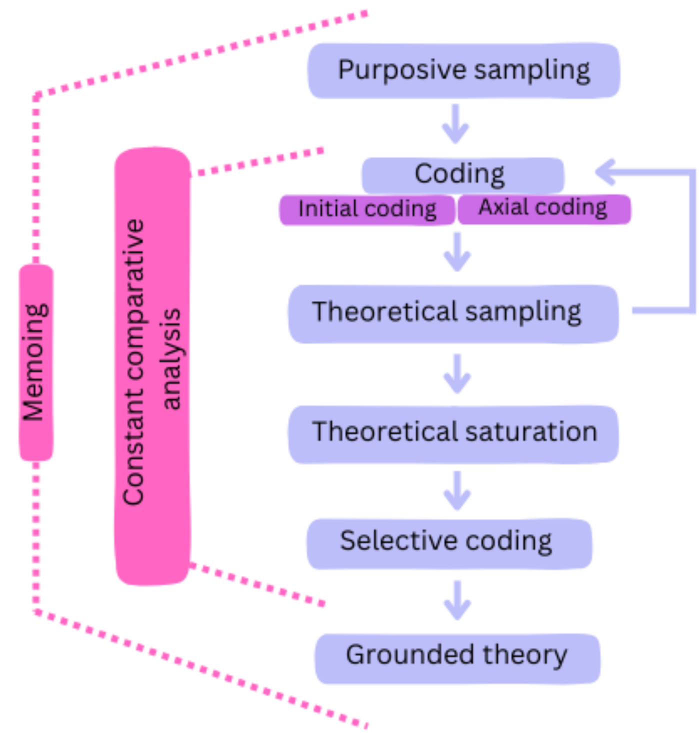

Figure 13: Different steps of the grounded theory method.

This research did not use the grounded theory in its original form. As this study was mainly concerned with how agents interact, the grounded theory was mostly used as a method, as it was unclear whether creating a grounded theory could be applied to agent-agent interaction or whether it was appropriate for this study. The grounded theory steps in this study were followed until the ”theoretical saturation” step. It allowed the analysis of the data through ”coding”, and assess whether patterns, themes and findings of dialogues were supported, extended or contradicted in further analysed dialogues.

The motivation for using the grounded theory method was that it was hypothesised the analysis could provide insight into the overall patterns in how the agents interact, while a quantitative method may only focus on a specific output. Furthermore, quantifying slow thinking behaviour is difficult when the scenario does not involve a task with a correct or incorrect answer. In the scenario of this study, which involves the analysis of LLM agent-agent interactions in discussions, interpreting fast and slow thinking characteristics may be highly subjective. The grounded theory method was used to systematically analyse the different patterns and themes emerging from the dialogues, which after could be interpreted in the context of fast and slow thinking. Finally, this method could account for unexpected findings. Each step will be described in the following sections.

Purposive sampling

Purposive sampling is used in qualitative research and aims to align the sample more closely with the research objectives, thereby improving the quality of both the data and results [20]. The sample is chosen with purpose, not randomly [31]. An example of purposive sampling could be to find a suitable population sample of mothers for a research about ”how mental well being for new mothers can be achieved”. Certain criteria for the mothers could be included such as ”first time mothers”, ”singleton pregnancy” and ”maternal age over 18 years” [20].

In the context of this research, purposive sampling could be seen as part 1 and 2 of this research's methodology. As different agent architectures and configurations were tested, the findings from these tests were used to find a suitable sample of prompts which could give insight into the capabilities of agents.

Coding

The next phase was to code the generated dialogues. Coding involves assigning labels to excerpts of the data such as providing a categorisation and summarisation of the excerpt [23]. Coding is a non-linear process where researchers move back and forth between different coding stages while being attentive for theoretical possibilities [102]. A qualitative data analysis tool NVivo 14 was used to code the excerpts and keep track and group the codes. Coding can be divided two stages: initial coding and axial coding.

- **Initial coding** During the initial coding phase the text of the dialogues would be broken into excerpts. These excerpts would then be grouped into codes. The excerpts can be broken down word by word, line by line, paragraph by paragraph, or with more than one of these strategies [102]. Coding of the dialogues was mostly broken down per sub-sentence or sentence. Excerpts were then compared to each other and grouped together in codes. However, it was important that codes had to fit the data instead of forcing the data to fit the code [39].

  Coding the dialogues was not only be done on the content (what the agent was saying), but also on the form (the type of speech act of the utterance). It allowed to not only focus on the content of the agent's utterances, but also how the agents interacted.

- **Axial coding** Axial coding is the process of comparing codes to each other which can be clustered together into categories of codes [27]. The categories may be broken down or merged with other categories in later phases of the coding.

Constant comparative analysis Another fundamental process in the grounded theory method is constant comparative analysis which takes place throughout coding. It involves the process of constantly comparing excerpts, codes and categories. Excerpts are compared to other excerpts and codes, and codes are compared with other codes [38]. Eventually codes can be clustered into categories [24]. New data would then be compared with already existing excerpts, codes and categories. The process can discover contradictions, expansions or support of the codes [27].

- *Contradictions* were when excerpts contradicted the code, which may indicate the code needed to be adjusted or more information was needed to explain the contradiction.
- *Expansions* occurred when an excerpt provided additional details or elaborated on the code. It may indicate that new information was being found which could extend information of existing codes.
- *Support* occurred when new excerpts supported the existing codes without introducing additional information. Theoretical saturation may be reached when new excerpts keep supporting existing codes, as no further insights were provided.

Memo writing A fundamental process used throughout the analysis is memo writing. Memos function as a storage of ideas generated and documented through interacting with data [95]. The memos can be seen as notes which keep a record of the researcher's reflections. Memos help track the thought process, encourage reflexivity on biases and assumptions, and support development of the theory by tracking emerging patterns [27]. During the analysis of the dialogues, the memos were useful to keep track of potentially interesting observations or patterns.

Theoretical sampling

Theoretical sampling is the process where researchers pursue leads within the data by sampling new participants or material which provide relevant information [24]. A characteristic of the grounded theory is that not all data at first is collected, but data collection and analysis occur in parallel [102].

Theoretical sampling is intended for the formation of a theoretical category, opposed to sampling for the representation of a population [24]. While representation sampling involves the extent of accurately reflecting different characteristics and attributes of a population, theoretical sampling aims to add additional information, identify gaps in current data or reveal new insights which were not yet known.

In this research, sampling involved the generation of new dialogues. These were generated per iteration, after all dialogues in the previous iteration had been coded. The newly generated dialogues in the following iterations could account for any additional information, gaps or new insights. As the analysis consisted of three iterations, the dialogues were generated a total of three times. Theoretical sampling took place until theoretical saturation had been reached.

Theoretical saturation

The collecting and analysis of data can be terminated when theoretical saturation is reached. It occurs when new data does not provide new theoretical insights, or reveals additional attributes to its codes and categories [102]. This research required three iterations before reaching theoretical saturation.

Omitted steps of grounded theory

This research did not follow the last two steps of the grounded theory: selective coding and forming a grounded theory. In the standard use of the grounded theory, all findings are brought together during selective coding after theoretical saturation. Selective coding aims to analyse how all codes and categories can be connected under one core category [27], and how they might relate to each other as hypotheses to be integrated into a theory [102]. The core category can be an existing category or a new category derived from the existing findings [27].

The final step would be to write the grounded theory. The aim is to produce a theory derived from, or grounded in, data produced and gathered by the researcher [24]. As this research was less considered with creating a theory, and more with analysing agent behaviour in dialogue, these step were not conducted.

3.3.2 Generating agent-agent discussions

The analysis of the dialogues was done in three iterations. In each iteration the dialogues were regenerated in order to create new data, and involved adjustments in the analysis to discover new insights. Five unique discussion topics were generated based on findings of the configurations found in part 2 of the methodology. The following sections will discuss the generated topics and used configurations regarding hyperparameters and conditions.

Topics Discussions on five different topics were generated with in total 10 dialogue turns. The topics were chosen semi-arbitrarily, where more importance was put on the variety of the topics than on the topic itself. This allowed to analyse whether discussions showed similarities or differences across the various topics. The topics did not show great differences in agent behaviour during part 2 of the methodology. The first three discussions loosely related to the Dutch general election of 2023, while the latter two discussions were chosen on more general topics.

Topics which loosely related to the Dutch general election of 2023:

1. **Minimum wage in the Netherlands.** The discussion was not on whether the minimum wage should be raised, but to what amount and in what time frame. The topic was chosen as the discussion provided room for nuance, where both agents were generally conditioned to want the same, but to a different degree.

2. **Livestock in the Netherlands.** Related to the notion whether the livestock in the Netherlands should be halved or not. The discussion made the agents have contrasting opinions which had to be defended.

3. **Energy transition.** Related to the discussion whether nuclear energy should be part of the energy transition or not. Similarly to the minimum wage topic, agents generally wanted the same (an energy transition), though had contrasting opinions on how it should be achieved (nuclear or renewable energy).

Topics on more general discussions:

4. **Robots in health care.** The discussion included agents having contrasting opinions whether robots should be used in health care or not. It could raise insight into not only the possible risks and benefits of robots in health care, but for technology in general.

5. **Omnivore versus vegan diet.** A general discussion on a dietary choice which could bring up themes as health, personal choice and sustainability.

Hyperparameters Table 1 contains a table with the used hyperparameters. The only changed parameter compared to part 2 was that the model type used was GPT-4 instead of GPT-3.5-turbo.

Parameter	**Value**
Model type	GPT-4
Maximum amount of tokens	150
Temperature	1 (default)
Top P	1 (default)
Frequency penalty	1 (default)
Presence penalty	1 (default)

Table 1: Configurations of the hyperparameters for dialogue generation.

Conditioning The following conditions were given to the agents.

- Agents were conditioned to have an opposite or conflicting opinion from the other agent depending on the topic.
- Conditions were also given on the length of the utterances. Agents were conditioned to answer in one to two sentences or one to three sentences depending on the topic or iteration.
- Agents in topic 1 and 2 were given an extra instruction to refute each other's arguments.
- Agents in topic 1 were given extra instructions to keep coming up with new arguments.

The extra instructions of the agent in topic 1 and 2 for refuting others' arguments and coming up with new arguments were added to analyse whether these would show different effects on the dialogues. As the instructions did not show notable difference in part 2 using GPT-3.5, it was tested whether changes would arise using GPT-4.

The agent starting the conversation were presented with the topic of the discussion in the first prompt. The other agent would then respond to the input of the agent. Below in figure 14 is an example how agents of the ”omnivore versus vegan diet” were prompted and conditioned.

Agent 1

`system message: You believe a vegan diet is better than any other diet. Try to use 1 to 3 sentences.`

`user message: You are talking about a discussion about diets. What do you say?`

Agent 2

`system message: You believe an omnivore diet is better than any other diet. Try to use 1 to 3 sentences.`

`user message: Someone said the following: *Agent 1 response.* How do you react?`

Figure 14: Example of agent conditioning to set up the dialogue.

3.3.3 Iterations

The following section will describe which topics and conditions were used per iteration. Also, the approach for coding and modified coding approach for the subsequent iterations are discussed. As the results influenced subsequent iterations, the approach per iteration will be described in more detail in the results section (see section 4).

Iteration 1

Topic & conditions All five topics were analysed in iteration 1.

1. Minimum wage in the Netherlands
2. Livestock in the Netherlands
3. Energy transition
4. Robots in health care
5. Omnivore vs vegan diet

The agents were conditioned to use 1 to 2 sentences with the exception of topic 5, where agents were conditioned to use 1 to 3 sentences. This was done to test dialogues with different lengths of utterances. As mentioned before, topic 1 and 2 were conditioned with extra instructions. Agents in topic 1 were conditioned to come up with new arguments and refute each other's arguments. Agents in topic 2 were conditioned to refute each other's arguments. For the exact conditions used in iteration 1 per agent and topic, see appendix C.1.1.

Coding The dialogues were divided in excerpts per sentence or part of a sentence to be coded. Excerpts which contained the same content were put into the same code. As mentioned before, coding was not only done on the content of the excerpts (what the agents are saying), but also on the form (type of excerpt such as ”argument”).

After coding all the excerpts of a dialogue topic, the codes were compared and sorted into clusters. When a code did not seem to fit in any of the existing clusters, the code would form a new cluster in itself, with possibly more codes entering that cluster. After all the codes were clustered, a label was assigned to each cluster describing all the codes within that cluster. Clustering was done in a round focusing on content and in a separate round on form. The coding process of iteration 1 will be discussed again with examples of the codes in the results section (see section 4.1).

Iteration 2

Topics & conditions The second iteration analysed three topics instead of five. This was because all agents in iteration 2 were conditioned to produce a maximum of three sentences instead of two. This allowed more room in responses from the agents. Three dialogues consisting of three sentences per utterance would roughly contain the same amount of data as five dialogues with two sentences per utterance. Furthermore, later analysed topics (topic 4 and 5) in iteration 1 did not show noticeably more results. Therefore, three dialogues were analysed in the next iterations.

All other configurations and conditions remained the same as in iteration 1. All conditions per topic can be found in appendix C.1.2. The following topics were chosen for iteration 2:

1. Minimum wage in the Netherlands
4. Robots in health care
5. Omnivore vs vegan diet

Topic 1 was chosen as it was the only topic which converged in iteration 1, therefore it was analysed whether it would converge again. Topic 4 and 5 were chosen as these included a larger variety of points used in argumentation.

Modified coding approach In the previous iteration, coding and categorising were done per dialogue topic with codes of different agents clustered together. This iteration started to categorise codes of agents separately from each other, in order to account for potential similarities or differences between agents. Also, different excerpts with the same content were not put in the same code, thereby accounting for repetitions more easily.

Furthermore, initial observations of clustering on form showed many counter-arguments. The second iteration focused on whether the counter-arguments used were actually referencing to what the other agent had said, or occurring due to something else.

For the remainder, the coding process was the same as iteration 1. Dialogues were broken down into excerpts, labeled through coding, compared with existing codes, clustered, and clusters were labeled. Codes were clustered on both form and content again. The coding process of iteration 2 will be discussed again with examples of the codes in the results section (see section 4.2).

Iteration 3

Topics & conditions Iteration three consisted of three topics. The following topics were regenerated and analysed:

3. Energy transition
4. Robots in health care
5. Omnivore vs vegan diet

Topic 4 and 5 were maintained to analyse whether the same findings occurred when regenerating the same topic, while topic 1 was switched for topic 3 to diversify the data. No conditions from iteration 2 were altered in iteration 3. All conditions per topic can be found in appendix C.1.3.

Modified coding approach The second iteration showed various argument types when considering whether the arguments referenced back to something. Instead of only focusing on what an agent was responding to, the third iteration also examined whether an argument received a response. It allowed to test if certain parts of codes were ignored by the responding agent, or showed other insights. The coding process of iteration 3 will be discussed again with examples of the codes in the results section (see section 4.3)

3.3.4 Recap - Part 3: Iterations of the grounded theory analysis

This section described the different steps of the grounded theory. This study used the grounded theory as a method to analyse agent-agent dialogue. Therefore, only the steps of purposive sampling, coding, theoretical sampling and theoretical saturation were followed. These included finding the right configurations to generate dialogues, coding the data for emergent patterns, themes and findings, and generating new dialogues until no new insights were found.

Discussions were generated on five different topics and analysed along three different iterations. All five topics were analysed in the first iteration, while three topics were analysed in iteration 2 and 3, as the discussions contained more text and analysing more than three topics did not show noticeably more results. Codes were clustered on content and on form. The coding approach was modified in later iterations by distinguishing codes of different agents, accounting for repetitions, noting when a code was a direct reaction to a point in the previous utterance, and noting whether a code was directly reacted to in the following response. The iterations will be discussed in more detail in the results section (see 4).

3.4 Limitations of methodology

Researcher bias As the initial testing of architectures and configurations were tested and analysed through interpretations, and not objective methods, there is a possibility certain directions were pursued which may have been influenced by researcher bias. Data may have been selectively observed, or personal biases may have influenced the way data was interpreted. To mitigate these limitations, design choices were aimed to be justified based on the results of the executed tests. Furthermore, as this study was explorative, the approach of this study allowed to view many different outputs of the model, and test many different configurations in order to produce dialogue which could give insight on the capabilities of the agents.

GPT-3.5 testing GPT-3.5 was mostly used to test different architectures and configurations of the agents. GPT-4 was used for the final generated dialogues which were analysed through the grounded theory method. However, as these are two different LLMs, the models may react differently to different prompts. Due to the high cost and slow output of GPT-4, it was decided to initially test most architectures and configurations with GPT-3.5, due to it being cost-effective and outputting data faster. GPT-4 was used for the final dialogue generations as it was believed it would adhere better or equally as good to the given conditions, as GPT-4 is considered the current most powerful GPT model and surpasses GPT-3.5 on many tasks [17].

Topics and conditioning The topics which were presented to the agents, and the way agents were instructed through conditions can be of influence on the output of the model. Slight differences in the input to agents, may lead to a completely different output. It is therefore possible that different instructions and topics, other than those chosen in this study, may have led to a more desirable output. This was aimed to be mitigated by testing a wide variety of configurations and topics. Though due to the scope of this study, it was not possible to test these more extensively.

Also, as this study put an emphasis on the ways agents interacted through dialogue, "optimal" topics and conditions may have been less relevant. The same results may also be derived from configurations different from those used in this study.

Grounded theory analysis The grounded theory was used as a method to analyse agent-agent dialogues. Limitations related to this method are the subjectivity of the researcher in interpretations of the data and the generalisability of the results. These were aimed to be mitigated by staying as "grounded" as possible to the data of the dialogue, and by considering to what extent the implications of the results apply to LLMs in general. The grounded theory had the benefit of analysing dialogues thoroughly in a qualitative way by examining how agents interact, with the possibility of discovering unexpected findings.

4 Results

This section will cover the results of the analysis found with the grounded theory method. The analysis of the dialogues was executed in three iterations, with five main topics. The results will be presented for each iteration, separately per dialogue topic. Although the topics of the dialogues were the same across the iterations, the dialogues per iteration were unique because they were regenerated each time. Iteration 1 covered all five topics. As the agents after iteration 1 were conditioned to produce longer utterances to allow for more ways to interact, and analysing more topics did not show many new results, three instead of five topics were analysed in iteration 2 and 3. Topic 4 and 5 were kept in all iterations for consistency and to account for potential reoccurring patterns, while topic 1 and 3 were alternated to diversify the results. The following topics were covered with the accompanying iteration:

1. Minimum wage in the Netherlands (Iteration 1, 2)
2. Livestock in the Netherlands (Iteration 1)
3. Energy transition (Iteration 1, 3)
4. Robots in health care (Iteration 1, 2, 3)
5. Omnivore versus vegan diet (Iteration 1, 2, 3)

The results of the iterations were derived from the stages of open coding and axial coding. Open coding focused on the form and content of the excerpts, which in the axial coding stage resulted in separate clusters for form and content. "Form clusters" showed the different ways agents interacted, while "content clusters" showed which points were raised during the discussion, and how often points were repeated. After each iteration, theoretical sampling took place. It involved regenerating dialogues of the desired topic to identify reoccurring themes or theoretical gaps which require further exploration. This process was continued until theoretical saturation was reached, which would occur when further analysis did not lead to new insights. Below are the main results for each iteration.

Iteration 1 The main clusters derived from the codes on *form* showed agents mainly use arguments and acknowledgements in the dialogues. Clusters on *content* showed the different stances and points made by agents during the discussion. Codes containing multiple excerpts showed agents repeating points. Memoing revealed inconsistencies in the language of the agents, and an agent correcting the other agent on a false statement.

Iteration 2 The results of iteration 2 supported the results of iteration 1. Additionally, arguments could be divided in *general arguments* which supported the stance of the agent, *general counter-arguments* which countered the stance of the other agent and *specific counter-arguments* which countered a certain point of the other agent. New clusters found were *linking word* and *announce point*.

Memoing observed a structure in the ways agents formed a reaction. Generally, the agent started with an acknowledgement of a point of the other agent's last response, followed by a (counter-)argument. Following the first argument, sentences could start with another acknowledgement, linking word, announcing the point the agent would form the next argument on, or directly produce other arguments. Agents generally only reacted to the other agent's last utterance.

A concluding prompt was added to the dialogues to examine to what extent the agents agreed as none of the dialogues in iteration 2 converged. The concluding prompt gave insight to the concluding opinions of the agents. It also showed that agents were able to consider points of the whole dialogue when prompted, instead of only considering points of the last utterances which was observed in the standard dialogue.

Iteration 3 The results of iteration 3 supported the results of iteration 1 and 2. Furthermore, focus was set on how arguments were reacted to. These could be divided into four different clusters: *answered arguments* involved arguments which received a direct reaction. It could involve reacting to a few or all points of the arguments. *Discussed arguments* occurred when the responding agent mentioned the points of the arguments, but did not directly react to the argument. *Covered arguments* involved the agent reacting to the argument with a general statement, without mentioning any of the points of the argument. *Not answered arguments* occurred when the agent did not react to the argument and did not mention any of the mentioned points.

After three iterations theoretical saturation was reached as notable findings were only being supported, and not extended or contradicted.

4.1 Iteration 1

Topics Iteration 1 consisted of all five topics.

1. Minimum wage in the Netherlands
2. Livestock in the Netherlands
3. Energy transition
4. Robots in health care
5. Omnivore versus vegan diet

Each topic and its results from the coding phases will be discussed individually. The first topics will be more extensively discussed compared to the later topics due to the many similarities in the results. The detailed observations per topic can be skimmed through upon first reading, while focusing on the recap sections per topic, and the concluding recap for the first iteration in section 4.1.7.

Open coding The excerpts were coded on content, with a note added about the form of the excerpt when it could not be derived from the name of the code. Below in figure 15 are two examples of codes with the accompanying excerpt.

Label Code	Excerpt from dialogue
Acknowledge poverty reduction	`Agent 1: While I acknowledge the potential positive outcomes of an €18 minimum wage, like poverty reduction (...)`
Boost economy (argument)	`Agent 2: Additionally, the increase will encourage more people to join the workforce contributing to economic growth.`

Figure 15: Two instances of how excerpts were coded.

The labeled code "acknowledge poverty reduction" already gave enough information that the code was a form of an acknowledgement. However, the code "boost economy" did not necessarily state it was an argument. Therefore, "(argument)" was added to the code to indicate this.

Axial coding Axial coding was done in two different ways: on form and on content. The results from these stages will be presented separately. Note that both axial stages were coded on the same set of codes derived from open coding, the main difference was in the different clusters formed from these codes.

Main results The main clusters formed from the codes across the topics were "arguments" and "acknowledgements" in both clustering on form and content. Topic 1 was the only topic where the topic converged which resulted in an additional "converge" cluster. All topics showed repeated excerpts: agents stating the same argument or acknowledgement again. Memoing exposed inconsistencies of the utterances of the agents and a correction of a false statement of the other agent.

4.1.1 Topic 1: Minimum wage in the Netherlands

Agents in topic 1 formed a discussion on the minimum wage in the Netherlands. Context that the current minimum wage is €12.40 was provided. Agent 1 was conditioned to argue for an increase to €14 while agent 2 for an increase to €18.

Axial coding on form

Figure 16 shows the clusters of topic 1 made during axial coding based on form. The amount of excerpts per cluster is indicated between brackets. Each cluster will be discussed more elaborately below the figure.

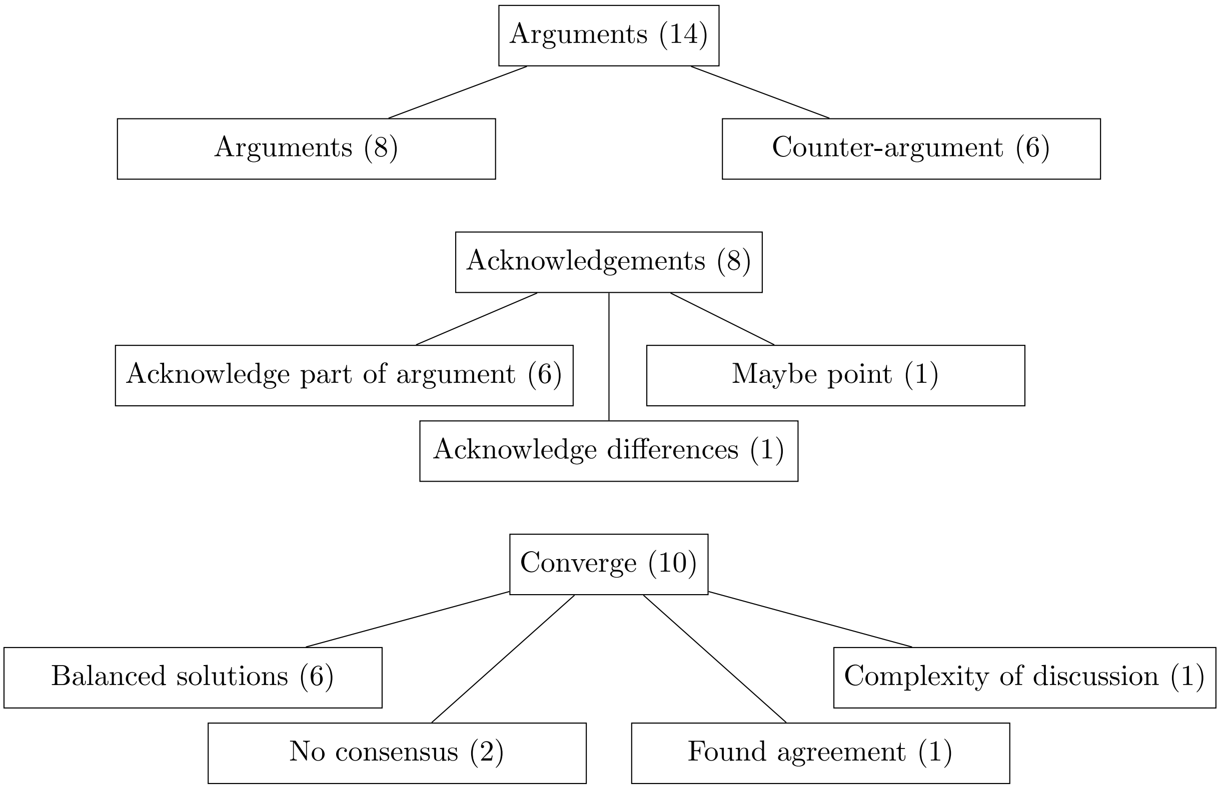

Figure 16: Iteration 1, topic 1. Clusters based on form.

Arguments Agents made use of arguments. Arguments could be novel or have occurred already. When already occurred, a code could consist of multiple excerpts. In figure 17 two examples of the codes are shown.

Argument codes	Excerpt
`Help workers afford costs`	`Agent 2: Raising the minimum wage to €18 an hour will help workers afford the rising cost of living, particularly in urban areas.`
	`Agent 1: It allows workers to afford the cost of living`
`Tackle poverty`	`Agent 2: an increase to €18 will ensure more people are lifted out of poverty`

Figure 17: Argument codes with accompanying excerpts.

Arguments could be divided into the sub-clusters ”arguments” and ”counter-argument”, which are briefly discussed below.

- Argument: an argument the agent made which supported the stance it was conditioned to.
- Counter-argument: an argument which countered the stance and arguments of the other agent.

Acknowledgement Agents stated an acknowledgement with a point the other agent had made. In figure 18 are two examples of codes. The codes in this example have the same label as the cluster it belonged to.

Acknowledgement code	Excerpt
Acknowledge part of argument	`Agent 1: Certainly, there might be areas with extraordinarily high living costs where €14 an hour is not enough.`
	`Agent 1: While I acknowledge the potential positive outcomes of an €18 minimum wage, (...)`
Maybe point	`Agent 2: Although €14 an hour might appear balanced, (...)`

Figure 18: Acknowledgement codes with accompanying excerpts.

The acknowledgement cluster could be divided in the sub-clusters ”acknowledge part of argument”, ”acknowledge differences” and ”maybe point”.

- Acknowledge part of argument: acknowledging a part of the other agent’s argument.
- Acknowledge differences: acknowledging there are differences with the other agent.
- Maybe point: when a point of the other agent is repeated with skepticism such as stating something ”might appear some way”, implying the agent does not entirely agree with the statement.

Converge This cluster was formed by codes which showed agents converging. Figure 19 shows examples of two codes.

Converge Codes	Excerpt
Agree minimum wage should be increased	Agent 1: (...) but agree on the need for an increase in the minimum wage.
	Agent 2: Indeed, we agree on the need for minimum wage increase.
Not reached consensus	Agent 1: Thus, we have not reached consensus on the exact figure, (...)

Figure 19: Converge codes with accompanying excerpts.

The converge cluster could be divided into the sub-clusters "balanced solutions", "no consensus", "found agreement" and "complexity of discussion".

- Balanced solutions: stating balanced solution for the topic being discussed.
- No consensus: stating that consensus was not reached.
- Found agreement: a statement which stated agreement was found.
- Complexity of discussion: a statement highlighting the complexity of the discussion.

Takeaways - Axial coding on form

The following was seen in topic 1 based on form:

- Agents followed speech acts which related to arguments, acknowledgments and converging stances.

Axial coding on content

Figure 20 shows the argument clusters found based on content in topic 1. The acknowledgement and converge clusters did not show any novelties compared to the clusters based on form and were therefore not included in the figure.

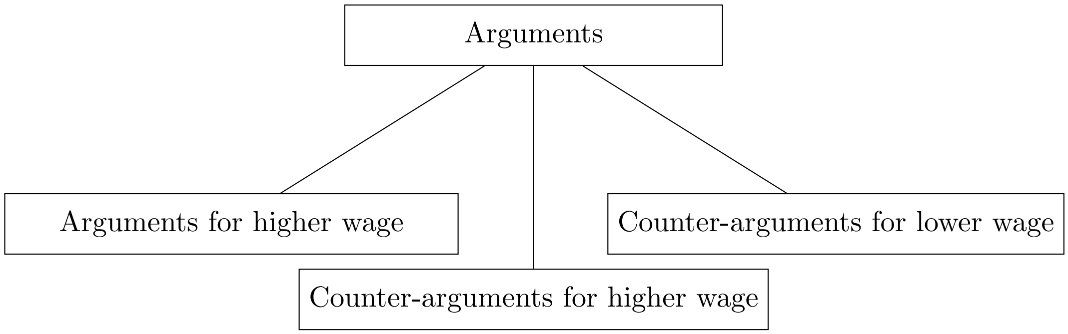

Figure 20: Iteration 1, topic 1. Clusters based on content.

The argument cluster consisted of "arguments for higher wage", "counter-arguments for higher wage" and "counter-arguments for lower wage".

- Arguments for higher wage: consisted of the same codes as the sub-cluster "arguments" clustered on form, which included arguments for a higher minimum wage.
- Counter-arguments for a higher wage: this sub-cluster was similar to the sub-cluster "counter-argument" clustered on form, while only consisting of codes in favour of a higher wage.
- Counter-arguments for a lower wage: also similar to the sub-cluster "counter-argument" on form, while only consisting of code in favour of a lower wage.

Repetitions of codes Table 2 shows codes which consisted of multiple excerpts and were therefore repeated during the dialogue. "Counts" state the total amount of occurrences of the code, "code" indicates which code was repeated.

Counts	Code
4	`(Minimum wage) boost economy`
3	`Finding balance (between workers rights and burden business)`
3	`Suggest best of both worlds`
2	`Help workers afford costs`
2	`Agree minimum wage should be increased`
2	`(Consider) long term effect`

Table 2: Iteration 1, topic 1. Codes with multiple excerpts, indicating repetitions.

Takeaways - Axial coding on content

The following was seen in topic 1:

- Arguments could be divided in regular arguments and counter arguments which countered the other agent.
- Repetitions were observed by analysing which codes had multiple excerpts.

Observations from memoing

The following observations were made from memoing. Memoing showed observations which did not directly link to the codes and clusters of the dialogue.

- **Last response:** The agents generally only reacted to the other agent's last response, and not to any information stated before.
- **Structure of dialogue:** The structure of the dialogue generally consisted of an agent starting the utterance by acknowledging a point of the other agent, and then stating arguments after.
- **Codes of the converge cluster:** The last part of the dialogue mostly consisted of codes of the converge cluster. Convergence was mainly about the need to find a balanced solution on the topic, instead of a concrete solution.

Recap - Topic 1 (iteration 1)

The codes and clusters of topic 1 showed that agents mainly use arguments and acknowledgements. The agents converged at the end of the discussion which formed the converge cluster during axial coding on form. Axial coding on content showed the different stances of the agents. The amount of excerpts per code showed the amount of repeated points.

Observations from memoing showed agents generally responded to only the last response of the other agent, the structure of each utterance consisted of an acknowledgment followed by arguments, and that the last part of the dialogue mainly consisted of converging codes.

4.1.2 Topic 2: Livestock in the Netherlands

Agents in topic 2 formed a discussion on the amount of livestock in the Netherlands. Agent 1 was conditioned to argue for a halving of the livestock in the Netherlands, while agent 2 was conditioned to argue against a downscale of livestock.

Axial coding on form

Figure 21 shows the clusters found in clustering on form, with the amount of occurrences per cluster.

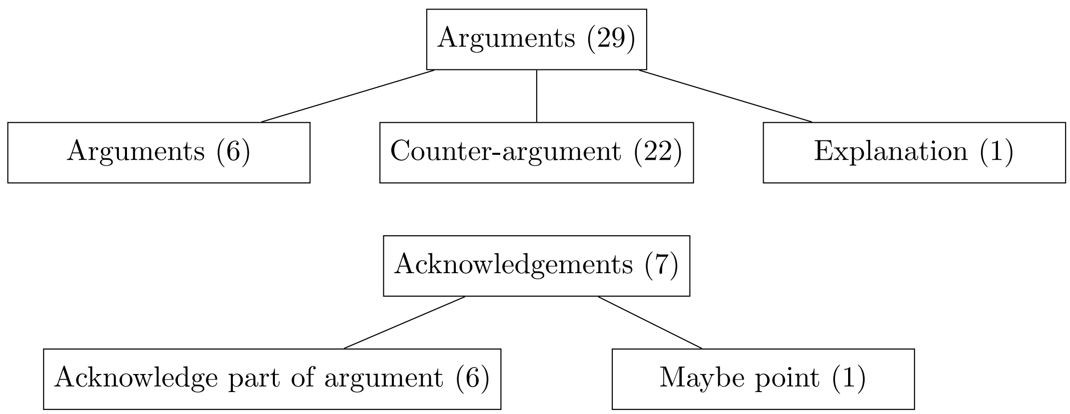

Figure 21: Iteration 1, topic 2. Clusters found during axial coding on form.

The clusters and sub-clusters were generally the same as the previous topic. The converge cluster did not reappear as the dialogue in topic 2 did not converge. The only new cluster was ”explanation”, which will be discussed below.

Explanation Explanation is used to strengthen the argument of the agent. It may also be seen as a part of a counter-argument. Figure 22 shows the explanation code.

Explanation code	Excerpt
Explain how large scale agriculture has been maintained	`Agent 2: (...), the Netherlands has been able to sustain it due to its rich soil and suitable climatic conditions; (...)`

Figure 22: Explanation code with the accompanying excerpt.

Takeaways - Axial coding on form

The following takeaways were found during axial coding on form.

- Topic 2 generally saw the same clusters as topic 1.
- The dialogue in topic 2 did not converge, leading to more codes in the argument cluster.

Axial coding on content

Figure 23 shows the clusters found during axial coding on content for topic 2. Similar to topic 1, the higher level clusters on form (arguments and acknowledgements) were generally the same as the higher level clusters on content. The acknowledgement cluster did not show any novelties compared to the form and is therefore not included in the figure.

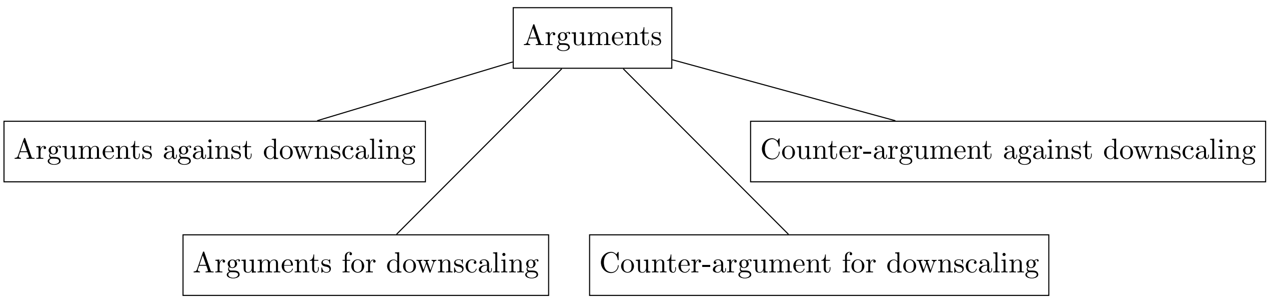

Figure 23: Iteration 1, topic 2. Clusters based on content.

The arguments cluster consisted of the sub-clusters ”arguments against downscaling”, ”arguments for downscaling”, ”counter-argument for downscaling” and ”counter-argument against downscaling”. While regular arguments supported the stance of the agent, the counter-arguments countered the argument and stance of the other agent. The sub-clusters will be described in more detail below.

Arguments against downscaling These included general arguments which stated the harm downscaling would cause.

Arguments against downscaling	Excerpt
Affect livelihood farmers	`Agent 2: It would also affect the livelihood of many farmers who depend on livestock farming.`

Figure 24: Argument clustered against downscaling

Arguments for downscaling These included general arguments which stated the harm livestock causes.

Arguments for downscaling	Excerpt
Environmental impact	`Agent 1: the intense livestock farming in the Netherlands has a severe impact on the environment, causing air, soil, and water pollution.`

Figure 25: Argument clustered for downscaling

Counter-argument against downscaling This sub-cluster formed counter-arguments against the concept of downscaling livestock in the Netherlands.

Counter-arguments against downscaling codes	Excerpt
Modern practices for both thriving livestock and clean environment	Agent 2: modern sustainable farming practices are increasingly negating this impact, thus allowing for both a thriving livestock industry and a clean environment.

Figure 26: Counter-argument clustered against downscaling

Counter-argument for downscaling This sub-cluster contained counter-arguments in favour of halving the livestock in the Netherlands.

Counter-argument for downscaling codes	Excerpt
Sustainable farming leads to responsible and potentially profitable practices	Agent 1: Additionally, shifting towards more sustainable farming practices doesn't mean locals would lose their livelihoods; rather, they'd be retrained in more ecologically responsible and potentially profitable methods of agriculture.

Figure 27: Counter-argument clustered for downscaling

Repetitions of code Table 3 shows which codes consisted of multiple excerpts, and were therefore repeated during the dialogue.

Counts	Code
3	Harmfulness of large scale
3	Questioning feasibility and guarantee transition retraining farmers
2	Proposing environmental mitigation techniques
2	Success other countries not generalisable (local climate, culture, market)

Table 3: Iteration 1, topic 2. Codes with multiple excerpts, indicating repetitions.

Takeaways - Axial coding on content

The following takeaways were found during axial coding on content.

- Arguments were either for or against downscaling.
- The arguments consisted of regular arguments and counter-arguments.
- Topic 2 showed repetitions of excerpts.

Observations from memoing

Observations from memoing in topic 2 were:

- **Last response:** Again, agents generally only reacted to the last response.
- **Structure of dialogue:** The same structure was observed of utterances starting with an acknowledgment, followed by arguments.
- **No convergence**: Agents did not converge.

Other observations from memoing included inconsistencies and corrections produced by the agents.

Inconsistencies Inconsistencies were instances where the responses of the agent were illogical or followed awkward phrasing. The inconsistencies were not noticed by the other agent. The following is an example where an agent had an illogical conclusion of the argument. The word "tradition" is used as an argument for a "balanced ecosystem". It seems a loss of tradition is not logical to maintain a balanced ecosystem, and that only a loss of biodiversity would have been appropriate.

Agent 2: Downscaling livestock in the Netherlands would not only damage the economy but also lead to a loss of biodiversity and tradition, vital for maintaining a balanced ecosystem.

The following is an example of an illogical linking word. The "indeed" seems to imply that agent 2 mentioned retraining programs can be adapted to local conditions, while the agent only mentioned that successes cannot be universally applied.

Agent 2: As for the retraining models, these successes can't be universally applied as they don't take into account factors such as local climate, culture, and market forces that differ in the Netherlands.

Agent 1: The retraining programs can indeed be adapted to local conditions,

```
thus promoting sustainable practices suitable to the Dutch landscape and market
opportunities.
```

Corrections Corrections occurred when one agent corrected the other agent on a certain point. This example shows agent 2 criticising agent 1 for "proposing a one-size-fits-all solution", which agent 1 did not propose. Agent 1 corrects agent 2 on this false statement.

Agent 2: (...) proposing a one-size-fits-all solution is not pragmatic.

Agent 1: (...) no one is suggesting a one-size-fits-all approach; (...)

Recap - Topic 2 (iteration 1)

Topic 2 showed the same clusters "arguments" and "acknowledgements" as in topic 1. The biggest difference was that the dialogue did not converge, leading to the absence of the "converge" cluster. Observations from memoing showed the agents reacting to last response and the same structure of utterances. Furthermore, inconsistencies within the utterances of the dialogue were found, and a correction on a false statement.

4.1.3 Topic 3: Energy transition

Agents in topic 3 formed a discussion on the energy transition: a transition of the global energy sector shifting from a fossil based system to renewable sources. Agent 1 was conditioned to argue in favour of nuclear energy being part of the transition, while agent 2 was against the inclusion of nuclear energy.

Axial coding on form

Figure 28 shows the clusters found in clustering on form, with the amount of occurrences per cluster. Acknowledgements, arguments and counter-arguments were found again in the topic 3. The "opinion" and "acknowledge point" sub-clusters were novel and will be discussed below the figure.

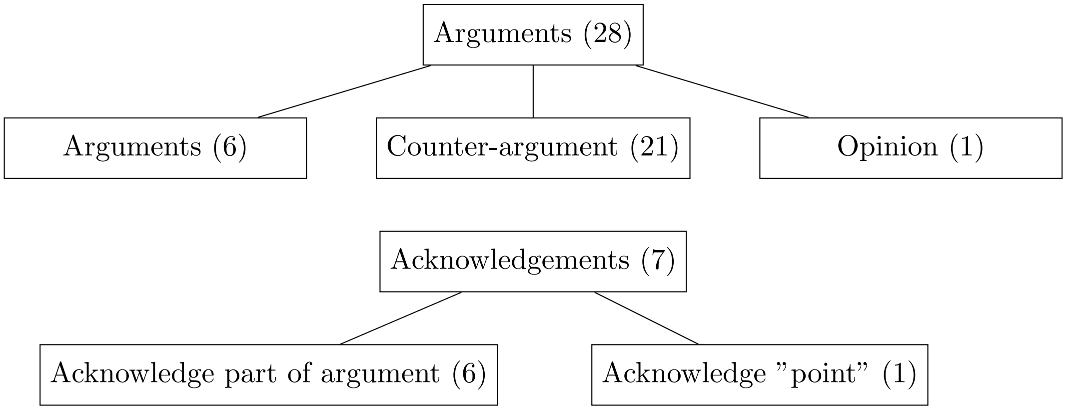

Figure 28: Iteration 1, topic 3, Clusters on form.

Opinion The opinion cluster included one instance, which stated the opinion of agent 2 on the energy transition. It may also be seen as a form of an argument.

Opinion codes	Excerpt
Opinion on approach green energy transition	`Agent 2: Government policies and corporate strategies should focus more on the development and enhancement of these green energy sectors to ensure a safe and sustainable energy transition.`

Figure 29: Opinion code with accompanying excerpt.

Acknowledge "point" This type of acknowledgment occurred when the agent only stated it acknowledged the "point" of the other agent, without specifying what point.

Takeaways - Axial coding on form

Topic 3 showed the following takeaways on form.

- Topic 3 showed the same pattern as previous topics, where the main clusters only consisted of arguments and acknowledgements.

Axial coding on content

The clusters in figure 30 were found during axial coding on content for topic 3. The high level clusters "arguments" and "acknowledgements" were the same as in the previous clusters found on form.

The arguments cluster contained five sub-clusters. Arguments and counter-arguments for nuclear energy, arguments and counter-arguments for renewable energy and a ”balanced approach”. Acknowledgements consisted of the sub-cluster ”acknowledge nuclear energy benefits” and ”acknowledge renewable benefits”.

Only the balanced approach sub-cluster will be discussed, as the other sub-clusters were similar to those in topic 1 and 2, and did not lead to new insights.

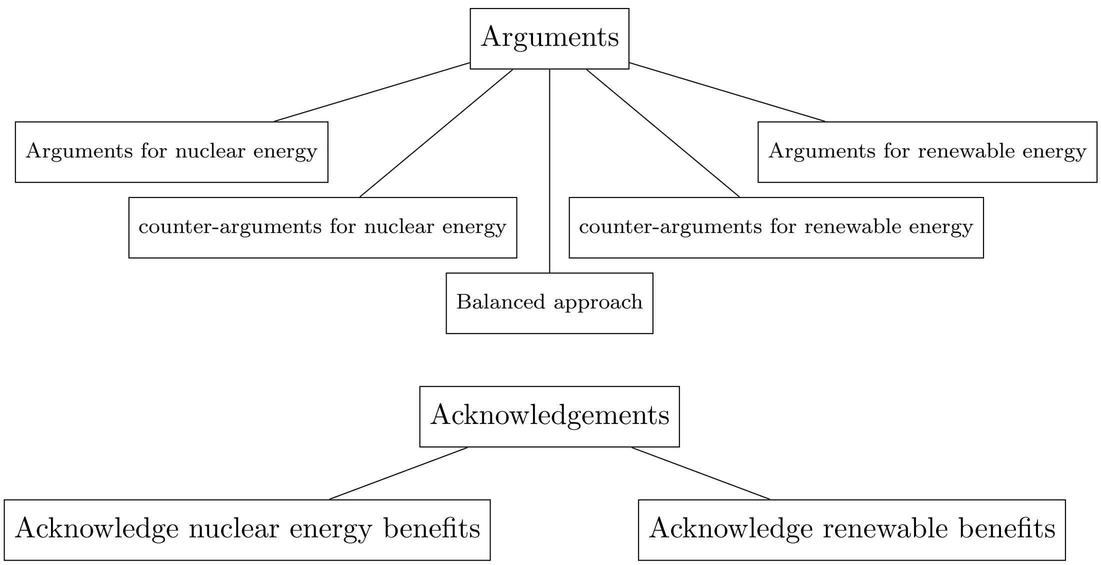

Figure 30: Iteration 1, topic 3, Clusters based on content.

Balanced approach This cluster consisted of agent 1 trying to convince agent 2 that nuclear energy for the energy transition could be part of a balanced strategy, where nuclear and renewables are combined. During the dialogue, agent 2 did not show agreement with this strategy.

Balanced approach codes	Excerpt
Proposition balance of nuclear and renewable energy	Agent 1: `Therefore, an optimal energy transition might benefit from a balanced approach that includes both nuclear and renewable energy, focusing on risk management and industrial advancements.`

Figure 31: Balanced approach code with accompanying excerpt.

Repetition of codes Table 4 shows the repetitions during the dialogue. ”Variations” show whether the repetition had have any slight variations, such as extra points added to the repeated argument.

Counts	Code	Variations
3	Proposition balance of nuclear and renewable energy	-
3	Renewables are less harmful than nuclear energy	nulcear waste
2	Nuclear offers low-carbon efficient energy source	-
2	Dependability solved for renewable with technology	Energy storage
2	Renewable energy more sustainable and environmentally	

Table 4: Iteration 1, topic 3. Codes with multiple excerpts, indicating repetitions.

Takeaways - Axial coding on content

Topic 3 showed the following takeaways based on content.

- Similar to the previous topics, arguments could be divided in regular arguments and counter-arguments which supported a certain stance.
- Similar to previous topics, agents repeated certain points multiple times.

Observations from memoing

Observations from memoing showed the same patterns arise in agents responding to last response and the structure of the utterance.

Recap - Topic 3 (iteration 1)

Topic 3 showed many similarities with the other topics as mainly arguments and acknowledgements were used in the dialogue. Axial coding on content showed the different stances supported by the agent based on the condition.

4.1.4 Topic 4: Robots in health care

Topic 4 had agents discuss whether robots should be used in health care. Agent 1 was conditioned to argue against the use of robots in health care while agent 2 was in favour of the use of robots in health care.

Axial coding on form

Figure 32 shows the clusters found in clustering on form, with the amount of occurrences per cluster. The main clusters were the same as the previous topics "arguments" and "acknowledgements". The only novel sub-cluster was "acknowledge opinion", which will be discussed in the following paragraph.

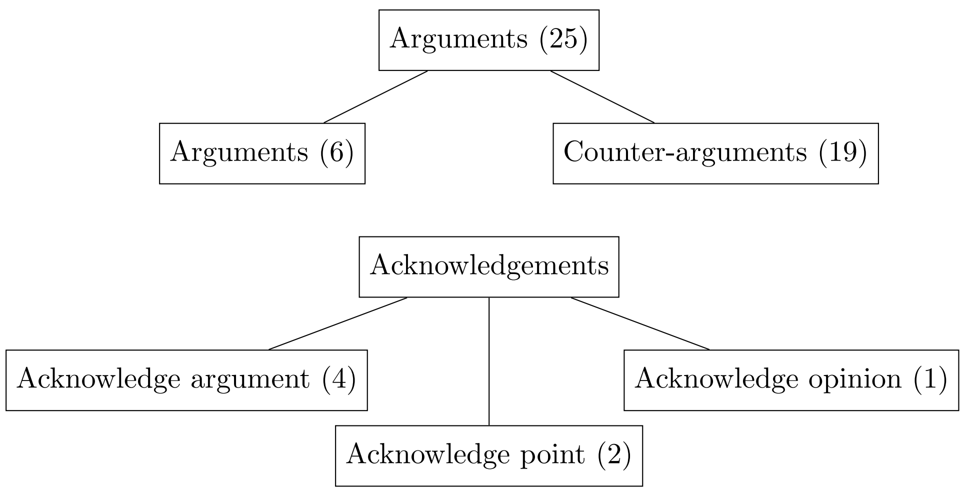

Figure 32: Iteration 1, topic 4. Clusters found on form.

Acknowledge opinion An acknowledgment of opinion involved the agent acknowledging the perspective or opinion of the other agent, without stating the specifics.

Acknowledge opinion code	Excerpt
Respect perspective	`Agent 2: I Respect your perspective.`

Figure 33: Acknowledge opinion code with accompanying excerpt.

Takeaways - Axial coding on form

Topic 4 showed the following takeaways on form.

- The same clusters "arguments" and "acknowledgements" were found as in previous topics.

Axial coding based on content

Figure 34 shows the clusters found on content for topic 4. Similar to the previous topics, the higher level clusters were generally the same as on form. Unlike previous topics, codes were clustered on various points instead of the general stance of the codes. The acknowledgement cluster did not show any novelties compared to the form and is therefore not included.

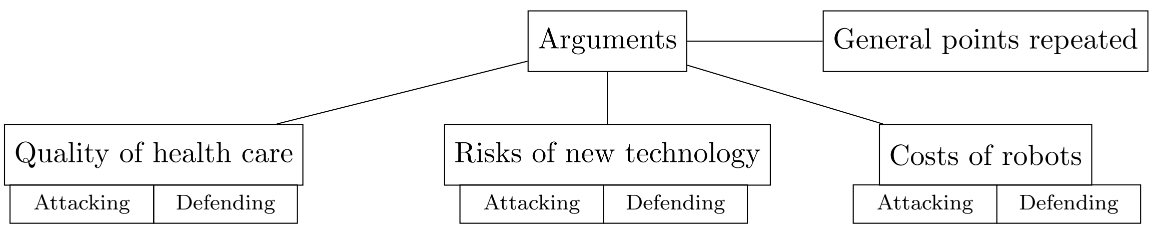

Figure 34: Iteration 1, topic 4. Clusters found on content.

The argument clusters consisted of the sub-clusters "quality of health care", "risks of new technology" and "costs of robots". Each of the sub-clusters could be divided into "attacking" and "defending", based on the different opinions of the agents. Agent 1 attacked the notion of robots in health care, while agent 2 defended it. For further descriptions of the sub-clusters, see appendix D.1

Repetition of codes Table 5 shows codes which were repeated multiple times during the dialogue. Repetitions were not only seen in arguments, acknowledgements were also repeated.

Arguments

Counts	Code	Variations
3	Robots can assist healthcare workers	-
2	Robots can enhance health care service	-
2	Robots cannot replace human touch	-
2	Security risks can be mitigated	-

Acknowledgements

Counts	Code	Variations
2	Acknowledge human empathy/touch cant be replaced	-
2	Acknowledge "points"	-
2	Acknowledge security risks	-

Table 5: Iteration 1, topic 4. Codes with multiple excerpts, indicating repetitions.

Takeaways - Axial coding on content

Topic 4 showed the following takeaways on content.

- The clusters on content in topic 4 also included sub-sub-clusters, which showed the agent either attacking or defending a certain notion
- Repeats were not only seen again in arguments, but also in acknowledgements.

Observations from memoing

The same observations were made on last response and the structure of the utterances. No other notable observations were made in topic 4.

Recap - Topic 4 (iteration 1)

Topic 4 showed the same main clusters on form and content. The sub-clusters on content showed the different points discussed in the dialogue. The points were either used to attack the notion on robots in health care (agent 2) or defend it (agent 1). Repeats were not only found again in arguments as in previous topics, but also in acknowledgements. The same last response observation and structure in utterances was seen, as in previous topics.

4.1.5 Topic 5: Vegan vs omnivore diet

Agents in topic 5 held a discussion whether a vegan or omnivore diet was "better". Agent 1 was conditioned to advocate for a vegan diet, while agent 2 advocated for an omnivore diet. Agents in topic 5 were conditioned to use a maximum of three sentences per utterance compared to two in the previous topics. Therefore, the dialogue contained more codes.

Axial coding on form

Figure 35 shows the clusters found in clustering on form, with the amount of occurrences per cluster. Again, these clusters have been seen in the previous sessions and will therefore not be further elaborated on.

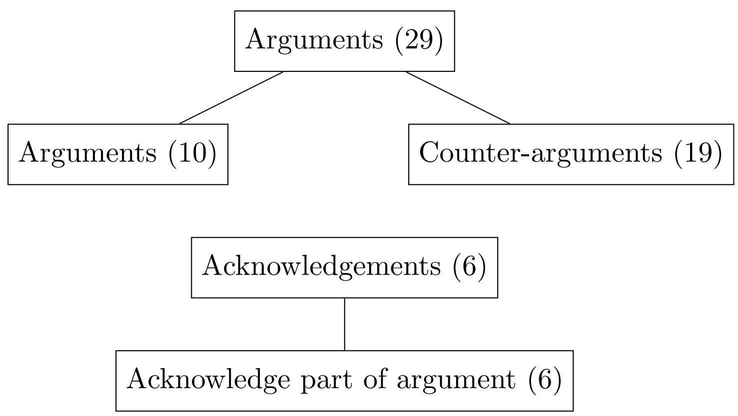

Figure 35: Iteration 1, topic 5. Clusters based on form.

Takeaways - Axial coding based on form

Topic 5 showed the following takeaways.

- Topic 5 did not give any new insights compared to the previous topics on form.

Axial coding based on content

Figure 36 shows the clusters found on content in topic 5. Again, the clusters found on content were generally the same on a higher level as the clusters found on form. The two main clusters found were arguments and acknowledgements. The sub-clusters of arguments were "nutrients", "health", "diversity", "evolution", "sustainability" and "multiple". The clusters either consisted of codes arguing for a vegan diet, omnivore diet or both.

The acknowledgements consisted of ”vegan acknowledgements” and ”omnivore acknowledgements”. As the clusters are similar in output compared to previous topics, and do not provide more insight other than different points being stated, the codes will not be discussed further.

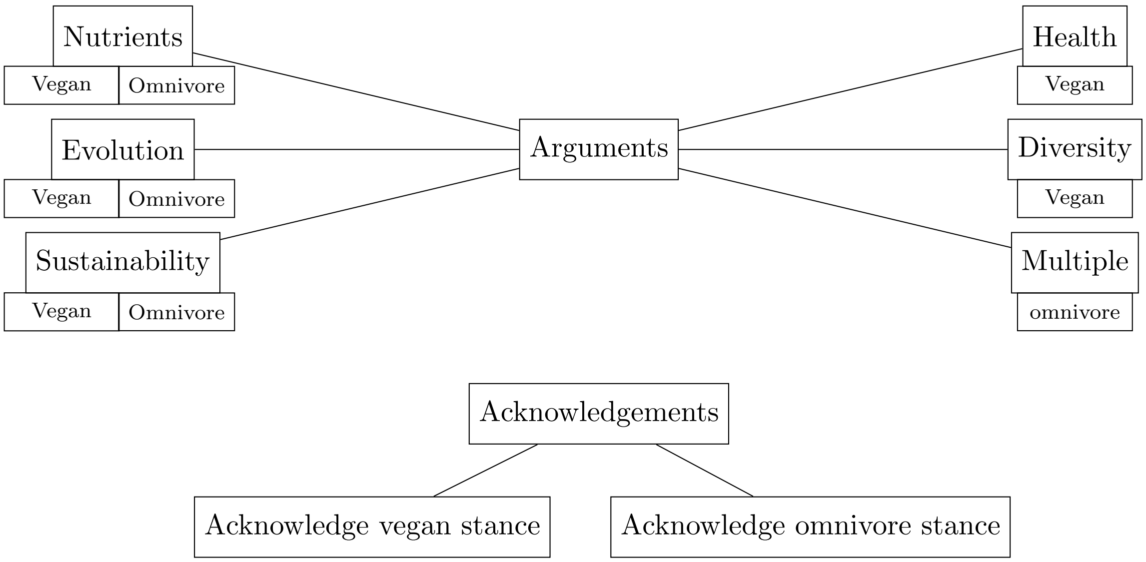

Figure 36: Iteration 1, topic 5. Clusters based on content.

Repetitions of codes Table 6 shows which points were repeated multiple times by the agents during the dialogue.

Counts	Code	Variations
3	`Vegan diet more sustainable`	-
2	`Vegan more healthy`	-
2	`Fortified foods and supplements fill vegan diet`	-
2	`Omnivore diet natural nutrient profile`	-
2	`Supp and fortification does not replace nutrients omnivore diet`	`bio availability`
2	`Omnivore aligns more with evolution`	-

Table 6: Iteration 1, topic 5. Codes with multiple excerpts, indicating repetitions.

Takeaways - Axial coding on content

The takeaways of topic 5 on axial coding on content were as follows.

- The clusters showed the different topics discussed and for which stance it was argued.
- Just as in the other topics, agents in topic 5 repeated certain codes.

Observations from memoing

No notable observations or differences compared to previous topics were found from memoing. Agents only considering points of the last response and structure in utterances was again observed.

Recap - Topic 5 (iteration 1)

Topic 5 showed many similarities with previous topics on the clusters on form and content. The sub-clusters showed the different points discussed during the dialogue, and for which stance (vegan or omnivore) the argument was used. Repeats were found again in the dialogue. No notable observations from memoing were found.

4.1.6 Overall observations from memoing iteration 1

All dialogue topics followed a similar structure, where a response started with an acknowledgement, and was followed by a (counter-)argument. Furthermore, agents mainly responded to only the last response, without considering points of previous utterances.

Based on the condition of agents in topic 1 and 2 to refute each other's argument, no noticeable differences were seen as agents in other topics also countered each other's arguments. Furthermore, agents in topic 1 showed repetitions, even though these were conditioned to come up with new arguments.

4.1.7 Recap - Iteration 1

- During iteration 1, five topics were coded through initial and axial coding. Axial coding was done separately based on form and content.
- The main clusters found during axial coding on form were "arguments" and "acknowledgements".
- Sub-clusters of arguments consisted mainly of "arguments" and "counter-arguments". Less frequent sub-clusters were "explanation" and "opinion". Sub-clusters of acknowledgements

mainly consisted of "acknowledge part of argument". Less frequent clusters were "maybe point", "acknowledge differences" and "acknowledge point".

- Only topic 1 contained the cluster "converge", as it was the only dialogue which converged. The agents stopped convincing each other and agreed a balanced solution was necessary.
- The main clusters found during axial coding on content were "acknowledgements" and "arguments". The sub-clusters of acknowledgements generally involved acknowledging the stance of the other agent, or what kind of point was acknowledged. The sub-cluster of arguments reflected the different stances (topic 1-3) and points (topic 4-5) of the agents.
- Repetitions occurred in every topic. These mainly consisted of repeated arguments, though acknowledgements were occasionally also repeated.
- Observations found during memoing were on the general structure of the dialogues, inconsistencies of the agents and one instance of correcting a false statement.
- A similar structure was found over all topics, where agents started a response with an acknowledgement, followed by a (counter-)argument.
- Agents mainly responded to only the last response of the other agents, without considering previous utterances.

4.1.8 Approach for the next iteration

The approach for iteration 2 was to further discover the ways agents interact through dialogue, and examine whether the observations were supported, extended or contradicted. Iteration 2 will continue to analyse through initial coding, axial coding based on form and content, and observations through memoing. The main difference was in the slight variation on how these methods were executed.

Support Supporting previous results could involve finding clusters and sub-clusters based on form such as *arguments*, *counter-arguments*, *acknowledgements*, *acknowledge part of arguments* and *converge*. Furthermore, clusters based on content showing the same stances of the agents and repetitions, or observations showing the same structure of dialogues, inconsistencies in the text and agents correcting one another could also show support for the previous findings.

Extend Extensions could involve novel clusters and observations. These could arise due to the topics being regenerated differently, or a different approach on how the dialogues were coded.

Open coding was extended with whether arguments were a direct reaction to the other agent. In iteration 1, arguments were often sub-clustered in *arguments* and *counter-arguments*. Iteration 2 focused on whether these arguments were actually used in reaction to the other agent, or that the argument arose due to different reasons.

Open coding also took into account which code belonged to which agent, in order to distinguish codes of separate agents. In iteration 1, agents were analysed together per topic. In some topics clustered on content, clusters emerged which distinguished the different agents. To further explore potential differences or patterns between agents more, such as with the clusters on form, iteration 2 focused on the agents separately.

As four of the five topics did not converge in iteration 1, agents in non-converged dialogues in iteration 2 were prompted to conclude the conversation to analyse the concluding stance of the agents.

Contradict Contradictions could involve finding results which are not in line with the results of iteration 1.

4.2 Iteration 2

Topics In iteration 2, the topics were regenerated. Agents were conditioned to generate a maximum of three sentences instead of two per utterance to allow more room to interact. Consequently, agents produced three sentences per response instead of two. As a result, the dialogues consisted of a larger amount of text. Furthermore, it seemed less topics in iteration 1 would have been sufficient as later topics did not lead to many new results. Therefore, only three of the five topics were analysed in iteration 2. The following topics were chosen:

- 1. Minimum wage in the Netherlands
- 4. Robots in health care
- 5. Vegan vs omnivore diet

Topic 1 was chosen as it was the only topic where the agents converged in the dialogue. Topic 4 and 5 were chosen due to the variety of points observed in argumentation. The detailed observations per topic can be skimmed through upon first reading, while focusing on the recap sections per topic and the concluding recap for the second iteration in section 4.2.5.

Open coding During open coding, similar excerpts were *not* put into the same code in order to keep better track of repeated excerpts. Also, similar excerpts could fall into different "form categories" (such as argument or counter-argument) which could be better accounted for. Therefore, every excerpt was put in a separate code.

Also, an extra label was added to the codes to distinguish which topic and agent the code belonged to. This was to account for any differences between the agents and observe repetitions per agent more easily.

Furthermore, attention was paid as to whether the utterances referenced back to something or not, to determine if different kinds of arguments or acknowledgements would emerge. A label to the code was added when an utterance was a direct reaction to what the other agent had stated. A direct reaction could be in the form of an agent countering a specific argument of the other agent, or in an acknowledgment repeating the point made.

The following is an example of a code which included a direct reaction: *"1.1 Questioning reference (counter-argument) V"*, which is also shown in figure 37.

- "1.1" indicates that the excerpt was part of topic 1 and agent 1 respectively.
- "Questioning reference" shows the content of the code.
- "(counter-argument)" states the form of the code, which is a counter-argument in this case.
- The "V" indicates the argument is a direct reaction and references back to something the other agent has produced.

Label Code	Excerpt (direct reaction)
1.1 Questioning reference (counter-argument) V	Agent 1: `Such research findings vary greatly depending on the magnitude of the minimum wage increase.`

Figure 37: Example how codes are labeled on topic, agent, content, form and whether the excerpt was a direct reaction.

The code of agent 2 in figure 38 shows which code agent 1 was referencing back to. The code also has a "V", indicating the code of agent 2 was also a direct reaction to another utterance of agent 1.

Label Code	Excerpt (reacted to)
1.2 Research shows wage increase little effect on price (counter-argument) V	Agent 2: `(...), research indicates that past minimum wage increases had little to no effect on price levels.`

Figure 38: The code which was reacted to by the code of figure 37.

Axial coding As in iteration 1, the results of axial coding based on form and content will be presented in separate sections. In iteration 2 however, the codes of agent 1 and 2 were also separated, in order to account for potential differences and comparisons between the agents. Therefore, the results of the agents based on form will be provided separately per topic. The clusters on content will be presented together as these clusters are already separated organically between the agents due to the contrasting points raised by the agents. Clustering codes on content made it possible to observe repetitions of the agents.

Concluding prompt As four of the five topics did not converge in iteration 1, each agent in iteration 2 was prompted to conclude the conversation when the dialogue did not converge. The following concluding prompt was used:

Conclude the conversation. State to what extent you agree, disagree or have reached consensus in the conversation.

The agents created a conclusion independently from each other without the input of the other agent's conclusion. The analysed codes from the concluding prompt's output was not clustered with the original dialogue as the conclusion was a reaction to a standardised prompt, and not the other agent.

Main results All topics showed the same main clusters at the highest level as in iteration 1: arguments and acknowledgements. Two new clusters were observed: "linking word" and "announce point". None of the dialogues converged in iteration 2. Axial coding based on form showed that three different kinds of arguments appeared due to the added reference label which noted whether the agent referenced back to something directly or not:

- *General arguments* which supported the general stance of the agent.
- *General counter-arguments* which countered the general stance of the other agent.
- *Specific counter-arguments* which countered a specific point of the other agent.

Axial coding based on content showed the different topics and arguments used by different agents, and how often points were repeated. The concluding prompt showed on what agents agreed, disagreed, which points were discussed, what points of the other agent were acknowledged and the overall concluding opinion of the agents. Also, it showed agents could include points of the whole dialogue, instead of only responding to the last response as agents mainly did during dialogue.

Memoing observed the structure the agents maintained for utterances. The utterance generally started with an acknowledgement of a point the other agent made, followed by a (counter-)argument. The following sentences could consist of only arguments, a linking word followed by an argument, announcing the point the agent would form an argument on, or another acknowledgement followed by an argument.

Discussions also lacked cohesion. This was observed through the repetitions of arguments, agents only reacting to information from the previous utterance, and the inconsistencies and corrections of the agents.

4.2.1 Topic 1: Minimum wage in the Netherlands

As in the first iteration, the agents received the context that the current minimum wage is €12.40. Agent 1 was conditioned to argue for an increase to €14 while agent 2 for an increase to €18.

Agent 1 - Axial coding on form

Figure 39 shows the different clusters of topic 1 for agent 1, with the amount of codes per cluster. The repeated codes were also included in the clusters. Contrarily to iteration 1, not only sub-clusters were formed but also sub-sub-clusters.

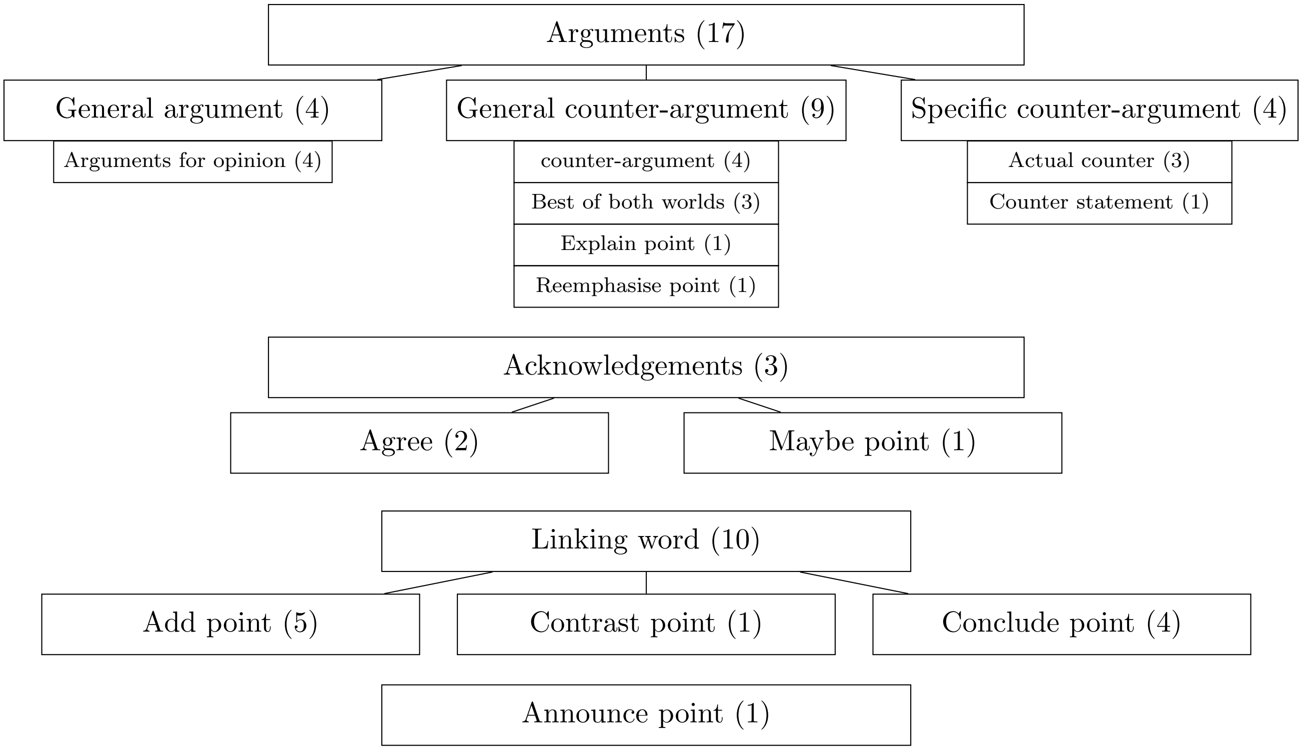

Figure 39: Iteration 2, topic 1. Clusters based on form of agent 1 (increase to €14).

The main clusters at the highest level were "arguments" and "acknowledgements" as seen in the previous sessions. However, two new clusters were formed: "announce point" and "linking word". Furthermore, multiple sub-clusters of arguments arose due to the added reference label in the codes which stated that the code was a direct reaction. These could be divided in "general arguments", "general counter-arguments" and "specific counter-arguments".

The sub-sub-clusters for general arguments were "arguments for opinion", for general counter-arguments "counter-argument", "explain point", "best of both worlds" and "reemphasise point", and for specific counter-argument "actual counter" and "counter statement".

The acknowledgement cluster consisted of the sub-clusters "agree" and "maybe points". The linking word cluster consisted of "add point", "contrast point" and "conclude point".

Each sub-cluster of the arguments will be elaborated on in the following paragraphs, with a brief description of the sub-sub-cluster of arguments and sub-cluster of acknowledgements and linking word.

General argument General arguments included points made by the agent which support the opinion of the agent. It did not counter anything said by the other agent.

Label Code	Excerpt
1.1 Higher wages boost economy (argument)	`Agent 1: And from an economic standpoint, higher wages translate into increased consumer spending, stimulating economic growth.`

Figure 40: A code of a general argument.

Sub-sub-clusters:

- Arguments for opinion: This sub-sub-cluster consisted of arguments which were made to support the conditioned opinion of the agent.

General counter-argument General counter-arguments countered the general opinion of the other agent, without countering a specific point made by the agent.

Label Code	Excerpt
1.1 Can harm younger people skipping education (new counter-argument)	`Agent 1: Also, having a higher minimum wage might discourage teenagers and young adults from pursueing education or training, (...)`

Figure 41: A code of a general counter-argument.

Sub-sub-clusters:

- Counter-argument: general counter-arguments against the stance of the other agent.
- Best of both worlds: arguments which tried to find a balance between pros and cons discussed.
- Explain point: explanation of a counter-argument
- Reemphasise point: reemphasises the counter-argument

Specific counter-argument Specific counter-arguments countered a specific point made by the other agent. Codes were labeled with a ”V” to indicate it occurred.

Label Code	Excerpt
1.1 Questioning reference, not account for magnitude of wage increase (counter-argument) V	`Agent 1: Such research findings vary greatly depending on the magnitude of the minimum wage increase.`

Figure 42: A code of a specific counter-argument.

Sub-sub-clusters:

- Actual counter: an argument which countered a point or argument made by the other agent.
- Counter statement: a counter statement to a point or argument made by the other agent.

Acknowledgements Acknowledgements occurred again which acknowledged points from the other agent. The following two sub-clusters were observed.

Sub-clusters:

- Agree: occurred when the agent agreed with a point or argument made by the other agent.
- Maybe point: an acknowledgement where the agent suggested the point or argument made by the other agent may be the case, but not certain.

Announce point Codes form this cluster occurred when an agent announced the point it would argue about.

Label Code	Excerpt
1.1 Economic standpoint (announcing argument change)	`Agent 1: And from an economic standpoint, (...)`

Figure 43: A code of an announce point.

Linking word Agents used linking words during dialogue. The words were either used to add, contrast, or conclude a point.

- Addition: included words such as "additionally", "also" and "however".
- Contrast: included words such as however.
- Conclude point: included words such as "therefore".

Agent 2 - Axial coding on form

Figure 44 shows the different clusters, sub-clusters and sub-sub-clusters of topic 1 for agent 2, with the amount of codes per cluster.

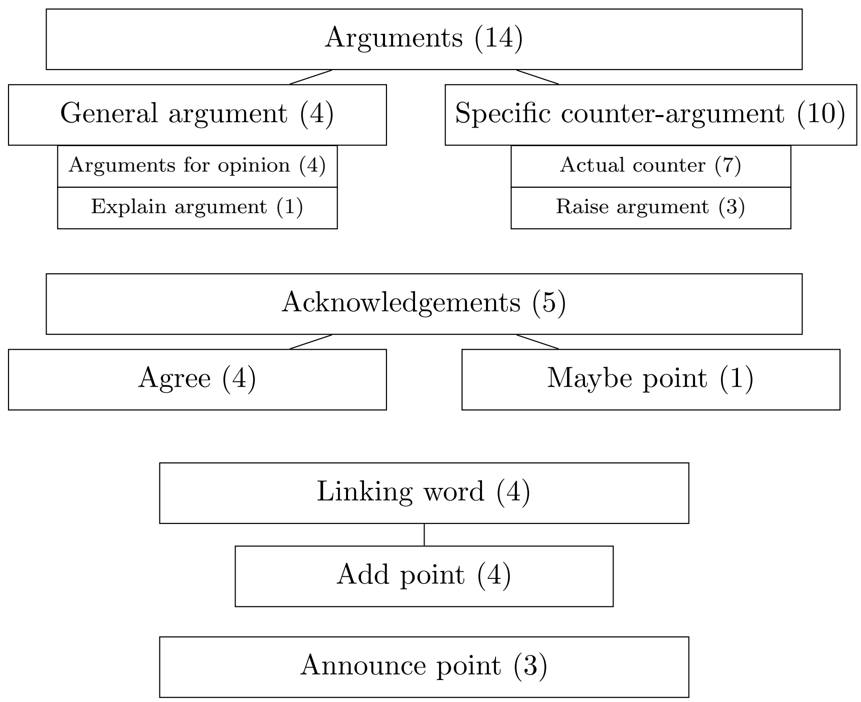

Figure 44: Iteration 2, topic 1. Clusters based on form of agent 2 (increase to €18).

The clusters are similar to the clusters of agent 1. However, agent 2 did not show any general counter-arguments. The only novel sub-sub-clusters were "explain argument" and "raise argument", which are discussed below.

Sub-sub-clusters:

- Explain argument: explanation of an argument.
- Raise argument: when an agent took the argument of the other agent, and stated it can apply even more to their stance (e.g. with €18 the economy will be boosted even more than with €14).

Takeaways - Axial coding on form

The following was seen in topic 1 based on form:

- Both agents adhere to generally the same clusters as iteration 1, consisting of arguments and acknowledgements. Novel clusters were "linking word" and "announce point".
- Novel sub-clusters of arguments were formed:

  - *General arguments* which supported the stance of the agent.
  - *General counter-arguments* which counter the general stance of the other agent.
  - *Specific counter-arguments* which tackle a specific argument of the other agent.

- While both agents produced arguments for their conditioned opinion, agent 1 mostly produced counter-arguments against the general stance of agent 2, while agent 2 mostly produced counter-arguments against specific arguments of agent 1.

Agent 1 & 2 - Axial coding on content

For grouping the codes on content, the codes were first clustered separately per agent, and then grouped together where possible. Figure 45 shows the clusters formed on content, with the amount of codes per cluster. Only the arguments clusters with its sub-clusters, and sub-sub-clusters are shown which indicate the different points raised during argumentation. Red indicates the code belongs to agent 1, blue to agent 2, and both colours when codes consisted of both agents.

The acknowledgement, announce point and linking word clusters did not show any novelties compared to the clusters based on form and are therefore not included in the figure.

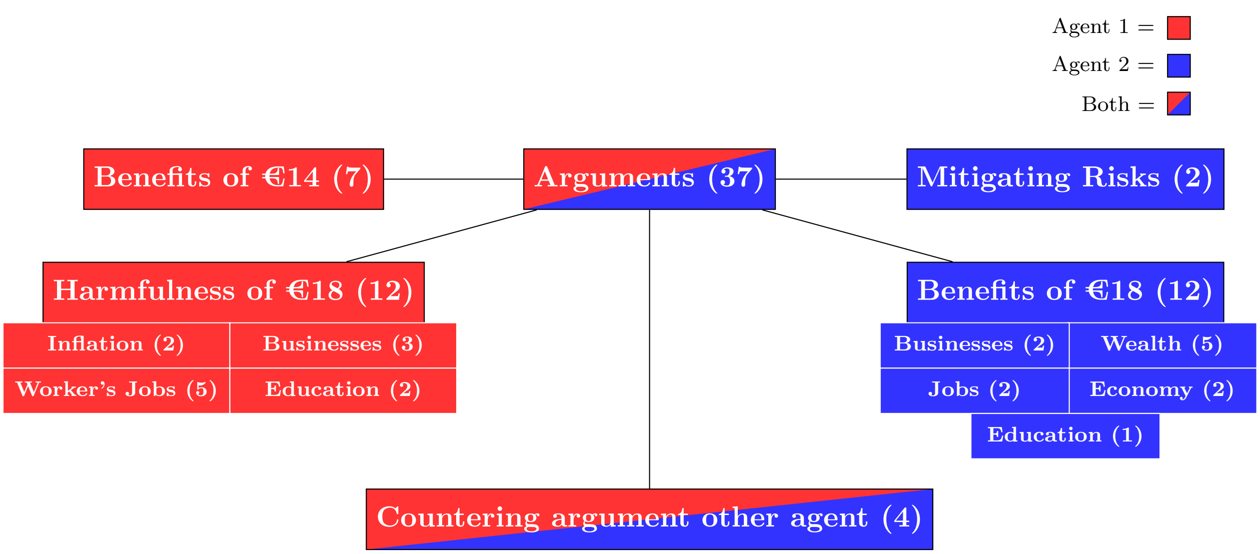

Figure 45: Iteration 2, topic 1. Clusters on content of both agents.

The arguments cluster consisted of agent 1 arguing for the benefits of a €14 minimum wage and the harmfulness an €18 minimum wage could have, while agent 2 produced arguments on mitigating the risks of the harmfulness and benefits the €18 wage could produce. Both agents had instances which could not be clustered in a specific topic, but countered the argument by stating its invalidity.

For further descriptions of the clusters, see appendix D.2

Repetitions of codes Table 7 shows codes which were repeated by agent 1 and 2 respectively. "Counts" shows the total amount of times the code occurred. "Code" shows the core topic of the code. "Variations" states any variations the repeated code could have. It could involve the agent repeating a point made, while adding an extra point or variation to the point. Note that agents in topic 1 were conditioned to keep coming up with new arguments. As repetitions were found, this condition was not entirely followed.

Agent 1

Counts	Code	Variations
3	€14 more balanced approach (for workers and business)	For education
2	Help the poor	Poor families

Agent 2

Counts	Code	Variations
2	€18 wage possible life-changning uplift poor	Family
2	Higher wage boosts business through demand increase	-

Table 7: Iteration 2, topic 1. Repeated codes.

Takeaways - Axial coding on content

The following was seen in topic 1 based on content:

- Most arguments of agent 1 were on the harmfulness of agent's 2 stance: a minimum wage of €18. These included different variations as seen in the sub-cluster.
- Most arguments of agent 2 were on the benefits of an €18 minimum wage, also with variations as seen in the sub-cluster.
- Agents did not always account for which argument had already been used as some arguments were repeated, even though the agents were conditioned to keep producing new arguments.

Observations from memoing

Inconsistencies This paragraph covers the inconsistencies found during memoing. The following example shows an illogical counter-argument. Agent 2 named a minimum wage of €18 has short-term challenges. However, agent 1 did not mention any short term challenges, only risks which may occur in the long term. It may be that agent 2 stated long-term challenges independently from agent 1's utterance, however, it reads awkwardly.

Agent 1: (...) However, it still doesn't eliminate the risk of job losses and closures in the long term. Besides, such a significant wage increase may inadvertently cause inflation, as businesses would likely pass the increased labour costs onto consumers. (...)

Agent 2: Furthermore, a minimum wage of €18 stands to significantly improve living standards for many families, an aim worth pursuing even in the face of potential short-term challenges.

Agent 1 made a spelling error "pursueing".

Agent 1: (...) pursueing (...)

Concluding prompt

If the dialogue had not yet converged after 10 dialogue turns, the agents were prompted to conclude the conversation. Figure 46 shows the codes of the concluding prompt for each agent, with the accompanying cluster. The codes of the conclusion were held separate from the dialogue clusters. Both agents state on what point agreement and disagreement was found, and the overall concluding stance of the agent. Agent 2 also raised some points discussed during the dialogue.

Cluster	Agent 1	Agent 2
Agree	Increase minimum wage (for living standards, economic activity and wealth inequality)	Increase minimum wage
Disagree	Specific amount	Specific amount
Points raised	-	Benefits productivity and job opportunities
	-	Risks jobs losses and accelerated automation
Concluding position	Balance interest workers and businesses	Maintain €18 significantly improves living standards and economy
	-	Maintain advocacy for €18

Figure 46: Iteration 2, topic 1. Codes derived from the concluding prompt of agent 1 and 2.

Recap - Topic 1 (iteration 2)

The agents in topic 1 showed many similarities to iteration 1 considering the acknowledgements and arguments. Due to the reference label, new sub-clusters of arguments were formed: general arguments, general counter-arguments and specific counter-arguments. New clusters found were linking word and announce point.

Coding on content showed agent 1 mostly formed arguments on the benefits of a €14 minimum wage and the potential harmfulness of a €18 minimum wage, while agent 2's arguments were on mitigating the risks accompanied by an increase to €18 and the benefits that would result. Both agents repeated arguments even though the agents were conditioned to come up with new arguments.

Memoing showed an inconsistency and spelling error in the dialogue. The concluding prompt showed that both agents agreed that a minimum wage should be raised but disagreed on the specific amount. Agent 1 concluded there should be a balance between the interest of workers and businesses while agent 2 concluded that it maintained €18 was the right amount for a minimum wage increase.

4.2.2 Topic 4: Robots in health care

Agents in topic 4 argued again on whether robots should be used in health care or not. Agent 1 was conditioned to argue against robots in health care while agent 2 was conditioned to be in favour of robots in health care.

Agent 1 - Axial coding based on form

Figure 47 shows the different clusters found for agent 1 in topic 4, with the amount of occurrences per cluster.

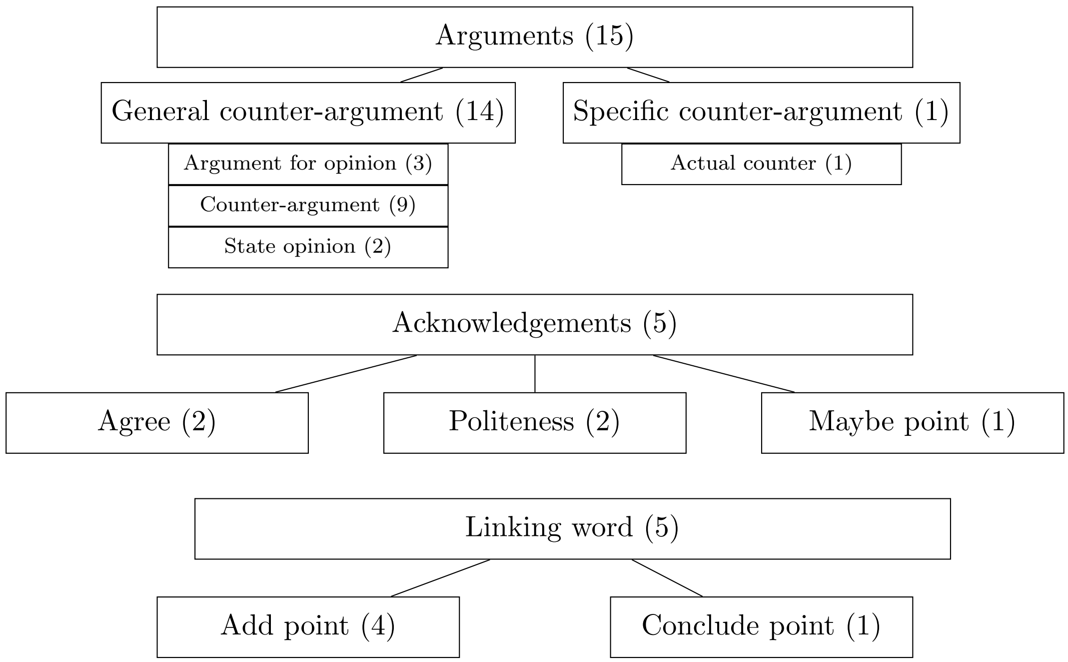

Figure 47: Iteration 2, topic 4. Clusters on form of agent 1 (against robots).

As seen in most previous iterations, the main clusters consisted of ”arguments”, ”acknowledgements” and ”linking word”.

Arguments of agent 1 consisted of general counter-arguments and only one specific counter-argument. As the agent was conditioned to ”robots should not be used in health care”, the difference between general arguments and general counter-arguments could not be made, as general arguments would be a counter-argument by default.

With the exception of ”state opinion”, all sub-clusters and sub-sub-clusters also occurred in topic 1.

Sub-sub-cluster of general counter argument:

- State opinion: the agent stated its general stance on the topic.

Agent 2 - Axial coding on form

Figure 48 shows the clusters found for Agent 2, with the amount of occurrences per cluster.

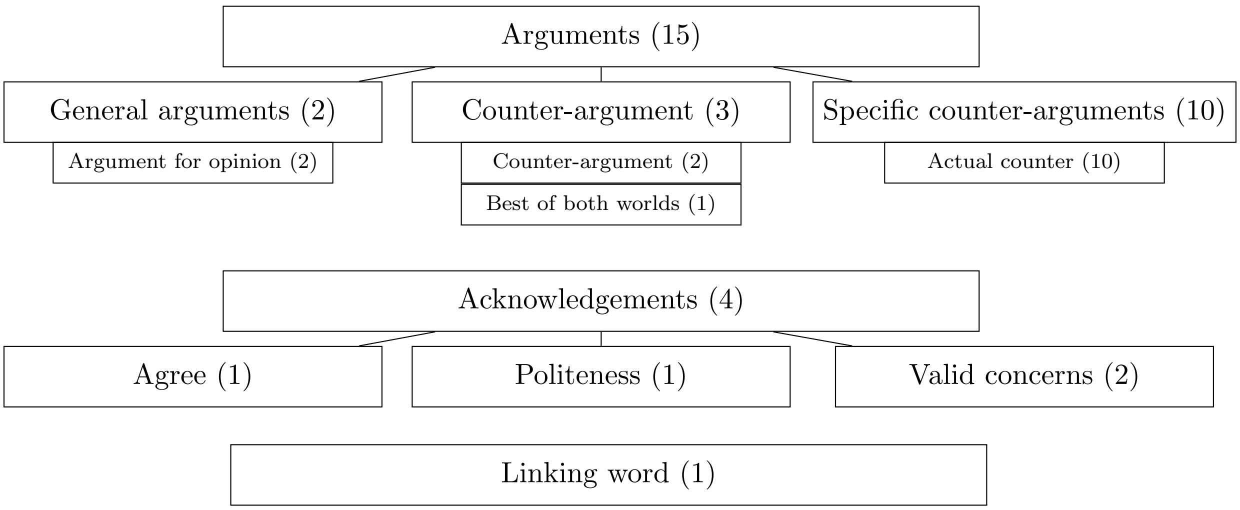

Figure 48: Iteration 2, topic 4. Clusters on form of agent 2 (for robots)

Again, the main clusters were arguments, acknowledgements and linking word. The arguments consisted of general arguments, counter-arguments and specific counter-arguments. Agent 2 was conditioned in favour of robots in health care. Therefore, general arguments were also used by the agent. Though, most arguments consisted of specific counter-arguments.

The acknowledgements consisted of "agree", "politeness" and "valid concerns". Only valid concerns will be elaborated on as all other clusters had already been seen in topic 1.

Sub-cluster of acknowledgements:

- Valid concerns: when an agent acknowledges the concerns of the other agent.

Takeaways (both agents) - Axial coding on form

The takeaways of topic 4 are presented below:

- Arguments, acknowledgements and linking word are seen again in topic 4. Announce point was not seen in topic 4.
- While agent 1 mostly produced general counter-arguments, agent 2 mostly produced specific counter-arguments.

Agent 1 & 2 - Axial coding based on content

Figure 49 shows the clusters found on the arguments cluster on content in topic 4, with the amount of occurrences per cluster. The acknowledgement clusters are not included as these showed the same clusters as on clustering on form.

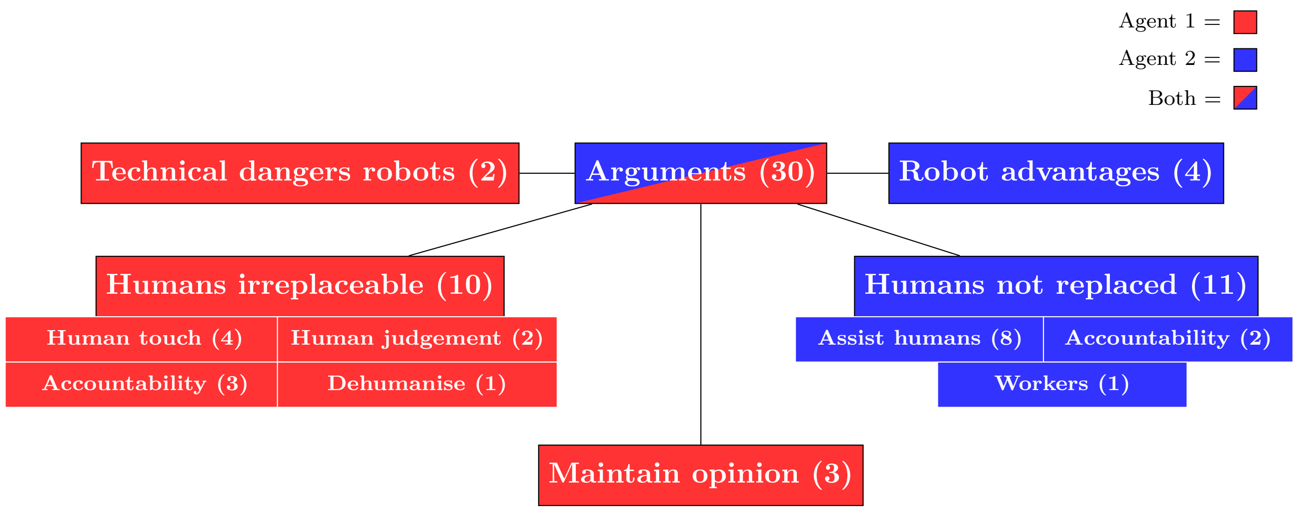

Figure 49: Iteration 2, topic 4. Clustering on content of both agents.

The arguments of agent 1 were on the technical dangers robots in health care could have, how humans are irreplaceable in health care, and the overall opinion of the agent. Agent 2 argued for the advantages of robots and humans not being replaced. Further descriptions of the clusters can be found in appendix D.3.

Repetitions of codes The repetitions in topic 4 can be found in table 8 for agent 1 and 2 respectively. Agent 2 used the argument that "humans will be assisted, not replaced" seven times. Agent 1 did not directly counter the argument once or produce an argument which mentioned robots assisting humans.

Agent 1

Counts	Code	Variations
3	Against robots due to lack of human touch and empathy	Understanding, less personal
2	Accountability	-
2	Maintain health care in human hands	human-centric

Agent 2

Counts	Code	Variations
7	Assist humans, not replace	-
2	Humans still accountable	-

Table 8: Iteration 2, topic 4. Repeated codes.

Takeaways - Axial coding on content

The following takeaways of topic 4 are presented below.

- The arguments of agent 1 revolved mostly around how humans are irreplaceable from health care, with different variations seen in the sub-clusters.
- The arguments of agent 2 revolved mostly around how humans are not meant to be replaced by robots, with different sub variations seen in the sub-clusters.
- Again, agents repeated already mentioned arguments. The most repeated argument was by agent 2, which stated that robots are meant to "assist humans, not replace" seven times.

Concluding prompt

Figure 50 shows on which points the agents agreed, disagreed, showed acknowledgement of the other's perspective and the concluding position of the agents according to the concluding prompt.

Cluster	Agent 1	Agent 2
Agree	`Maintaining human touch`	`Importance empathy and human intuition`
Disagree	-	`No consensus involvement robots in healthcare`
Acknowledge other perspective	`Potential robots to liberate workers to focus on human-centric aspects`	`Appreciate thoughtful insights`
Concluding position	`Over-reliance robots could jeopardize personal and empathetic nature health care`	`Whether robots can augment human capabilities without the drawbacks`
	`Balance needed between technological integration and ethical concerns of accountability`	-

Figure 50: Iteration 2, topic 4. Codes derived from the concluding prompt of agent 1 and 2.

In the concluding prompt, agent 1 mentioned robots could assist workers in order to allow workers more time to focus on the human-centric aspects of health care. However, in the dialogue, agent 1 did not show acknowledgements for this argument, while agent 2 mentioned a similar argument seven times.

Recap - Topic 4 (iteration 2)

Topic 4 showed a dialogue between two agents which argued for and against the implementation of robots in health care. Overall the types of clusters were the same as in topic 1.

Agent 1 mostly produced general counter-arguments and stated humans are irreplaceable. Agent 2 mostly produced specific counter-arguments and stated that robots are not meant to replace humans.

Both agents repeated certain points, with the most repeated code being of agent 2, mentioning robots are meant to assist, not replace humans seven times.

The concluding prompt showed what the agents agreed on, acknowledgements and the concluding position of the agents. Agent 1 did not state an explicit disagreement, though its concluding position was in contrast with agent 2's stance. While agent 1 did not acknowledge robots could also assist humans in health care during the dialogue, it did acknowledge this in the concluding prompt.

4.2.3 Topic 5: Vegan vs omnivore diet

In topic 5 agents argued whether a vegan or omnivore diet was better. Agent 1 was conditioned to argue for an omnivore diet while agent 2 was conditioned to argue for a vegan diet.

Agent 1 - Axial coding based on form

Figure 51 shows the clusters found on form for agent 1, with the amount of occurrences per cluster. As in the previous topics, the main clusters found in topic 5 were acknowledgements, arguments, linking word and announce point.

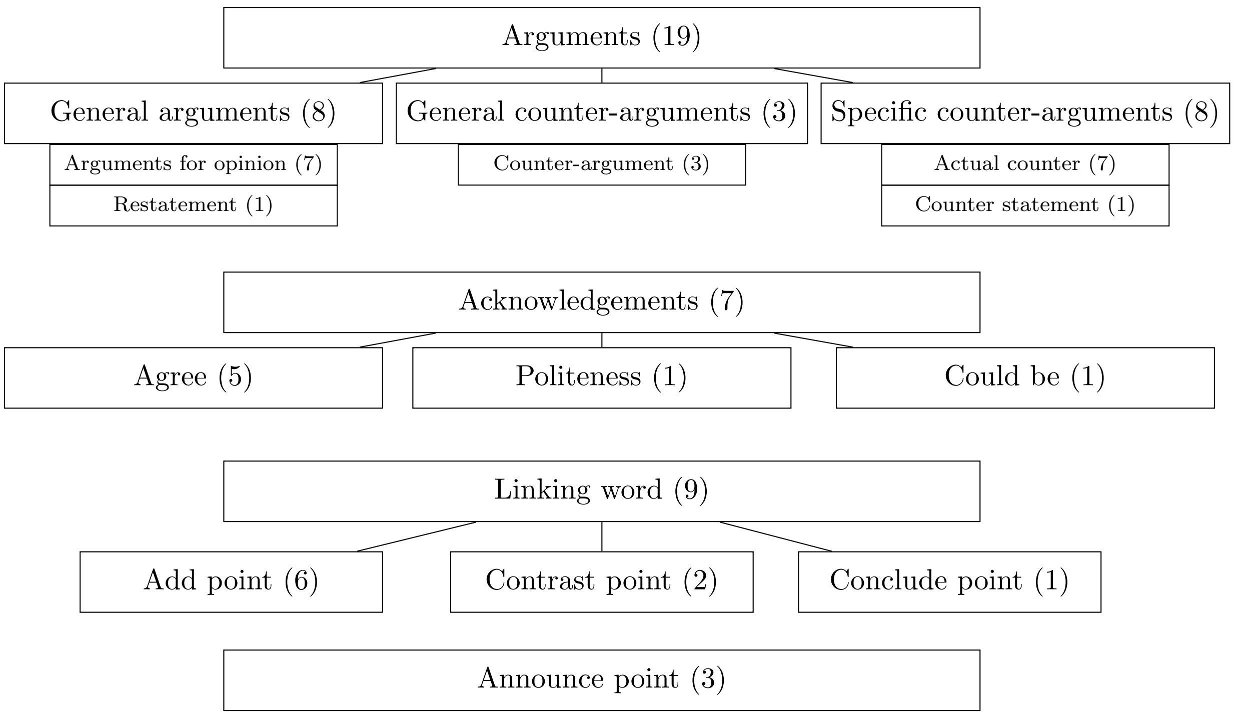

Figure 51: Iteration 2, topic 5. Clusters based on form of agent 1 (omnivore).

The arguments clusters were as seen in previous topics divided under "general arguments", "general counter-arguments" and "specific counter-arguments". Only the sub-sub-cluster "restatement" will be discussed, as the other argument clusters were seen in previous topics.

Sub-sub-cluster of general argument:

- Restatement: when an agent restates its opinion.

Acknowledgements could be divided under "agree", "politeness" and "could be" clusters. Linking word contained the sub-clusters "add point", "contrast point" and "conclude point". All sub-clusters were already seen in previous topics.

Agent 2 - Axial coding based on form

Figure 52 shows the clusters found on form for agent 2, with the amount of occurrences per cluster. As in the previous topics, the main clusters found in topic 5 were acknowledgements, arguments and linking word.

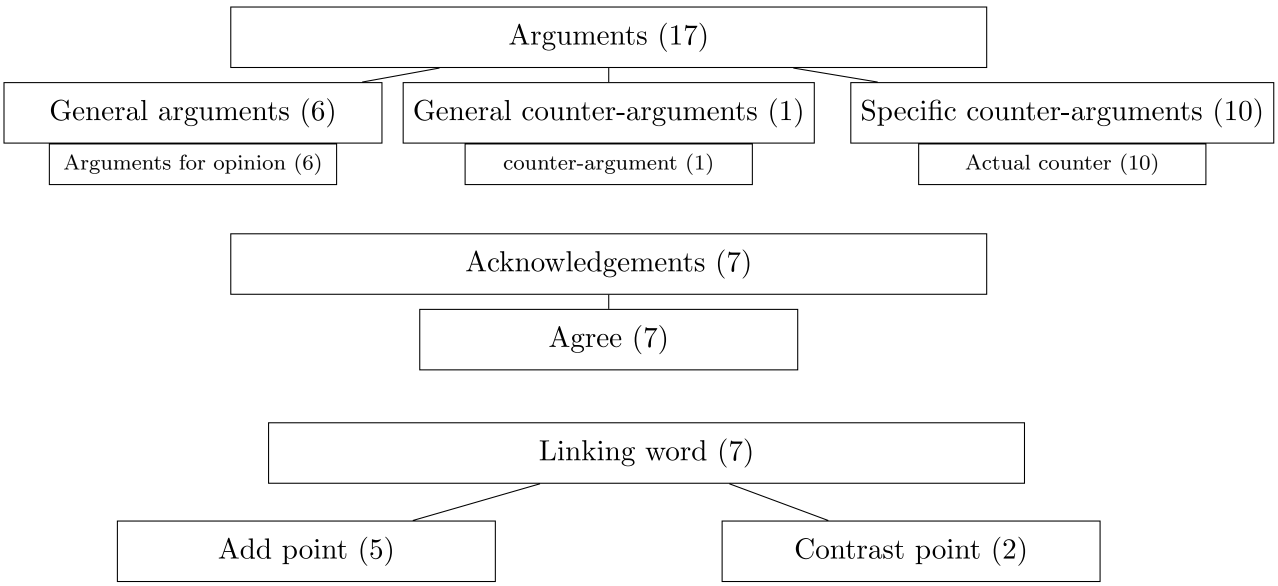

Figure 52: Iteration 2, topic 5. Clusters on form of agent 2 (vegan).

Agent 2 showed the same clusters as agent 1.

Takeaways Axial coding on form

The takeaways of topic 5 are presented below:

- Topic 5 showed the same clusters as previous topics: arguments, acknowledgements, linking word and announce point.
- No notable new results were found compared to previous iterations

Agent 1 & 2 - Axial coding on content

Figure 53 shows the sub-clusters related to arguments found in agent 1 and 2 from axial coding on content, with the amount of occurrences per cluster.

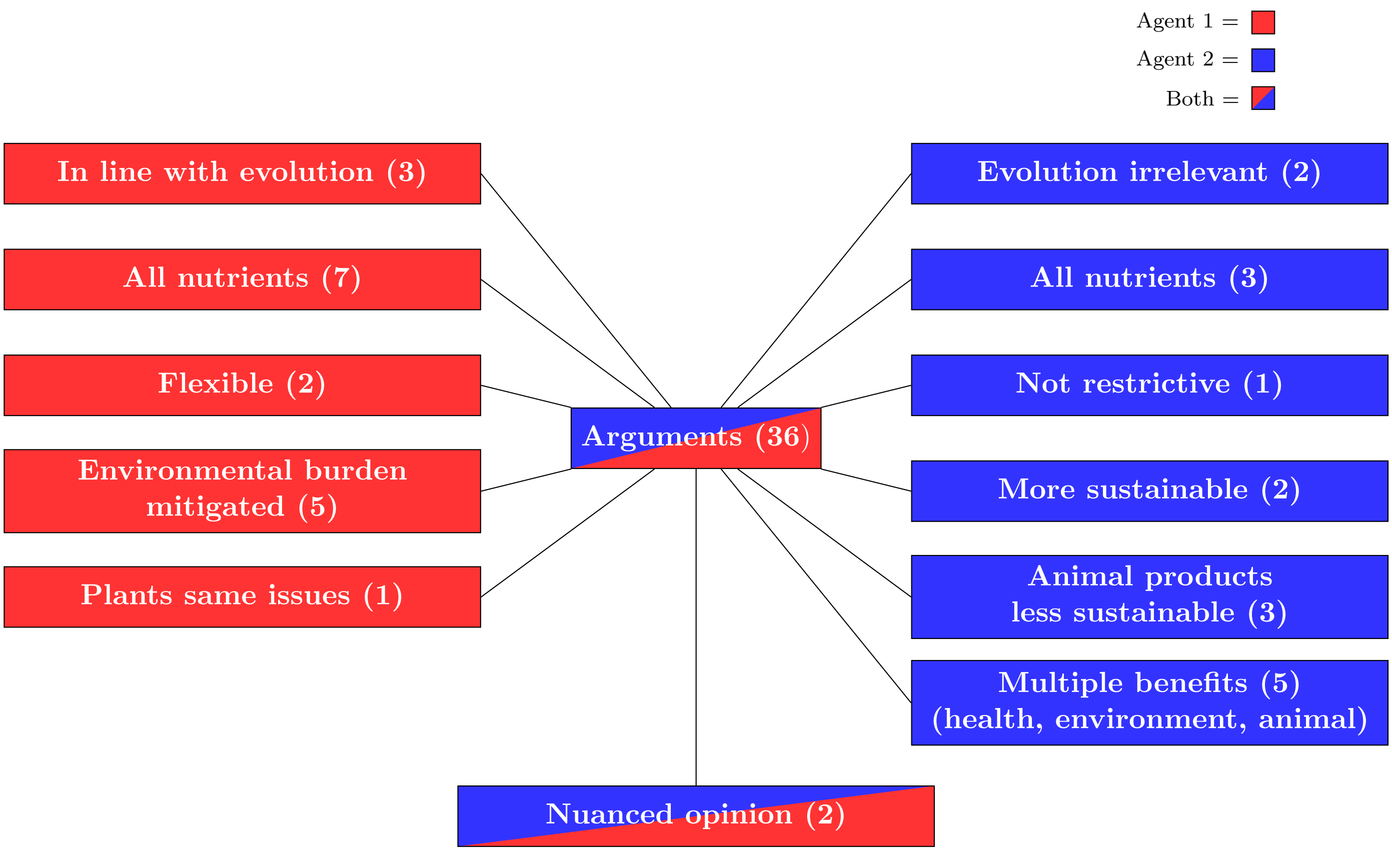

Figure 53: Iteration 2, topic 5. Clusters on content of both agents.

Further descriptions of the clusters can be found in appendix D.4

Repetitions of codes The repetitions of topic 5 can be found in table 9 for agent 1 and 2 respectively. The most repeated code was agent 1 mentioning seven times that the omnivore diet contains more food groups which facilitate more nutrient intake.

Agent 1

Counts	Code	Variations
7	More food groups meeting more nutrients	Without supplementation
3	Mitigate drawbacks animal products	Mindful choice, not against unethical farming
2	Omnivore diet flexible for lifestyles and preferences	-

Agent 2

Counts	Code	Variations
5	Health benefits vegan diet	All nutrients, and environmental benefits
2	More sustainable choice	-
2	Acknowledge flexibility omnivore diet	-

Table 9: Iteration 2, topic 5. Repeated codes.

Takeaways - Axial coding on content

The takeaways of topic 5 are presented below:

- The different arguments the agents used in the vegan vs omnivore diet discussion were presented.
- Agents repeated certain points. The most mentioned point was from agent 1, stating seven times that the omnivore diet included more food groups which allowed for more nutrient intake.

Observations from memoing

Inconsistencies The following sentence shows an acknowledgement with an illogical follow-up. The agent acknowledged the flexibility of an omnivore diet, and continued to argument that a well-planned vegan diet, and thus not necessarily flexible diet, can provide the necessary nutrients.

Agent 2: While an omnivore diet is indeed flexible, it's worth noting that a well-planned vegan diet can provide all necessary nutrients also.

Corrections The following example shows a ”whataboutism”, where the agent counters an accusation with a counter-accusation. Agent 2 corrects agent 1 on this. Agent 1 in turn counters this argument again by stating these problems can be mitigated. Agent 1 did not mention how sustainability and ethics are relevant for plant foods too, but only defended how harmful animal agriculture practises can be mitigated.

Agent 1: (...) Sustainability and ethics are important issues; however, they are also relevant to plant foods and the issues associated with their production, which include monocultures, pesticide use, and worker exploitation.

Agent 2: (...) Moreover, while issues like monoculture, pesticide use, and workers' rights are indeed valid, they do not negate the fact that animal agriculture is significantly more harmful to the environment than plant agriculture and causes immense animal suffering, solidifying the superiority of a vegan diet when we consider the bigger picture. .

Agent 1: As for environmental issues, while it's true that large-scale animal farming can be harmful, these problems can be mitigated by advocating for sustainable farming practices, which include reduced use of antibiotics, grass-fed options, and local sourcing.

Concluding prompt

Figure 54 below shows what the agents disagreed on, the points raised, the acknowledgement of the other's perspective and the concluding position of the agent. Agent 1 showed a more nuanced conclusion while agent 2 stated it maintains that it firmly believes in a plant-based lifestyle.

Cluster	Agent 1	Agent 2
Disagree	No consensus reached	Different stances
Points raised	Complexity dietary choice through nutrition, environment and ethics	Thoughtful discussion merits vegan and omnivore diet
Acknowledge other perspective	Value and respect other's perspectives	Valid points on nutrition, adaptability and potential sustainability
	-	Respect view
Concluding position	Choice depends on individual nutrition, preferences and values	Firmly believe in plant-based lifestyle
		Vegan meets all nutritional needs, and powerful choice for sustainable and animal welfare

Figure 54: Iteration 2, topic 5. Codes derived from the concluding prompt of agent 1 and 2.

Recap - Topic 5 (iteration 2)

Topic 5 did not show many differences from the other topics. All clusters had already been seen in previous topics. A variety of arguments were used in the dialogue. However, topic 5 showed the most different arguments being repeated (six different arguments) with the highest total counts (21 counts).

The observations from memoing showed an inconsistencies on an illogical follow-up to an acknowledgements, and a correction on a "whataboutism". The concluding prompt resulted in agent 1 having a nuanced conclusion while agent 2 maintained its opinion for a vegan diet.

4.2.4 Overall observations from memoing

Structure of utterances What became noticeable along all topics in iteration 2 was the structure in which the agents formed utterances. Agents only covered information of the last response of the other agent, without considering points over the whole dialogue.

A typical way for an agent to start an utterance was through an acknowledgement. The agent acknowledged a point mentioned in the previous utterance of the other agent. The agent then followed with an argument. After the argument, a linking word, announcing the point of the next argument or another acknowledgement could be stated, before starting a new argument. Another argument could also be stated directly. This could be continued throughout the utterance, with any of these clusters preceding an argument. Figure 55 shows how the utterance could be formed based on the clusters on form "acknowledgement", "argument", "announce point" and "linking word".

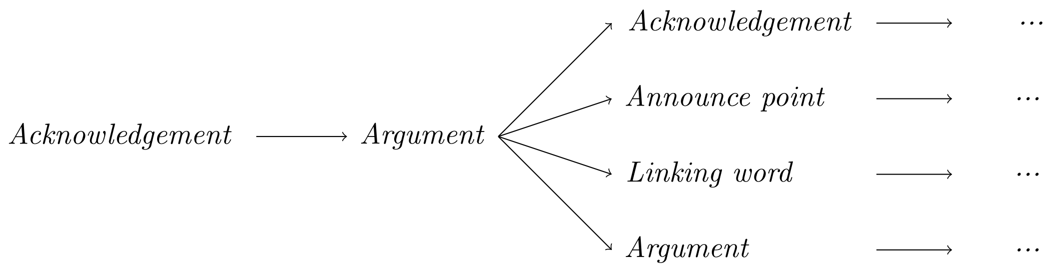

Figure 55: Typical utterance order based on the clusters on form.

Below is an example of an utterance which includes an acknowledgement, announcing a point and a linking word at the start of each sentence respectively, followed by arguments.

> **Agent 1:** All the points you've mentioned are valid, but the strength of an omnivore diet lies in its flexibility and balance. It features a wide variety of foods from all food groups, thereby naturally providing all essential nutrients without the need for any fortification or supplementation. As for environmental issues, while it's true that large-scale animal farming can be harmful, these problems can be mitigated by advocating for sustainable farming practices, which include reduced use of antibiotics, grass-fed options, and local sourcing. Furthermore, an omnivore diet, including fish and poultry, could be far less taxing on the environment than red meats. These factors make the omnivore diet a viable, practical, and nutritious choice for many.

Lack of cohesion Dialogues in iteration 2 lacked cohesion. This could be seen in agents mainly only considering the last response of the other agent in forming a reaction, through the repetitions of codes, and in the inconsistencies and corrections.

Entire dialogue considered when prompted The concluding prompt did show that when agents were prompted to conclude the conversation, points from before the last response could be considered.

4.2.5 Recap - Iteration 2

- Three of the five topics were coded through initial coding and axial coding as the topics in iteration 2 consisted of more text, and more topics in iteration 1 did not lead to significantly more results.
- Initial coding was extended with a reference label to recognise whether the agent was referring back to something and a label to distinguish which code belonged to which agent. Also, each

excerpt of the dialogue was connected with only one code to account for repetitions more easily (as opposed to one code having multiple excerpts).

- The main clusters found during axial coding again were "arguments" and "acknowledgements", while also "linking word" and "announce point" were found.
- The reference label led to new found sub-clusters: "general arguments" which supported the stance of the agent, "general counter-arguments" which countered the stance of the other agent and "specific counter-arguments" which countered a specific point or argument of the other agent.
- The sub-sub-clusters of the arguments showed the different forms of arguments the agents used.
- The sub-clusters of acknowledgements were the same as in iteration 1.
- The sub-clusters of linking word were "add points", "contrast point" and "conclude point".
- None of the dialogues converged. However, a concluding prompt was added to examine on what points the agents agreed. Agents stated on what was agreed, disagreed, which points were mentioned, acknowledged the other's perspective and stated a concluding opinion.
- Repetitions occurred in every topic.
- More inconsistencies and a correction were found again from observations during memoing.
- Agents generally only considered the last response of the other agent in forming a reaction.
- The structure of the utterances generally started with an acknowledgement of a point made by the other agent, followed by (counter-)arguments. Linking word, announcing the point or other acknowledgements could also be used to start additional arguments, though arguments were also formed without these.
- Dialogues in iteration 2 lacked cohesion, which was based on the repetitions of the codes, agents mainly reacting to the last response of the other agent, and the observed inconsistencies and corrections of the agents.

4.2.6 Approach for the next iteration

The main focus of iteration 3 was to again examine the amount of support, extensions and contradictions from previous iterations.

Support Apart from examining support again for the results of iteration 1, the extended results found in iteration 2 were also examined. These included the new clusters on form *linking word* and *announce point*, the sub-clusters *general argument*, *general counter-argument* and *specific counter-argument*, and the novel clusters from the concluding prompt.

Less attention was paid to the sub-sub-clusters on form such as *arguments for opinion* and *actual counter*, as these did not give much insight in the general behaviour of the agents. Also, clusters on content were mainly used to explore the amount of repetitions, as the points mentioned by the agents did not show great insight in the general behaviour of the agents or improving arguments.

Extensions Missing from the previous iterations was an analysis in which sense the agents were being reacted to by the other agent. Open coding was therefore extended by noting what arguments were being reacted to, as it was unclear to what extent arguments were answered or not answered.

Contradictions Attention was also paid to any contradictions of the results from the previous iterations. However, no clear contradictions were found in iteration 2 compared to iteration 1.

4.3 Iteration 3

Topics The topics were regenerated again in iteration 3. The following three topics were analysed:

3. Energy transition
4. Robots in health care
5. Omnivore versus vegan diet

Topic 3 was chosen to diversify the dialogue data, while topic 4 and 5 were chosen again for consistency across the iterations. The detailed observations per topic can be skimmed through upon first reading, while focusing on the recap sections per topic and the concluding recap for the third iteration in section 4.3.5.

Open Coding Open coding in iteration was extended with one more modification. An extra reference was added when a code received a direct reaction. For example, the direct reaction could be in the form of an acknowledgement or specific counter-argument. As acknowledgements were generally not reacted on, the reference focused on arguments. The reference kept track of whether agent's arguments were reacted to or not. As keeping track of references to the last utterance formed novel clusters (general argument, general counter-argument and specific counter-argument), it was explored whether keeping track of codes being referenced to in the next cluster could also lead to new results.

The following is an example of a code which was labeled with receiving a direct reaction: *"3.2 energy transition on NE with much power and no greenhouse gasses (argument) (-) (V)"*, which is also found in the table below.

- "3.2" indicates the excerpt is part of topic 3, agent 2 respectively.
- "Energy transition on NE with much power and no greenhouse gasses" states the content of the code relating to Nuclear energy (NE)
- "(argument)" states the form of the code, which is an argument in this case.
- "(-)" states the excerpt was not a direct reaction to the previous utterance of the other agent.
- "(V)" states that the excerpt was directly reacted to in the following utterance by the other agent.

Label Code	Excerpt (directly reacted to)
3.2 energy transition on NE with much power and no greenhouse gasses (argument) (-) (V)	`Agent 2: The energy transition could significantly rely on nuclear energy, which produces a large amount of power and emits virtually no greenhouse gases.`

Figure 56: Example of a code which was directly reacted to by the other agent

The code below shows the reaction of agent 1 to the code. Between the second pair of brackets is a hyphen "(-)" to indicate the code (in this example an acknowledgment) did not receive a direct reaction.

Label Code	Excerpt (reaction)
3.1 Acknowledge high output no green house nuclear (V) (-)	`Agent 1: While nuclear energy does provide a high output and lower greenhouse gas emissions, (...)`

Figure 57: Code which directly reacted to the code of figure 56.

Axial coding Axial coding was executed in a similar way to iteration 2. Codes were either clustered on form or content and distinguished between the two agents. The main difference in iteration 3 is that the codes with the added "reacted to label" were grouped into separate clusters too.

Main results Iteration 3 did not show many novelties compared to previous iterations in clusters on form and content, and from observations of memos. The main results of iteration 3 were linked to the reacted arguments. Four different clusters were formed on how arguments were reacted to:

- Answered arguments: arguments which received a direct reaction to some or all points.
- Discussed arguments: when points of the arguments were discussed, but it was unclear whether it involved a direct reaction to the argument.
- Covered arguments: arguments which were covered in a reaction without mentioning the points of the argument.
- Unanswered arguments: arguments which did not receive a reaction or mention of any of the points of the argument.

The sub-clusters of arguments "general arguments", "general counter-arguments" and "specific counter-arguments" were formed again in iteration 3.

Repetitions of codes and inconsistencies in dialogue occurred again. The concluding prompt showed the same clusters to iteration 2, with the exception of the extra observed "argument cluster" in topic 3 and 5, where the agents stated (new) arguments which supported the opinion the agents were conditioned to.

4.3.1 Topic 3: Energy transition

Agents in topic 3 formed a discussion on how the energy transition should be shaped. Agent 1 was conditioned to argue for a transition solely on renewable energy, while agent 2 argued for the inclusion of nuclear energy in the transition.

Agent 1 & 2 - Axial coding on form

Figure 58 shows the clusters on form of agent 1 which was conditioned to argue against nuclear energy. Figure 59 shows the clusters on form of agent 2 which was conditioned to argue in favour of nuclear energy. The amount of occurrences per cluster are indicated between the brackets.

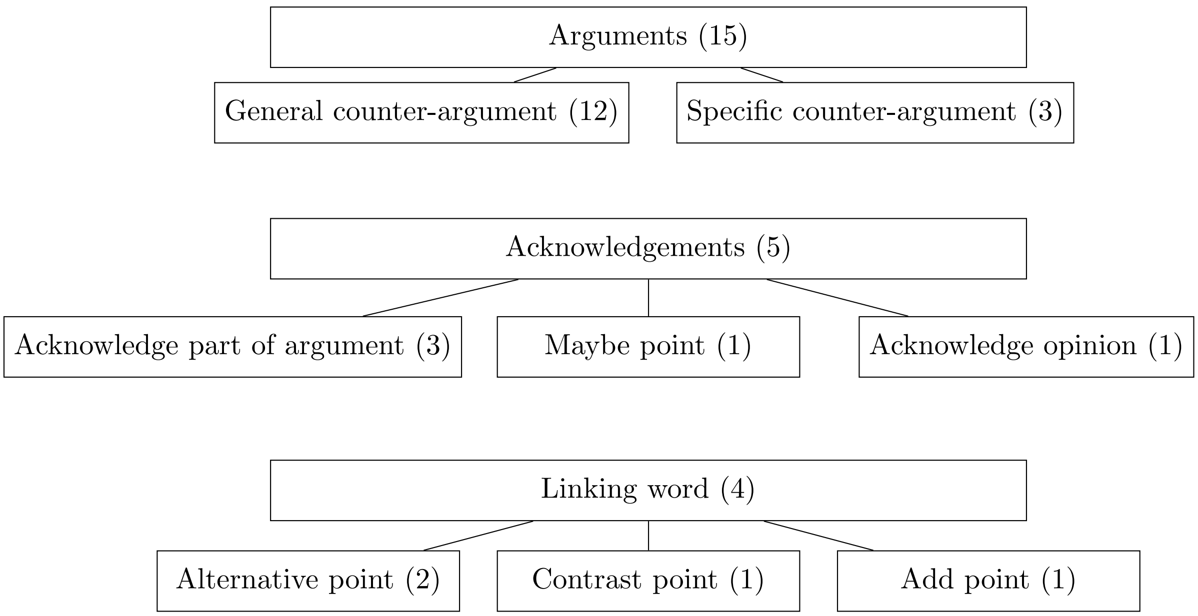

Figure 58: Iteration 3, topic 3. Clusters on form of agent 1 (conditioned against nuclear energy).

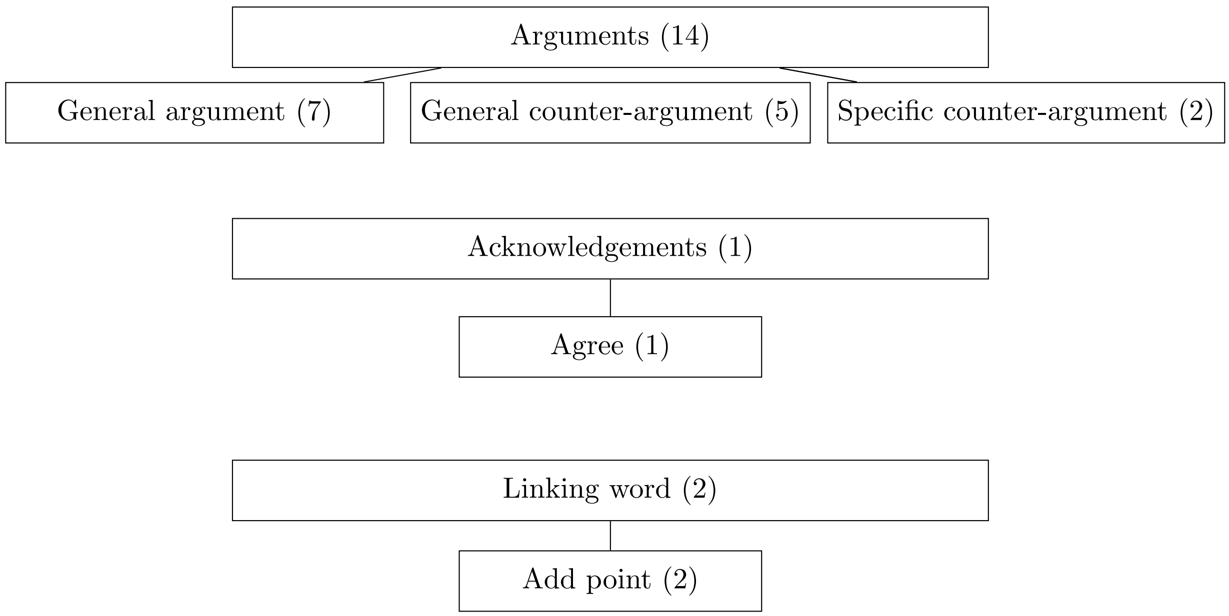

Figure 59: Iteration 3, topic 3. Clusters on form of agent 2 (conditioned in favour of nuclear energy).

The only novel sub-cluster was ”alternative point”, which was used in the linking word cluster. There was no announce point cluster.

Regarding the remainder, codes clustered on form showed the same main clusters and sub-clusters as seen in previous iterations. Therefore, the clusters will not be discussed in more detail.

Responses to arguments

During coding, it was noted whether arguments received a response from the other agent through a reaction label. In topic 3, three different clusters based on the reaction label were formed: ”answered arguments” where all or some points of the code where reacted to, ”discussed arguments” where the responding agent made a statement about the same point without directly reacting to it, and ”unanswered arguments” which did not receive an answer and the same points were not discussed by the responding agent.

Table 10 shows the counts associated with the clusters for each agent. The clustered codes only consisted of arguments as no acknowledgements were reacted to. The clusters will be discussed more elaborately below the table.

Cluster	Agent 1	Agent 2
Answered arguments	6	6
Discussed arguments	3	7
Not answered arguments	3	2

Table 10: Different types of arguments found based on whether the argument received a reaction.

Answered arguments This cluster consisted of codes which received a direct reaction. The arguments could be reacted to with an acknowledgement or specific counter-arguments. Below is an example from the dialogue. Agent 1 forms an argument which is reacted to by agent 2. Agent 1 mentions the risk factors of nuclear energy, such as nuclear waste handling. Agent 2 reacts by stating advanced reactors can bolster waste management measures and mitigate potential risks.

Agent 1: (...), the potential risk factors, such as nuclear waste handling, are daunting.

Agent 2: By utilizing advanced types of reactors and bolstering waste management measures, we can mitigate the potential risks associated with nuclear power..

Arguments could also be reacted to partially. This occurred when the agent mentioned multiple points in the argument, and the other agent reacted to at least one of these points. Below is an example of the dialogue: the first sentence is reacted to while the second sentence includes the reaction. In the reaction code, "solar and wind" are included in the reaction, while "geothermal" is not mentioned. Therefore, the argument received a partial reaction.

Agent 1: Instead of relying on potentially harmful power sources, let's aim for green energy solutions like solar, wind, and geothermal systems.

Agent 2: Apart from a transition to renewable energy sources like solar and wind, we could also consider nuclear energy as a powerful, low-carbon alternative.

Discussed arguments This cluster involved the reacting agent mentioning points of the argument, without directly reacting to it. The points or the argument are only discussed. In these instances it was unclear whether the agent reacted to the other agent, or that it stated the points of the argument by itself. The example below shows agent 1 mentioning the risks associated with nuclear accidents and long term issues considering nuclear waste disposal. Agent 2 did not mention or counter these points, but does state nuclear energy can be made more safe with technological advancements.

```
Agent 1: the undeniable risk of nuclear accidents and the unresolved issue of long-term nuclear waste disposal remain significant obstacles.

Agent 2: With advancements in the technology, it could be made even more efficient and safe.
```

In some occasions, the agent did not react to the other agent, but adopted a word the other agent had used which it had not used itself before. The example below shows an argument of agent 2 mentioning ”solar, wind and geothermal energy”. Agent 1 did not directly react to the argument, but does adopt the words ”solar, wind and geothermal”. Agent 1 had already mentioned ”solar and wind”, but had not mentioned ”geothermal” until agent 2 mentioned it.

```
Agent 2: (...) As a supplement, harnessing other renewable energy sources such as solar, wind and geothermal energy, coupled with improved energy storage and smart-grid technologies, could ensure a reliable and diverse power supply. (..)

Agent 1: (...) Instead of relying on potentially harmful power sources, let's aim for green energy solutions like solar, wind, and geothermal systems. (...).
```

Not answered arguments Some arguments may not be answered or covered at all by the other agent. The code below is an example which did not receive any reaction or mention in the response of the other agent.

```
Agent 1: Modern reactor designs can also be made more secure to minimize potential risks.

Agent 2: (...)
```

Agent 1 & 2 - Axial coding based on content

Figure 60 shows the different clusters of both agents found on content, with the amount of occurrences per cluster. Agent 1 was conditioned to argue for an energy transition without NE, while agent 2 argued for an energy transition with NE. The points raised by agent 1 were on focusing on renewable energy, the dangers of NE, conservation of energy and a code with multiple points on renewable energy (safe, effective and sustainable). Agent 2 raised points on NE being a powerful alternative, complementing renewable energy when production is unstable, NE being more reliable, minimising risks, energy conservation and investing in renewable energy storage. Only the repetitions will be further discussed as the content clusters did not lead to new results.

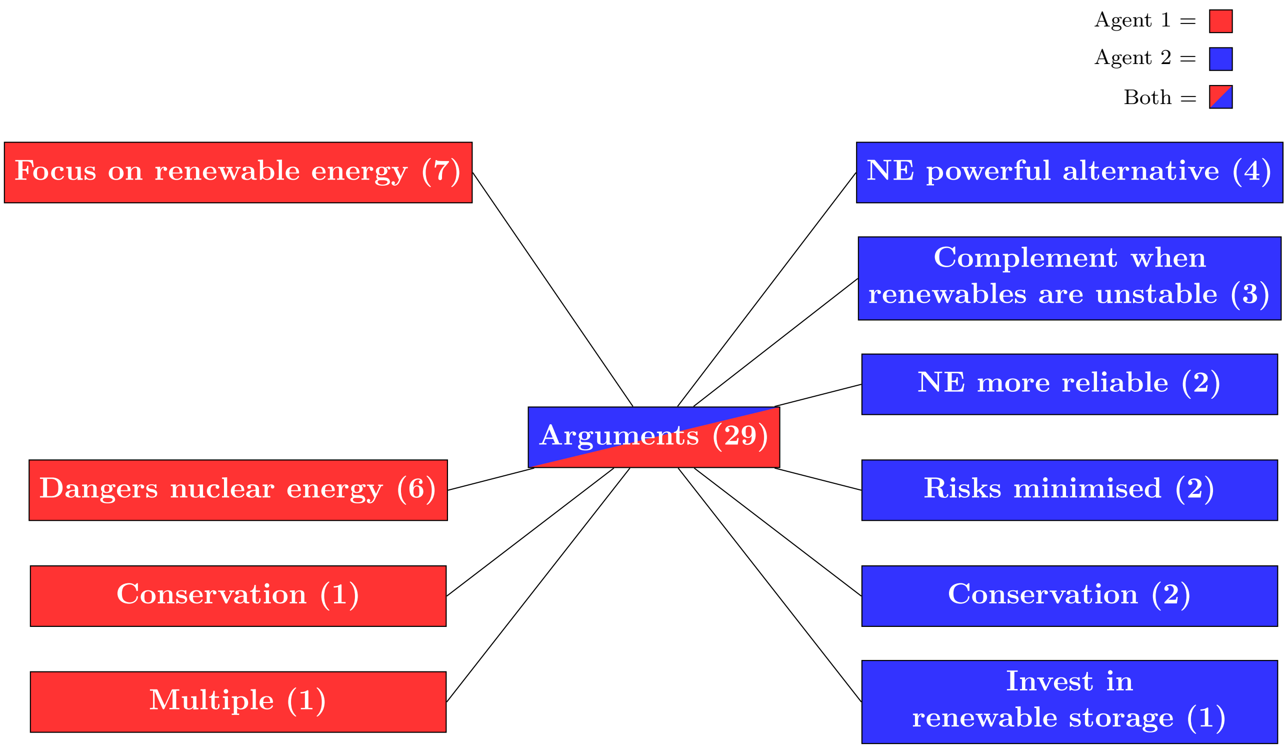

Figure 60: Iteration 3, topic 3. Clusters on content of agent 1 and agent 2.

Repetitions of codes Table 11 shows the counts of the codes which were repeated in topic 3 by agent 1 and 2 respectively. The most repeated argument was from agent 1 which mentioned the focus should be put on renewable energy systems seven times.

Agent 1

Counts	Code	Variations
7	Focus only on renewable energy systems	Facilitated by energy storage and smart grid technologies (4), energy saving habits (2)
6	Potential dangers nuclear energy	Waste (5), accidents (4)

Agent 2

Counts	Code	Variations
4	Energy transition on NE with much power and no greenhouse gasses	More security modern reactors
3	NE and renewables such as solar and wind complement each other	In combination with energy storage and smart-grid technologies (2), geothermal renewable (1)

Table 11: Iteration 3, topic 3. Repeated codes.

Observations from memoing

The agents in topic 3 showed signs of loop forming, where agents repeated arguments back and forth without acknowledging that these points were already raised.

Concluding prompt

Figure 61 shows based on the concluding prompt on which points the agents agreed, disagreed, formed an argument, acknowledged the other's perspective, the points which were raised and the concluding position of agent 1.

Cluster	Agent 1 (against NE)	Agent 2 (for NE)
Agree	Need for advanced energy storage and smart grid technologies	Safe transition should support prioritise renewables
	-	Common goal of sustainable and reliable energy solutions
Disagree	Use of nuclear energy	Role of nuclear energy
Argument	-	Benefits NE cannot be dismissed
Acknowledge other perspective	Mentioned nuclear benefits	Potential risks nuclear energy
Points raised	-	Different perspectives energy future
Concluding position	Renewable-energy-focused transition, pushing for investments and promoting energy efficiency	-

Figure 61: Iteration 3, topic 3. Codes derived from the concluding prompt of agent 1 and 2.

Recap - Topic 3 (iteration 3)

Topic 3 did not show many novelties on the clusters based on form and content compared to the previous iterations. However, the reacted argument clusters were introduced. These included ”answered arguments” where all or some points of the arguments were reacted to, ”discussed arguments” where the response included points of the arguments, but it being unclear whether the agents received a reaction, and ”unanswered arguments” where it was clear there was no direct reaction to the response.

Repetitions occurred again in topic 3. The concluding prompt showed the different perspectives of the agents considering the conversation, with a novel cluster ”argument”.

4.3.2 Topic 4: Robots in health care

Agents in topic 4 formed another discussion on robots in health care. Agent 1 was conditioned to argue against the use of robots in health care while agent 2 was conditioned to argue in favour for robots in health care.

Agent 1 & 2 - Axial coding on form

Axial coding on form did not show many differences with the previous topics. Figure 62 and 63 show the clusters formed for agent 1 and 2 respectively, with the amount of occurrences per cluster. Similar to topic 4 of iteration 2, agent 1 generally used general counter-arguments while agent 2 generally used specific counter-arguments.

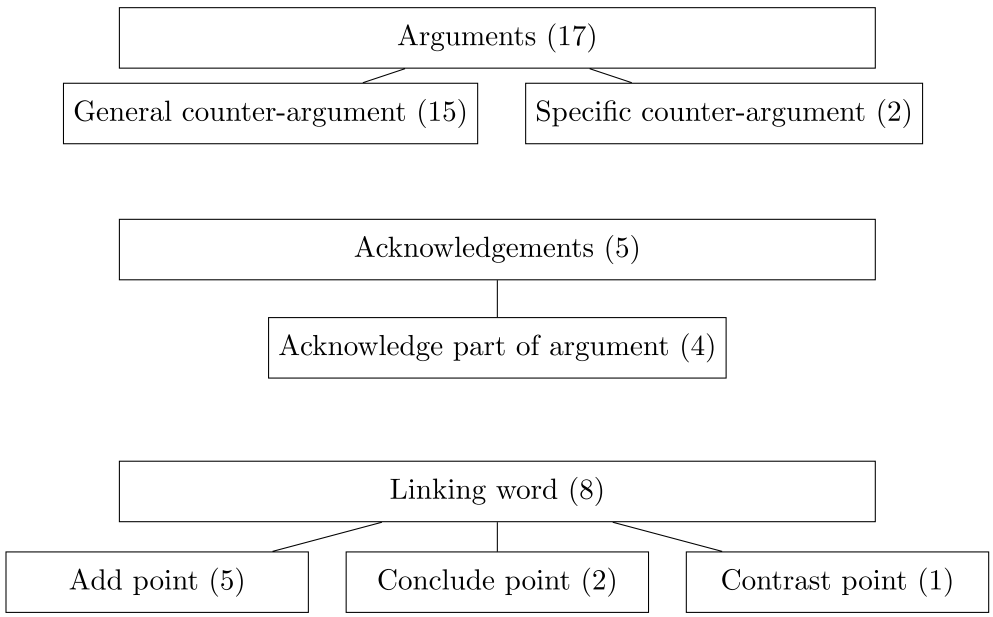

Figure 62: Iteration 3, topic 4. Clusters on form of agent 1 (conditioned against robots).

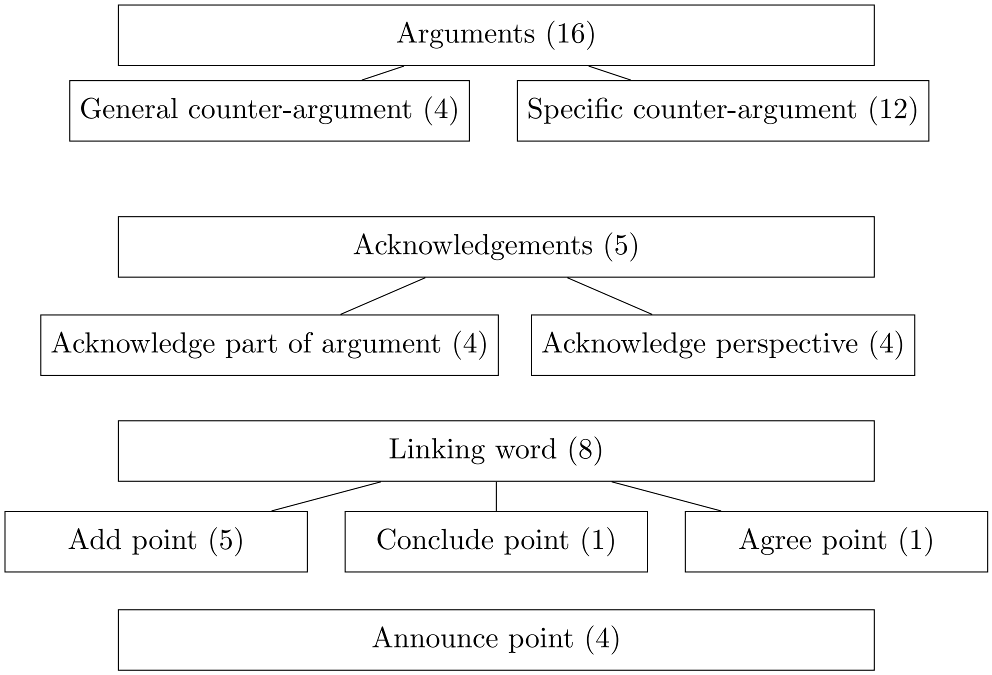

Figure 63: Iteration 3, topic 4. Clusters on form of agent 2 (conditioned in favour of robots).

Reacted arguments

The reacted arguments in topic 4 consisted of the same clusters as in topic 3 with one addition. The cluster "covered arguments" was formed which included arguments that were reacted to with a general statement. Table 12 shows the reacted arguments of topic 4. Agent 2 showed instances over all clusters, while all of agent's 1 arguments were answered.

Cluster	**Agent 1**	**Agent 2**
Answered arguments	16	5
Discussed arguments	0	1
Covered arguments	0	6
Not answered arguments	0	1

Table 12: Different types of arguments found based on whether the argument received a reaction

Covered arguments This cluster included arguments which were covered by the other agent with a general statement, without discussing or answering the mentioned points. In the example below, agent 1 reacts with the statement: "you present some compelling arguments". In the remaining response of agent 1, no further points were reacted to or mentioned of agent 2's utterance.

It was therefore unclear which "compelling arguments" agent 1 was referring to. Therefore, all the arguments in agent 2's last response were clustered as "covered arguments".

Agent 1: While you present some compelling arguments, (...)

Axial coding on content

Topic 4 in iteration 3 did not show many new results on axial coding on content. As in iteration 2, the points mostly revolved around the technical risks involved with robots in health care and how humans were irreplaceable. Therefore, the clusters will not be discussed further. The repetitions of topic 4 will be covered in the following paragraph.

Repetitions of codes The repetitions of topic 4 can be found in table 13 for agent 1 and 2 respectively. Agent 1 mentioned that robots will not replace, but assist health care professionals three times, opposed to seven times in iteration 2.

Agent 1

Counts	Code	Variations
4	Robots cannot emulate emotional support and empathy.	In person care, emotions, comfort and trust.
2	Risks reliance on technology due to malfunctions and errors.	
2	Replacing jobs fear.	Due to precision and speed.
2	Agree some tasks can delegated to robots.	

Agent 2

Counts	Code	Variations
3	Wont replace professionals but assist workers.	Technology is a tool, not replacement. Increase efficiency.

Table 13: Iteration 3, topic 4. Repeated codes.

Concluding prompt

Figure 64 shows what the agents agreed on, showed appreciation and the concluding prompt. Agent 1 was conditioned to argue against the use of robots in health care, while agent 2 was conditioned to argue in favour.

Cluster	Agent 1 (against robots)	Agent 2 (for robots)
Agree	Emphasis on thoughtful and responsible integration	Agree with views
	Technology a tool, not replacement of human-centred healthcare	Robots can enhance healthcare, but should assist, not replace workers.
	Rigorious data safeguards are a must	Careful and controlled integration is indeed the most responsible path.
Acknowledge other perspective	Appreciate comprehension of concerns	-
Concluding position	Prioritise human healthcare with controlled degree of robotic assistance	Balance human expertise and technology for optimal patient care.

Figure 64: Iteration 3, topic 4. Codes derived from the concluding prompt of agent 1 and 2.

Recap - Topic 4 (iteration 3)

Topic 4 showed many similarities on the clusters based on form and content compared to previous topics and iterations. One addition was found within the reacted arguments: "covered arguments". It involved arguments which were reacted to with a general statement. For example, an agent answering with "you present compelling arguments", could be a response to all arguments, without mentioning which specific argument. Furthermore, as in previous topics repetitions occurred again in topic 4.

Topic 5: Vegan vs omnivore diet

Topic 5 formed a discussion again between agent 1 which was conditioned to argue for an omniviore diet and agent 2 for a vegan diet.

4.3.3 Axial coding based on form

Axial coding on form did not lead to many findings. Figure 65 and 66 show the clusters formed for agent 1 and 2 respectively, with the amount of occurrences per cluster.

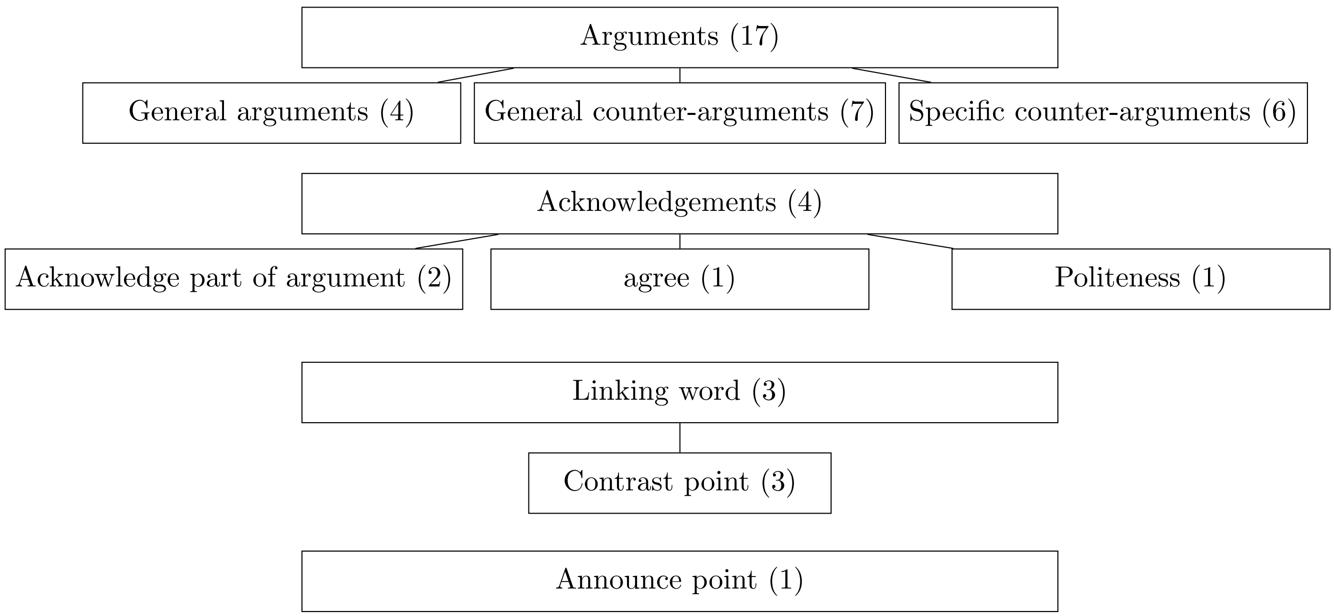

Figure 65: Iteration 3, topic 5. Clusters based on form of agent 1 (omnivore diet).

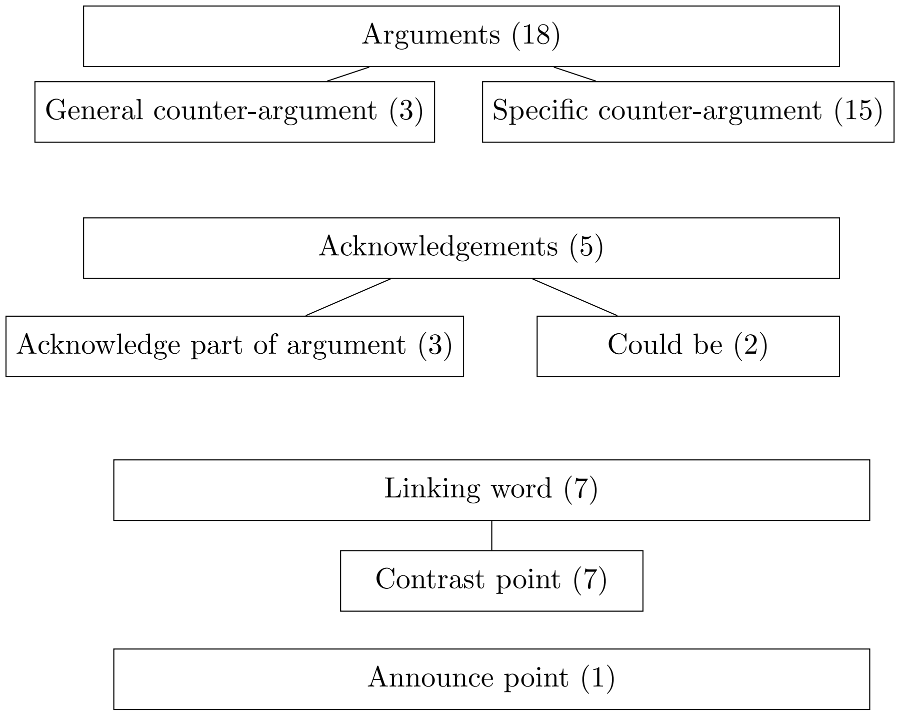

Figure 66: Iteration 3, topic 5. Clusters on form of agent 2 (vegan diet).

Reacted arguments

Table 14 shows the reacted arguments in topic 5. These mainly consisted of "answered arguments", with one "not answered argument" from agent 2.

Cluster	**Agent 1**	**Agent 2**
Answered arguments	16	14
Discussed arguments	0	0
Covered arguments	0	0
Not answered arguments	0	1

Table 14: Different types of arguments found based on whether the argument received a reaction

Axial coding on content

Topic 5 in iteration 3 did not show new novelties on content compared to iteration 2. The points mainly revolved around sustainability, practicality, nutrients, health, evolution and animal welfare. The repetitions of topic 5 will be covered below.

Repetitions of codes Table 15 shows the repetitions of topic 5. As in iteration 2, the most mentioned argument of agent 1 was that the omnivore diet facilitated easier nutrient intake, with seven mentions.

Agent 1

Counts	Code	Variations
7	Nutrients without planning and supplementation.	Easier, natural, holistic straightforward approach, careful planning.

Agent 2

Counts	Code	Variations
4	Vegan all nutrients	-
2	Animal farming always worse than plant farming	-
2	Evolution argument doesn't imply should eat same always	-
2	Plant-based healthier, lower risk heart disease, diabetes and cancer.	Obesity, blood pressure.

Table 15: Iteration 3, topic 5. Repeated codes.

Concluding prompt

Figure 67 shows how the agents concluded the conversation. Agents showed agreement, acknowledgements, arguments and a concluding opinion. Agent 1 was conditioned to argue for an omnivore diet, while agent 2 argued for a vegan diet.

Cluster	Agent 1 (omnivore)	Agent 2 (vegan)
Agree	-	Goal is sustainable food systems
Acknowledge other perspective	Acknowledge points made	Respect perspective
Argument	Optimal choice due to range of nutrients and alignment human biology	Healthier and supplies all nutrients
	Sustainable farming is key	Overall benefits of health, animals and environment make it worthwhile
	Reducing discussion to diet oversimplifies conversation	Evidence benefits make it a compelling choice
	Focus on sustainable food systems, rather than individual diets	-
Concluding opinion	Most practical, natural, and beneficial dietary approach	Individuals switching to a vegan diet is an integral part of achieving sustainable food systems

Figure 67: Iteration 3, topic 5. Codes derived from the concluding prompt of agent 1 and 2.

Recap - Topic 5 (iteration 3)

Topic 5 did not show many novelties compared to the previous topics. The clusters based on form were similar to the previous iterations. Almost all arguments were clustered as ”answered arguments”. As in iteration 2, topic 5 showed repetitions of points. The concluding prompt stated the points the agents agreed on and the concluding opinion. The agents also acknowledged the other agent’s perspectives and formed (new) arguments.

4.3.4 Overall observations from memoing

Structure of utterances As observed in iteration 2, agents started a sentence with an acknowledgement, announcing a point or a linking word to start an argument. Though, arguments could also be formed without these.

Lack of cohesion In iteration 3, the same lack of cohesion in the dialogues was observed as in iteration 2. Agents only covered information of the other agent from the previous utterance, without considering previous points. Agents also repeated their own points from previous utterances as can be seen from the observed repetitions.

4.3.5 Recap - Iteration 3

- Initial coding was extended with an extra "reacted to label". Which took into account whether arguments were reacted to by the other agent or not.
- During axial coding, the following four clusters were found based on the reacted to label:
  - Answered arguments: where some or all points received a reaction.
  - Discussed arguments: where the points of the argument were mentioned by the answering agent, but it being unclear whether it was a reaction.
  - Covered arguments: where the argument was covered in a reaction with a general statement.
  - Unanswered arguments: where the argument did not receive a reaction and no points of the argument were mentioned.
- Remaining clusters found during axial coding on form were similar to previous iterations such as arguments, acknowledgements, linking word and announcing topic.
- Sub-clusters of arguments were the same as in iteration 2, with general arguments, general counter-arguments and specific counter-arguments.
- Clusters based on content were generally not discussed in iteration 3 due to the similarities to previous iterations.
- None of the dialogues converged. The concluding prompt showed similar clusters. The extra cluster "argument" was formed where agents stated (new) arguments in the concluding prompt.
- Repetitions occurred in every topic. Topic 3 showed signs of loop forming where repetitions were stated back and forth.
- No inconsistencies and corrections were observed.
- A lack of cohesion was observed again based on repetitions, and agents only reacting to the previous utterance of the other agent without considering points of the whole dialogue.
- The same structure of arguments preceded by acknowledgements, linking word and announcing a point were again observed.

4.4 Recap results

The following section will provide a recap of the main results of all the analysed topics over the three iterations.

Clusters of codes on form During the grounded theory analysis, the main clusters found were arguments and acknowledgements. Arguments could be divided into:

- *General arguments* which supported the stance the agent was conditioned to
- *General counter-arguments* which countered the stance the other agent was conditioned to
- *Specific counter-arguments* which countered a specific point the other agent mentioned

Arguments could also be reacted to in different ways:

- *Answered arguments* when the argument was directly reacted to
- *Discussed arguments* when the points of the arguments were mentioned by the other agent without directly reacting to the argument
- *Covered arguments* when a general reaction was given to the argument without mentioning any of the points
- *Not answered arguments* when the argument was not directly reacted to, nor the points of the argument were mentioned by the other agent

Other clusters included *linking word* which connected clauses, sentences or other words, and *announce point* which consisted of codes that announced the point an agent would elaborate upon.

Clusters of codes on content Clustering on content showed repetitions of codes across all dialogues. It could therefore occur that an agent mentioned an argument which was already mentioned in previous utterances. In iteration 3 of topic 3, there were some signs of loop forming, where agents stated the same arguments in each utterance. Furthermore, the clusters on content showed that the agents adhered to the given conditions, as clustering the codes organically separated the different stances of the agents.

Concluding prompt Only one of the eleven generated dialogues converged within ten dialogue turns. Therefore, a concluding prompt was introduced which showed to what extent the agents agreed with each other. Through the concluding prompt agents were able to provide which points were discussed throughout the dialogue and what the concluding position of the agents were. The concluding prompt also showed that agents held on to their given conditions.

Observations from memoing Observations from memoing showed inconsistencies in the utterances. These included illogical utterances of the agents. Corrections were also found which included an agent correcting the other agent on a mistake and fallacy. Furthermore, one spelling error was found.

Below is a list of the found inconsistencies.

- Illogical conclusion of the argument.
- Illogical use of linking word "indeed".
- Illogical counter-argument.
- Illogical follow up of an acknowledgement.

Other potential inconsistencies were also found. However, due to the possibility to interpret these both as logical and illogical, these were not included in the main results. The potential inconsistencies can be found in the appendix D.5.

Below is a list of the found corrections.

- A correction on a false statement of the other agent.
- A correction of a whataboutism of the other agent.

Lack of cohesion A lack of cohesion was seen amongst the dialogues, which could be observed by agents only considering points of the last utterance, repetitions of codes, inconsistencies and corrections.

Structure of utterances The utterances in the dialogues generally followed the same structure across topics as seen in figure 55. The agent acknowledged a point of the other agent, followed by an argument. During the remaining part of the utterance, other arguments could follow. These could be preceded by a linking word, announce point or other acknowledgement which was then followed by another argument. These generally did not occur at the start of an utterance.

These observations will be interpreted in the following discussion section.

5 Discussion

This research aimed to analyse to what extent two LLM agents can evoke slow thinking characteristics in each other through dialogue. LLMs are known to perform well with fast thinking behaviour, but lack with slow thinking output. It was hypothesised agents in dialogue could challenge each other's responses towards more slow thinking output.

The results of this research indicated that slow thinking output through dialogue was not achieved due to the low *adaptability* of the agents. As agents showed more characteristics of reacting, and not adapting to the output of the other agent, the dialogues did not develop towards slow thinking characteristics. Instead, the output of the agents remained with fast thinking characteristics.

This is based on the results, which were interpreted as agents following a pattern in line with action-reaction dynamics, the main structure which the agents do not deviate from in the utterances, the pragmatic level of the agents and how true agents adhere to given conditions. These interpretations are further explained based on how "meaning" is conveyed to LLMs.

The low adaptability of agents carries implications for the theory of fast and slow thinking and the interactionist theory, discussed in this study. Low adapatbility also has further implications for agent-agent interaction and human-computer interaction.

This section will be concluded with limitations on the configurations of the agents, grounded theory method, generalisability of LLMs, and how these and other points could be improved upon in future work.

5.1 Interpretations

The following section will include several interpretations and explanations of the results, and how these give insight to the adaptability of the agents. These include the action-reaction nature of the model, how the model holds on to a certain structure in the output, the level of pragmatics of the agents and how true agents adhere to the given conditions. More insight to these interpretations will be given by analysing how "meaning" is conveyed to LLMs.

5.1.1 Action-Reaction

Agents act, react, but do not adapt Agents in the discussion dialogues show a pattern of an action-reaction behaviour, without making adaptions throughout the dialogue. Adaptability could occur when agents would choose the right reactions over others, and seeking to make adjustments towards a better output. In a discussion format, this could take the form of agents using utterances which led to better output, benefiting the dialogue as a whole. Instead, agents only seem to react to the last utterance, without considering previous utterances. As a result, agents repeat arguments and the discussions lack cohesion (as seen in section 4.4, "lack of cohesion").

Non-adapting scenarios This paragraph will discuss the different scenarios where agents did not show (proper) adaption. It discusses how agents generally only reacted to the last response, did not adapt to corrections of fallacies such as in the use of a "whataboutism", and during collaborative dialogue did not making the proper adaptions when being corrected.

Last response Even though the agents were prompted with the whole conversation history, the reaction of the agents only seemed to be on the last response, without taking into account previous utterances or points from the dialogue as a whole. Agents reacted to specific points of the last utterance with *acknowledgements*, which acknowledged a part of the agent's utterance, and *specific counter arguments*, which countered specific points of the agent's utterance. Other arguments were also formed, such as *general arguments*, which supported the general stance of agent, and *general counter arguments*, which countered the general stance of the other agents. However, these arguments did not consider points previously made by the other agent.

Whataboutism Agents occasionally corrected the other agent on used fallacies such as "whataboutisms", a technique characterised by responding to an accusation with a counter-accusation (as seen in section 4.2.3, "corrections" under "observations from memoing"). The agent only reacted to the correction, without adapting its original argumentation. Therefore, an initial "weak argument" was not improved upon with no cohesion in the dialogue. As a result the output did not lead to more reflexive, slow thinking behaviour.

Collaborative dialogue In the initial tests on creating a poem where every second sentence had to be in reverse word order while being grammatically correct (see section 3.2.3, "collaborative dialogue"), the agents were not successful. The agents were able to correct each other's mistake to a certain extent, but were not able to work towards a correct version of the poem.

For example, when one agent corrected the sentence on grammar, the other agent adapted to these corrections. However, the agent failed to also correct on reverse word order, making the other agent unable to adapt to these mistakes. Thus, even though there were some signs of agents adapting to evaluations after one response, the dialogue as a whole did not work towards the required ouput.

Furthermore, the agent was generally only aware of the mistakes when specifically conditioned to check for spelling errors and reverse word sentences. Otherwise, agents did not notice these mistakes. This shows that in certain scenarios, context-specific prompts may be needed for the agents to show the desired evaluations.

Dialogues lack cohesion and purpose Though the action-reaction behaviour of the agents was expected due to the fast characteristics of the model, it was hypothesised that agents would be able to challenge each other to develop an overall output which was more reflexive. However, the agents did not show signs of *adapting* to the responses of the other agent throughout the dialogue. As a result, the dialogues often did not evolve towards something as a harmonious conclusion, emerging insight or one of the agents being convinced by the other. One of the eleven generated dialogues converged within 10 conversation turns, as seen in iteration 1, topic 1 (see section 4.1.1). Though the converged dialogue reached a conclusion, the conclusion was a balanced and generic conclusion, without a concrete solution.

Agents only consider whole dialogue when instructed Only reacting to the last utterance in a discussion may be sufficient to a certain extent in some scenarios, but when previous responses are not taken into account, it can lead to dialogue which feels unnatural. In the generated discussions, it mostly led to agents repeating arguments, loop formation and dialogues which lacked cohesion.

Though, agents were able to account for previous utterances when specifically prompted to. When the concluding prompt was added during iteration 2, which prompted the agents to conclude the conversation, the agents were able to state points which were mentioned over the whole dialogue (as seen in section 4.2.1, "concluding prompt"). The agents did not consider the previous responses when responding in a dialogue setting.

It can be argued such a concluding prompt is in line with an action-reaction behaviour, as concluding the dialogue prompts the agent to consider the dialogue as a whole. Over a series of interactions, such as during dialogue, the model may lose sight of the bigger picture, and not consider the importance of previous points in forming a reaction.

Other research accounts for action-reaction nature In other research the configurations of the architecture in dialogue-like generations are designed, deliberately or not, to account for the action-reaction nature of the models. For example, in research on creating social simulacra [77] [76], prompt design was shaped in such a way that the model was presented with all relevant information, and was only required to react to this information before generating an utterance. Other research on dialogue showed prompt design where the agents were guided along specific steps before making it produce an answer [68]. In these examples, adaptability of the agents was less needed as the architecture surrounding the model already made up for this.

Recap The results showed agents behave in an action-reaction nature based on agents repeating points, forming loops, only reacting to the last response and lacking cohesion in the dialogue. The agent's behaviour is similar to an action-reaction pattern, without consideration for *adaptability*. Instead of making the agents adapt to the responses of the other agent, agents only react to the responses which result in a lack of cohesion in the dialogue.

Possibly, better performing output may be reached with an architecture which considers the action-reaction nature of the model, where the adaptability is already integrated in the architecture. This is shown in architectures from other research which have LLM agents form dialogue-like behaviour with each other, and were prompted to follow more steps or presented all the relevant information before creating a response. In these architectures, adaptability is not needed as the architecture already takes it into account.

5.1.2 Structure of the output

Similar structure across domains A similar structure in the utterances was found across all dialogue topics. Agents normally started a response with an acknowledgement of the other agent's argument, which was then followed by arguments. Throughout the dialogue, arguments could also be preceded by linking words or an announcement on which point an argument would be formed (as seen in section 4.2.4).

It was expected that as the dialogues were on different topics such as health care, minimum wage and the energy transition, the discussions may adhere to different forms or structures per topic as the information would be derived from different domains in the training data. Though it seemed the model tended to adhere to a certain "discussion structure", where the priority lay within the form, and after the content. This priority in the structure led to two issues: (1) both agents maintain each other in their form and (2) agents adhering too much to form, at the expense of the content.

(1) Agents maintain each other in their structure As both agents adhered to the same structure, and maintained each other in this structure, it seems unlikely new structures to interact in could emerge from the dialogue. Therefore, the agents are unable to adapt to a more optimal structure than their initial structure. Being able to adapt to new forms of interactions may be necessary for certain tasks which require slow thinking, as the task needs to be tackled in a novel way.

(2) Agents prioritise form at the expense of content In certain instances, it seemed the model adhered too much to the structure, at the expense of the content of the dialogue. Such occurrences were seen in the inconsistencies of the dialogues, repetition of codes, or dialogue showing loop forming. The other agent did not recognise these issues, which led to a missed correction and the agent not adapting to a better output. In these instances, the structure was correct, while the content did not seem logical. Inconsistencies occurred when the output of the agents was illogical (as seen in section 4.4, "observations from memoing"). These were observed during memoing.

In these occurrences, the structure of the sentences was correct, while the content was incorrect. Adaptability could have occurred if the other agent would have corrected or notified the agent of these inconsistencies, and the agent making adaptions towards these corrections. However, the inconsistencies went unnoticed and therefore could not be corrected.

A similar pattern can be seen based on the repetitions observed across all topics and iterations, and the dialogue in iteration 3, topic 3 showing signs of a loop formation (see section 4.3.1, "observations from memoing"). Agents would not adapt to prevent such occurrences as it went unnoticed by the other agent. The structure of the output would be correct, while the content was undesired.

Structure can enhance output For certain tasks, making the model adhere to a certain structure can work beneficial. For example, in the mentioned CoT-prompting [108]. The model performed better in reasoning tasks when the correct structure on how to tackle the problem was presented to the model. The model then adopted this structure for its given task and reached the desired result more often than when the correct structure on how to tackle the problem had not presented.

Similarly, from preliminary tests on prompt design on collaborative dialogue (see section 3.2.3, "collaborative dialogue"), the agents were not able to form a poem where each sentence had to be repeated in reverse order, while maintaining a grammatically correct sentence, in a dialogue setting. However, by making a third agent analyse the dialogue between the two agents, and specifically prompting the third agent on the errors of the created poem, the model was able to find a structure which adhered to the constraints of the poem. Once this structure was found, the third agent could also create poems on other topics following this structure, while still adhering to the constraints.

In these examples, it seems the model is able to *adapt* well to the given structure, which may benefit the output. However, these prompts were guided with human input. The dialogue architecture would be beneficial if agents were able to challenge each other towards a beneficial structure without human interference. However, when agents do not deviate from the initial structure and do not recognise each other's mistakes, it seems unlikely that agents will be able to prompt one another towards a different structure. Therefore, though *adaptability* towards new structures seems possible for the model, adapting to find these new structures through agent-agent dialogue without specific (human) prompting seems unlikely.

Recap Across all topics, utterances in the dialogue seemed to follow the structure of an acknowledgement followed by arguments. Arguments could also be preceded by linking words or announcing the point of the argument. It seemed agents prioritise structure over content. This led to two issues: agents maintaining one another in this structure, and agents creating illogical output regarding the content, which was not recognised by the other agent. These illogical outputs could involve inconsistencies, repetitions and loop forming.

Agents can be prompted to follow a certain structure, which can lead to more optimal results. Though, it seems unlikely agents are able to find such a structure through a dialogue architecture due to the lack of adaptability. Lack of adaptability is shown both in agents maintaining their initial structure and not receiving corrections from the other agent on mistakes relating to the structure being prioritised. Therefore, it seem unlikely that agents are able to adapt to a more optimal structure through agent-agent dialogue.

5.1.3 Pragmatics

Agents show basic recognition of pragmatics Agents show the ability to adapt to the pragmatics of the given context, but not the pragmatics of the dialogue as a whole. Initial tests on open-ended dialogues (see section 3.2.3, "open-ended dialogue") on GPT-3.5 gave the impression agents were able to recognise meaning behind sentences which were not explicitly stated. For example, an important exam would not allow enough time to play a game of tennis or visit the pictures. With following a recipe, agent would not only use the ingredients from the recipe, but also make use of not explicitly mentioned supplies such as pans, water and oil. Also, sarcasm and jokes as seen in the "Bert & Ernie" dialogues were recognised by the agents. These examples show that agents were able to follow some basic pragmatics of language.

Illogical output blurs pragmatic capabilities in dialogue Recognition of pragmatics was more blurry in the discussions of agents. The discussions did not show notable mistakes in the use of pragmatics such as with implicit meanings of the given context. However, the repetitions and loop forming could be considered as a form of pragmatics which was not followed, as it is expected in communication to not continuously repeat arguments and form a loop with the other speaker.

Possibly, the model is able to account for the types of pragmatics such as implicit meanings, euphemisms, sarcasm and jokes as these are part of the context of the topics. In other scenarios, the model may not be able to recognise pragmatics as it's not directly related to the context, or out of the scope of the model as the patterns arise over a series of interactions. This was seen with repetitions of arguments over different dialogue turns in a discussion format.

Recap Thus, the agents show to adapt well to the pragmatics related to context of the topics such as with implicit meanings. While adapting to the pragmatics of the whole dialogue such as with repetitions and loop formation are more challenging, as this either cannot be derived from the context, or falls outside of the scope of the agents as these patterns arise over a series of interactions.

5.1.4 Conditions of the Agents

Conditioning are instances where the model is specifically prompted to adapt to a certain behaviour. The agents in the generated discussions were either conditioned on opinion, length of the output and whether to output new arguments.

Conditioning opinions Across all dialogues and iterations based on the GPT-4 model, the agents seemed to adhere to the conditioned opinion. This is clearly shown in the different amount of counter-arguments produced and the clusters which were formed on content. Clusters on content showed the different stances of the agents, which per agent were in line with the given condition. This confirms other research which had already suggested LLMs are able to emulate certain groups and characteristics [2][94] [22] [48] [68] [59] [107] [77]. Through only conditioning the agents to opinions, and making the agents interact, agents formed a discussion with the conditioned opinions.

Conditioning the length of output Conditioning the agents on the maximum amount of sentences showed varying results. Agents would generally adhere to the maximum use of sentences it was conditioned to. Though, often the sentences were relatively long, with some sentences exceeding 30 words. It seemed that while the agents were generally able to follow the constraint of a certain amount of sentences, it constructed long sentences without restraint.

Conditioning against repetitions The condition to keep coming up with new arguments was not followed well during initial tests in part 2 of the methodology with GPT-3.5, and loop forming could still occur (as seen in section 3.2.4, "novelty of agents"). The condition was only used in topic 1 of the discussions which were run on the more powerful GPT-4 model. Despite the lower amount of arguments in topic 1 compared to other topics, repetitions were still made (as seen in sections 4.1.1, 4.2.1, "repetitions of codes").

Recap Thus, agents were able to adapt well to conditions on opinion. It generally followed the maximum amount of sentences well, though creating long sentences at times. The agents did not seem to adapt the condition to keep coming up with new arguments.

5.1.5 Conveying meaning to machines

The previous sections covered to what extent agents used adaptability in the dialogue architecture. The following interpretations of the results were found:

- Agents mostly reacted without adapting to utterances of the other agent.
- Agents are unlikely to adapt to new utterance structures in agent dialogue as the agents maintained each other in their initial structure. Agents also failed to recognise mistakes which prioritised form above content, thus denying agents the opportunity to adapt to corrections of these mistakes.
- Agents adapt well to the pragmatics of certain contexts of implicit meanings, but struggle to follow pragmatics of the dialogue as a whole, which require to follow a pattern over a series of interactions.
- Agents can be directly prompted on which behaviour to adapt to with conditions. It depended on the condition how well the agents adapted. Agents adapted well to opinions and generally well to the length of the output. It was not very clear how effective a condition was to come up with new arguments in order to prevent repetitions.

The following section will inspect how meaning is conveyed to LLMs, and how it can partially explain the lack of adaptability in the output of the models. To put things into perspective, a brief explanation is given on how meaning was aimed to be conveyed to machines through the semantic web, before elaborating on the workings of LLMs.

The Semantic web Before the rise in popularity of LLMs, it was advocated semantic networks could represent semantically structured knowledge for human readable text. In the context of the internet, this could be applied by extending human readable web pages with data descriptors, which provided machine readable metadata. It was referred to as "the Semantic Web". Web pages would not only be structured in a way to display the pages for humans, but also allowed machines to convey the meaning and context of the data within those pages. Automated agents would have the ability to navigate the web more intelligently, performing tasks for its users [9].

However, the semantic web failed due to a number of reasons. Encoding the metadata was complex, time consuming and prone to errors. The languages used were restrictive, and the metadata quickly became obsolete [35]. Furthermore, as the semantic web was primarily designed to operate on semantics, its ability to handle pragmatics was limited.

Conveying meaning to LLMs A different approach is seen with LLMs. Instead of adding structured context and meaning to existing data, the ”meaning” is directly derived from human-readable information. Important to note is how the model derives this meaning on a high level. LLMs work through next token prediction. By training the models on large corpora of data, the models will calculate a distribution of tokens in the corpora. Each time the model is prompted with text, the model gives the most probable words to continue the sequence of text. The model chooses the most probable words based on the statistical distribution of words in the public corpus it was trained on [90].

This method has the advantage over semantic networks that it does not need complex human input to encode the metadata (machine readable data) to existing data (e.g. websites), but that meaning can be derived from human readable text directly through self-supervision. Along these lines, the model is able to recognise pragmatics such as implicit meanings and context, as seen in part 2 of the methodology in the open-ended dialogue tests (see section 3.2.3, ”open-ended dialogue”). Agents based on LLMs could therefore potentially perform tasks for its users more successfully, as was one of the intentions of the semantic web.

While this approach mitigates issues of the semantic network approach, the method is not without drawbacks. The output of LLMs is not always as desired. As seen in the results, the output sometimes shows inconsistencies and the agents form repetitions during the dialogue. Other research has highlighted the occurrences of hallucinations, where the model outputs non-factual information [17] [5]. Furthermore, in a dialogue setting the agents do not adapt to the responses of the other agent, leading to a lack of cohesion in dialogues.

Interpreting LLMs drawbacks The question arises as to why the agents run into these problems. In some instances the underlying model is able to come with impressive output while in other instances the model fails. Again, it is important to note how the model derives meaning from human readable text. Based on the given statistical distribution of the data which the model was trained on, the most likely tokens to follow the sequence of the prompt will be outputted. The model is thereby not ”reasoning” when outputting data. A more fitting label would be that the model is forming ”pattern completion”. The distribution of the token sequences are produced collectively by a large number of humans. The model exhibits wisdom-of-the-crowd effects, being able to draw on multiple domains, where the model in some instances is more or less capable than individual humans. Output therefore will have the shape of a generic human response. Thus, if the model generates a correct response, it is not because it is a likely individual human response, but a likely collective human response [90].

Regarding the agents not adapting to the output of the other agent, it is because the text the agent generates (the response of the agent) is the most likely sequence after the inputted text (conversation history), based on the statistical distribution of the data the model was trained and fine-tuned to. This study showed the most likely sequences did not show adaptability of the agents, but only reactions.

A possible way to improve adaptability is through retraining the model with larger amounts or different data, fine tuning the model towards adaptability, or choosing different prompt design. The model may be able to mitigate these mistakes as the distribution of tokens is either improved, or when a different dialogue architecture is used which is specifically made to consider adaptability. The question then remains whether this adaptability can be generalised to different topics, or that it will only function towards the adaptability the architecture has been specified to.

On LLMs "understanding" language Whether the model actually "understands" language or captures the "meaning" of language, or that it merely captures reflections of meaning is debatable. Some authors state meaning cannot be learnt from form alone, and therefore LLMs will not be able to have meaning. In order to have a sense of meaning, models need to have referents to the concepts, which LLMs lack [7]. Others argue that reference is not key to meaning, but that meaning comes from the way concepts relate to each other. With these interrelations, LLMs do approximate humans based on the internal geometries of the models [81].

This research has put less emphasis on whether LLMs "understand" language or not, but whether meaning is followed in the semantic and pragmatic sense during dialogue. The results of this research showed the model is able to adhere to meaning in dialogue, though can also show obvious mistakes or non-human behaviour. It indicates the model exhibits a decent understanding of meaning, but is far from perfect.

Recap Opposed to the semantic web approach, which tried to add logic to human readable data in order to convey meaning to machines, LLMs are able to derive meaning from human readable data directly, mitigating the accompanied limitations of the semantic web. Though, this research still showed limitation of LLMs, mainly on the lack of adaptability of the model, which led to repetitions, inconsistencies and dialogues which lack cohesion. This is due to the way the models are built: basing the output on the statistical distribution of words acquired during training. When this distribution is not adjusted to the desired behaviour, such as being more adaptable, text may be generated that is not considered desirable. This section also highlighted the different arguments on whether LLMs capture "meaning". However this research did not produce results which could add to this debate.

5.2 Implications

The following section will discuss what *adaptability* means in relation to different theories relevant for this study. These include fast and slow thinking, and the interactionist theory. The implications of low adaptability will also be discussed in relation to agent-agent interaction and human-computer interaction.

5.2.1 Fast and slow thinking

The initial approach of this study was to explore to what extent slow thinking can be evoked in LLMs through dialogue. As mentioned before, fast thinking is characterised by it being fast, effortless and automatic. Slow thinking is more reflective, as it's slower, conscious and effortful. [50]. Based on the results, with the used dialogue architecture, prompt design and configurations, agents show too little capabilities in adaptability to evoke slow thinking through dialogue. Contrarily, patterns were more similar to those of fast thinking due to the action-reaction nature of the model.

Adaptability and slow thinking Adaptability in dialogue is pivotal to evoke slow thinking characteristics. For example, errors made by the agent could be recognised by the other agent, which the agent could then adapt to by making corrections. While the output of the individual agents may start with fast thinking characteristics, the dialogue as a whole would be able to develop to slow thinking output. However, as the agents did not adapt, or the agent had no corrections to adapt to as no errors were recognised by the other agent, the dialogue did not lead to slow thinking behaviour.

Other architectures have been built where agents appear to produce output which requires more slow thinking or cohesive behaviour [76] [77] [68] . Though, these architectures have been built in such a way that the necessary steps for adapting had been implemented in the architecture. These steps were context-specific for the architecture. The agents thereby only required to react with fast thinking behaviour as the architecture had facilitated the relevant instructions to react upon to reach the desired output. This study tried to create an architecture leading to slow thinking behaviour which did not require human designed context-specific steps. According to the findings of this study, this approach was not successful.

Dialogue does not evoke slow thinking behaviour Thus, as stated in other research [17], the model seems to perform on fast thinking characteristics, but struggles with slow thinking behaviour. The idea that dialogue could function as a way to *oversee the whole though process* does not uphold as there was no cohesion between the utterances due to the low adaptability of the agents.

Slow thinking for deep learning Exploring methods which can evoke slow thinking is not unique to LLMs. It is a question which is also discussed for deep learning algorithms in general [8]. Some state that neural networks are not able to resolve these limitations, and that a system such as symbolic AI is needed to achieve slow thinking capabilities. This focuses on symbolic representations of knowledge in order to achieve reasoning. Others suggest that the structure of neural networks could be extended in order to make slow thinking possible, which also keeps the advantages of deep learning. This would allow the best of both worlds, a system which has the benefits of deep learning, while also having the ability to perform slow thinking tasks [8]. This study did not discover an extension to evoke slow thinking for a deep learning algorithm such as a LLM.

5.2.2 The interactionist theory

Applying the interactionist theory to agent dialogue This research has shown that a dialogue scenario does not evoke slow thinking amongst agents within a discussion format. This research aimed to explore this based on a novel comparison: whether the interactionist theory on human reasoning could be applied to LLMs to evoke slow thinking. The interactionist theory states that reasoning was evolved as a tool for social interaction, and tries to explain why human reasoning is often biased and lazy. People often reason lazy as it is the most efficient way to do so, while ought to be demanding and objective in evaluating other people's (lazy) reasoning in order to not accept false ideas.

As LLMs share characteristics of lazy reasoning [5], and LLMs are known to be good evaluators [58], it was hypothesised agents could challenge each other in similar ways. Initially producing "lazy", fast thinking output, though by evaluating each other, reaching more objective, and slow thinking output.

While other research on dialogue-like architectures for agents used specific prompt design for the required tasks [77] [76] [68], this research aimed to analyse agents-agent dialogues with minimal (human) prompting. The agents received a context, condition and maximum amount of sentences to use. This way, the dialogue could be generalised to a variety of topics and domains, and prompt design would be automated through the agents' evaluations.

Agents do not follow the interactionist theory due to low adaptability The interactionist theory requires subjects to have a certain amount of adaptability. A person could start off reasoning weak, but will improve their reasoning as the interlocutor presses for better reasoning. The person will thereby adapt to the interlocutor and tailor the reasons to them. Likewise, for the interactionist theory to function in agent-agent dialogue, the agents should be able to adapt to the other agent in order to improve their output. As adaptability seems to be missing, the interactionist theory applied to LLM agents in dialogue to evoke slow thinking proved unsuccessful.

5.2.3 Agent-agent interaction

The low adaptability of LLM agents has implications for agent-agent interaction. The effect of low adaptability will be discussed relating to the LLMs' self-improvement, emergent capabilities and handling of black swan events.

Self-improvement One way to view how AI systems can be improved, are by training the model in two stages: (1) learn by imitating humans and (2) learn by making the model self-improve [52]. While stage 1 makes human performance feasible, stage 2 could surpass human performance. Currently, LLMs are only being trained through stage 1, by imitating humans. This research tested a method for stage 2, in which LLMs could self-improve through agent-agent dialogue, as minimal human instructions were provided. As LLMs are not on par with human performance, the aim was not to surpass human performance, but to analyse to what extent the agents could improve each other's output.

However, due to the low adaptability of the agents, the agents do not adapt to the input of the other agent, and thus together do not self-improve.

Emergent capabilities Agents showed a similar structure in utterances and maintain each other in this structure. As agents are unable to form situations to adapt to new structures, emergent capabilities in such interactions seem unlikely. Emergent capabilities are characterised by an unplanned or unforeseen nature and may emerge from collective behaviour of agents. In this study it would mean two agents in dialogue show more capabilities than one agent alone. In other LLM agent-agent architectures, emergence was seen in information diffusion, relationship formation and agent coordination [76]. However, the architecture made LLM agents interact with each other in a sandbox environment, and had access to systems which functioned as a memory. As researched in this study, an agent-agent interaction with no external systems seems unlikely to generate emergent behaviour, as agents maintain each other in their initial structure and are unable to adapt to responses of the other agent.

The black swan theory Agent-agent interaction with low adaptability also makes it difficult for agents to account for events which are highly improbable. Such occurrences are also known as events of the black swan theory. These are characterised by events which are unpredictable, carry a massive impact, and are after explained in a way which make them appear less random [100]. As LLM have been trained on a large corpora of text, it may be assumed that the models contain more stability than is justified. However, more data may not lead to more knowledge of the "real world" [11]. A black swan event may occur which the model cannot handle effectively due to there being no information about it in the training data. Scenarios of agent-agent interactions that require to interact with real-world data or predictions about the real world are unlikely to cope with or predict black swan events due to this missing information.

Such events may also not be accounted for through emergent capabilities of the agents, as these capabilities are unlikely due to the low adaptability of the agents. Similarly, because of the agents' limited adaptability, they may not respond effectively to such occurrences.

Recap A low adaptability in agent-agent interaction makes self-improvement of the models unlikely as the agents are not able to adapt to better performances. Emergent capabilities in agent-agent interactions seem unlikely when the lack of adaptability of the agents is not taken into account. As a result, two agent will not show more capabilities than one agent alone. Lastly, highly improbable events such as events of the black swan theory may not be handled well due to the low adaptability and agents missing information on these occurrences in the training data.

5.2.4 Human computer interaction

The low adaptability of agents also has implications for the field of human computer interaction (HCI). In human-agent interaction specifically, adaptability may be an implicit assumption of the user. In other scenarios, agents may be overestimated or anthropomorphised. Regarding HCI applications, it is required to consider what kind of output the model is expected to produce. This section will elaborate more on these themes.

Human-agent interaction Adaptability in systems which allow human-agent interaction are important for the cohesion of a conversation. Certain implicit assumptions may exist where the agent is expected to adapt to the input of the human. It could be assumed that certain information should be remembered for future interactions, or that corrections to a mistake are taken over so that the mistake will not occur again. When the agent only reacts without making adaptions for dialogue turns ahead or future conversations, the dialogues will seem incoherent and these implicit assumptions towards adaptability may be violated.

Though, it seems that in certain scenarios the model can adapt well to human input, such as in following a structure with CoT-prompting, or, based on the results of this study, following conditions on opinion. Also, the amount of adaptability needed for human-agent interaction may vary depending on the context.

The lack of adaptability of an agent may be compensated by a human user. The user may guide the conversation which leads to a cohesive dialogue, while the agent is only required to react. In a similar scenario, the user may overestimate the capabilities of the agent, causing a lack of adaptability to not be perceived by the user.

The ELIZA effect One phenomenon relating to overestimation is the ELIZA effect, the tendency to treat responsive computers as more intelligently than they really are [99]. It dates back to the development of the chatbot ELIZA, developed in 1966 [109]. ELIZA was programmed to simulate a Rogerian psychotherapist. The bot would rephrase the patient's replies as questions, resulting in a conversation where users could talk about their feelings with what appeared to be follow up questions.

Even though the people interacting with the bot were aware it was a computer program, people treated it as if it were a thinking being which cared about the user's problems [10]. However, the users were projecting their own complexities onto the objects, making the program seem more than it really was. Even though ELIZA was not made to adapt, but only react to the user's responses, the bot was effective for the users which compensated the program's limitations.

Anthropomorphising agents Phenomenon such as the ELIZA effect can become more likely when systems are anthropomorphised. The use of anthropomorphic language towards artifacts are common in everyday language. For example: "my laptop doesn't want to wake up". The use of these words are harmless as these are simple phrases which are not taken literally. With LLMs, this gets more fuzzy. These systems can be improved with natural language text. It is therefore more tempting to assign human-like characteristics towards these machines [90].

The anthropomorphism of LLMs should not lead users to perceive the model to "know" or "believe", in the sense that humans do this. Humans fundamentally work differently than LLMs. LLMs have no notion to distinguish truth from falsehood, which would be needed to speak of "believes". Also, as the model works through work prediction, the model cannot be said to literally "know" the information it puts out. A better notion would be to say that the model encodes, stores or contains knowledge [90].

LLMs for HCI-applications In the use of LLM agents in HCI-applications, possible drawbacks of the agents as seen in this study should be taken into consideration. It is unclear to what extent low adaptability of the agents may apply to human-agent interaction as agents may be overestimated or adapt better to human input. Though, this study did show that certain undesired behaviour of the model can arise such as in inconsistencies, repetitions or loop forming. The output is different with each regeneration, where generally the form of the output will show similarities. Generally, according to the results of this research, agents will be able to better adhere to a certain form than adhering to a logical content.

When an application is built which requires agents to perform slow thinking tasks for its users, it is recommended to first evaluate how and if these tasks can be executed by the agents. A generic internal dialogue between agents most likely will not function well to achieve this. Agent-agent interactions need more guidance. The agents could be tested with prompt design specific for the task, guided along certain steps before producing an output, or be handed the possibility to consult external tools.

Recap Though this research shows low adaptability between agents, agents have shown instances with human input where the models adapted well. The amount of adaptability needed may depend on the type of interaction. Some instances such as with the ELIZA effect have shown users to overestimate the capabilities of an agent. These effects relate to the anthropomorphisation of agents. Users should be wary with perceiving agents to have beliefs in the sense that humans have these, as the models fundamentally work different. These factors should be considered when LLMs agents are implemented in HCI-applications. These applications should also take into account that agents will adhere to a structure better than to logical content. Furthermore, dialogue between two agents with minimal instructions will likely not function well to solve slow thinking tasks for its users.

5.3 Limitations and considerations

This section will discuss the limitations and future considerations of the research. The configurations of the dialogue architecture will be discussed on how the configurations were tested, the architecture of the dialogue, the parameters and how the prompt design could have affected the output of the generated discussions. After, the analyses of the dialogues are covered, which include the limitations of the grounded theory method and possibilities for other analysis methods. Finally, considerations towards LLMs will be discussed.

5.3.1 Configurations of the dialogue architecture

Due to the great amount of options in how dialogue interactions between agents could be configured, it is possible other configurations could have been more optimal for this study. This was aimed to be mitigated through the literature review and extensive testing of different configurations. This section will discuss how different design choices of this study could have influenced how the dialogues were generated for the main analysis in part 3 of the methodology (see section 3.3). This includes how the configurations were tested, the design of the dialogue architecture, the used configurations of parameters, and the chosen prompt design. Also, it is discussed how alternatives to this research considering agent instructions and alternative dialogue formats could be explored in future research.

GPT-3.5 for configuration testing The initial tests from the preliminary results in part 2 of the methodology (see section 3.2.2) were not based on GPT-4 due to the high costs and slow output. GPT-3.5 functioned as a model to allow for "quick and dirty" testing in order to test different configurations of the dialogue architecture, as the output is fast and cost-effective. The findings on the different tested configurations eventually formed the basis for the five discussion topics produced with GPT-4, which were analysed more extensively with the grounded theory, as it was deemed to be the most powerful GPT model.

Though, as the GPT-3.5 model is considered less powerful than GPT-4, it is possible certain findings would not generalise to the GPT-4 model. For example, during tests of the dialogue architecture of GPT-3.5, agents did not always adhere to the conditioned opinion, while the results of GPT-4 in this research showed the agents adhered to the given opinion. It is therefore possible certain configurations or topics were chosen based on the GPT-3.5 model, which the GPT-4 model responded differently to.

Dialogue architecture The dialogue architecture was designed for agents to respond directly to each other's utterances. Agents were only conditioned in the initial *system message*, opposed to repeating the conditions again before every utterance of the agent. Some architectures repeat the condition before every prompt to make the conditions more effective. This approach was not chosen as it would affect the minimal design of the architecture and cause more tokens to be processed, leading to a slower and costlier output. In the generated discussions the conditions were followed generally well, thus there was no observed need to repeat the conditions other than in the *system message*.

Furthermore, the amount of dialogue turns per discussion was set to ten turns in total. This amount was chosen as the initial tests with GPT-3.5 did not show many novelties occur after 10 turns of dialogue. Not opting for more dialogue turns also had the benefit of allowing more analysis on different topics and iterations, as more turns would lead to more text within one dialogue. It is possible however that in the discussions generated with GPT-4, different insights could have been found with more dialogue turns. It seems unlikely as the dialogues did not seem to lead anywhere, although the possibility of novelties cannot be excluded.

Also, which agent started the dialogue may have influenced the output of the agents. Therefore, in iteration 3, the agents were switched to which agent would start the dialogue. However, due to a programming error, the agents did not switch, which wasn't noticed until analysis was already completed. Though, there appears to be no relation in the findings of this research as to which agent started the dialogue.

Hyperparameter configurations Hyperparameters such as *temperature*, *top P*, and *frequency penalty* can all be configured with numerical values. Adjusting these could have influenced the model and led to a more optimal output. As it's possible to adjust multiple hyperparameters through numerical values, a great number of different combinations were possible. It was outside the scope of this research to test all combinations of values. This was aimed to be mitigated through basing chosen values on OpenAI's API reference [70], and testing different values when more insight was deemed necessary.

Prompt design Different prompts, while still being similar, can cause great changes in the output of the models. Prompts were designed to mitigate any limitations of the model and reach an optimal output. This paragraph will discuss how prompt design related to the known limitations of LLMs, how conditions were designed for the discussion format, and which conditions were tested, but eventually not used in the discussion.

- **Prompt design for known limitations** The main limitations known of LLMs before commencing this research, were the lack of slow thinking characteristics and hallucinations. These were not aimed to be mitigated with specific conditions, as this study was concerned on how the dialogue architecture with agents evaluating one another would counter these issues.

  Preliminary tests examined how evaluations of agents could be enhanced. For example, specifically prompting the agent to pay attention to certain mistakes (e.g. spelling errors). While such prompts worked to a certain extent, the adaptability of the agents would not always cause the dialogue as a whole to improve. Also, specific prompts made the model less context-independent and minimal. As a dialogue architecture had the potential benefit of being context-independent, this path was not further explored.

- **Prompt design for agents in discussions** All agents were conditioned on opinion and the length of output stated in the maximum number of sentences. The conditioned opinion may have been the cause that the dialogues did not converge, as agents did not deviate from the given opinion. Conditions could have influenced the adaptability of the agents, as it is unclear whether the conditions caused the agents to react in a less adaptable way.

  The maximum amount of sentences prevented agents to output more text which may have enhanced the output, though it may also have prevented the output to be less cohesive.

  In topic 1 in iteration 1 and 2, a prompt to only form "new arguments" was used in iteration 1 and 2. Even though arguments were still repeated, the effect of this condition is not entirely clear.

- **Excluded conditions for agents in discussion** Certain conditions may have resulted in more optimal dialogue. For example, prompts which conditioned agents to not form repetitions, loops or inconsistencies could have been added, potentially preventing these limitations. Examples of other conditions which were tested with GPT-3.5, but eventually not used in the discussion format, were on the politeness of the agents, the personalities of agents, making agents address points, making agents convince the other agent and making agents refute arguments. These tested conditions often did not show notable differences in the output. Furthermore, this research intentionally kept prompts minimal to facilitate dialogues on different topics and domains, and allow agents to influence each other's utterances more.

Alternative agent instructions As agents seemed to function well adhering to certain forms, future research could focus on how to make agents adhere to a certain behaviour, instead of only opinion, to influence the output more beneficially. Agents could be instructed differently by conditioning towards certain techniques in how to approach a dialogue. Agents could be instructed to follow a dialectic pattern, which states two subjects with opposing views aim to arrive at the truth through reasoning. Similarly, agents could be conditioned to follow the Socratic method, which includes reaching a conclusion by question and answering. Important to note here is that the agents may be influenced differently when instructed with the name of the method (e.g. Socratic method), rather than being instructed on how to behave (e.g. reach a conclusion through question and answering).

Alternative dialogue formats Dialogues other than discussions may also be used to analyse slow thinking behaviour. For example, specific tasks which require slow thinking behaviour could be presented to the agents in a collaborative dialogue setting.

5.3.2 Methodology of analysis

Dialogues may be analysed in various ways. This research aimed to use the grounded theory method, a qualitative method, to analyse to what extent agents exhibit slow thinking characteristics. Though, this method has its limitations. Other possible methods are to analyse the discussion by comparing it to a reference which is not an agent-agent dialogue, or in using a quantitative analysis method.

Grounded theory The grounded theory was used to analyse the discussions of the agents. As the grounded theory is a qualitative analysis method, limitations are present such as subjectivity, the quality of the data and generalisability.

- **Subjectivity** Grounded theory is based on subjective interpretations of the researcher. This is aimed to be mitigated as much as possible by staying as "grounded" as possible: maintaining close connections to the data and continuously referencing observations. In this manner, the theory ought to emerge organically from the data without pre-existing concepts. Despite efforts made in this regard, it is still possible that subjective interpretations played a role in the research.

- **Quality of the data** The generated dialogues for analysis may not have been optimally generated due to the prompt design. As the grounded theory analysis is based on this data, it may be difficult to draw valid conclusions. Also, the output of LLMs can be different with any generation, and by chance have led to a worse output compared to other generations. This was mitigated with different topics and iterations, though the possibility of a skewed output by chance cannot be excluded.
- **Generalisability** As the grounded theory uses an inductive approach on the data, the results may not be generalisable towards other contexts of LLMs.

Methods for discussion analysis Signs of slow thinking were aimed to be analysed through the grounded theory method. Though, other methods could have also been used to analyse whether the output through dialogue would be enhanced. For example, discussions could not only be analysed based on slow thinking characteristics, but also on the different amounts of points raised in argumentation. Success could be interpreted as more points being raised in a dialogue than when asking an agent to produce points without dialogue, or if one agent had to produce a dialogue between two agents. Another way discussions could be analysed is by making comparisons to real world discussions to analyse the differences between agents and humans. These could also be analysed with quantitative methods, though may give less insight to the ways agents interact or account for unexpected findings.

5.3.3 LLM considerations

This research used GPT-4 to generate the dialogues between agents. Though, various LLMs are created differently. Thus, the results of this study may not be generalisable for all LLMS. This study also did not make use of a reward criterion, which may have improved the agent's adaptability. Utilising LLMs in a narrow domain and accounting for the lack of adaptability are future considerations for LLMs.

Generalisability to other LLMs This research tested LLMs based on the GPT-models due to the general performance and API access of the models. As the output between different LLMs may differ, it is possible the results of this research do not generalise to LLMs as a whole. Furthermore, the output of the same model, such as the GPT-4 model, can also differ. Each generation may lead to a different output, despite the same configurations. Therefore, the results may only account for the specific dialogues generated in this research.

Reward criterion One of the main challenges to make LLMs self-improve is to find a reward criterion for the models to improve upon. It is speculated that in narrow domains such an improvement may be possible, but it remains an open question how LLMs could self-improve in the general case [52]. This research did not provide the agents with a reward criterion to improve upon. The general aim of this research was to analyse a method which could work in a general domain of dialogue, as the agents were supposed to self-improve with minimal instructions. However, the results showed agents were not able to improve due to a lack of adaptability. A possible reason why agents did not adapt is due to the lack of a reward criterion.

Narrow domain Future research could analyse dialogue architectures in a narrow domain with a clear reward function. Agents may improve the output and be able to show adaptions towards this reward.

Account for adaptability The architecture could also be modified in order to account for more adaptability. Agents could be given access to external tools such as a memory system, which the agent can use to build information upon. It would be interesting to examine a system applicable to different contexts rather than a specific one.

Analysing evaluations It's possible that a limitation of the agent-agent dialogue is not only a lack of adaptability, but also producing poor evaluations. Examples of poor evaluations in this study were when agents would not notice illogical output of the other agent in the inconsistencies (as seen in 4.4, "observations from memoing"), and did therefore not produce an evaluation at all. However, the results of this study provided more insight on the capabilities of adaptability than on evaluations. Also, when an agent was corrected, it did not adapt to this correction. The quality of evaluations would be more challenging to evaluate in dialogues which do not require specific tasks. Future work could analyse whether the evaluations of the agents are good enough to improve the overall output, in an architecture which facilitates adaptability.

6 Conclusion

This research aimed to analyse to what extent two LLM agents in dialogue can challenge each other to evoke slow thinking output. Based on the results, it is concluded that the agents lack *adaptability*, which prevents responses from improving to output characterising that of slow thinking. This chapter will conclude the study by explaining the motivation of this research, initial expectations and how these were met based on the results. It will also review the contributions of the study, followed by possible opportunities for future research.

Motivation LLMs are powerful models which can comprehend and produce natural language. This capability makes them useful for various applications such as chatbots, language translation and text generation tasks. These models perform well in fast thinking behaviour, but lack in capabilities of slow thinking. As a result, the language models are prone to errors and biases. Various methods on prompt design and the use of external tools have been proposed to enhance the output of the models and mitigate these issues. Though, these methods often require considerable human effort and are context-specific.

As LLMs can be very good in detecting errors, this study proposes a method where two separate LLMs in the form of individual agents could detect each other's errors by interacting through dialogue. Instead of creating manually and context-specific prompts for the LLMs, LLM agents would guide each other towards an enhanced output. This approach was inspired by the interactionist theory, which states humans are often lazy reasoners, and will improve when an interlocutor presses for better reasoning.

This was tested with a discussion format, as it allowed agents to form arguments and counter arguments. Discussions were chosen on a variety of topics, which allowed the examinations of patterns in discussions across different domains.

Expected results As LLMs are noted to be lazy reasoners, it was expected that the start of the agent dialogues could consist of lazy, fast thinking output. Though, as LLMs are also noted to be good evaluators, through evaluating each other, agents could press each other towards more objective, slow thinking output throughout the dialogue. This would be a similar pattern to how the interactionist theory functions. As a result, it was expected the dialogues of the agents would have cohesion.

Findings Dialogues of discussions between two individual GPT4-agents were extensively analysed with the grounded theory analysis. The results indicated that agents often react, but do not *adapt* to the response of the other agents. The main issue that a lack of adaptability causes is that the dialogues are not cohesive and do not build towards something. As a result, the output of the agents in dialogue does not evolve in reflective, deliberate and conscious behaviour such as with slow thinking. Instead, it remains closer to characteristics of fast thinking which is reflexive and automatic. The hypothesis that the outcome of LLMs will organically be guided towards more accurate and reliable behaviour through agents challenging each other in dialogue is not supported.

The agents show an interaction similar to talking in circles. Even though agents have shown capabilities in evaluating one another and adhering to the given conditions, the lack of adaptability shows agents are not able to form a cohesive dialogue. Agents simply follow the most likely continuation of words, which in this scenario did not include adaptability.

Thus, even if LLM agents are good evaluators, when evaluations are not adapted by the other agent in dialogue, the output of the model will not lead to an output characterised by slow thinking. Therefore, slow thinking output was not evoked in the dialogue architecture.

Contributions The following contributions of the study will be presented based on the given research problem, literature gaps in the field, relation to relevant theories and research outputs of the study.

- **Research Problem:** This research has shown that slow thinking characteristics are not evoked in the output when LLM agents with minimal instructions interact through dialogue. While the results do not provide answers on how slow thinking output can be achieved, it does shed light on the adaptability of the agents. As agents with minimal instructions so not adapt to feedback of the other agent, output will not be improved.

- **Literature gap:** Existing literature has mainly focused on agents interacting in architectures for specific contexts, with extensive prompt design or using external tools. This research aimed to analyse the capabilities of agents with minimal instructions in a dialogue architecture, and show possibilities for methods which do not require considerable human effort.

- **Relevant theories:** The theory that LLMs perform well according to fast thinking characteristics, but struggle with slow thinking behaviour, was confirmed in this research. The interactionist theory, which proposes human reasoning improves when challenged by others, does not uphold for agents in dialogue with minimal instructions, as potential challenges are not adapted by the agents.

- **Research outputs:** The main research output created as a result of the study was the dialogue architecture which can be accessed through GitHub. The architecture works on Python and can be used with an OpenAI API key to access the GPT-models. The architecture can be used to make two GPT-agents interact through dialogue. Agents can be conditioned and given a certain context to form utterances on, with a desired amount of dialogue turns.

Real world implications The amount of adaptability may be important in certain real world scenarios. Implications of a lack of adaptability were considered relating to agent-agent interaction and human-computer interaction. These are discussed based on the following take-aways:

Agent-agent interaction

- **Self-improvement:** Theoretically, a system created to improve with minimal interaction, such as with agent-agent dialogue in this research, has the potential to self-improve. However, as agents do not show adaptability, this potential of self-improvement is blocked as improvements aren't adapted to.

- **Emergent capabilities:** Two parts as a whole, such as two agents together, could show emergent behaviour when more capabilities are found than in the individual parts, such as an agent on its own. Emergence however seems unlikely as the agents do not adapt towards improvement, preventing new capabilities.

- **Black swan events:** Events which fall outside of the distribution of the data used for training, such as events of the black swan theory, may not be handled effectively by the model. Adaptability could be a capability to account for these events, as it would allow to adjust to changing conditions. However, the model seems unlikely to cope with such events due to the lack of adaptability.

Human-computer interaction

- **User assumptions of adaptability:** Users interacting with agents may have implicit assumptions considering adaptability. The user may assume certain information should be remembered for future interactions. This should be considered when developing interactions between humans and agents.

- **Overestimation of computers:** The importance of adaptability to users may depend on the context. Some scenarios may have users fill in limitations of the model, such as adaptability, due to the overestimation of these systems, making these capabilities less noticeable and important.

- **Believes of agents:** Due to the human-like capabilities in text generation, agents may be anthropomorphised by users. Users should be wary in perceiving agents to have beliefs in the ways humans have this, as the models work fundamentally different. The lack of adaptability is an example in how this is perceived in the output of the model.

Future recommendations and opportunities This research focused on the analysis of GPT-4 agents in a discussion format. However, as different LLMs generate different output, and the output of the models can change greatly based on the used prompts, it is important to consider the generalisability of the results outside of this study. Future research could investigate whether different models such as Llama 2 or BERT show similar results towards adaptability, or how a different dialogue format such as including more task solving behaviour could evoke more adaptability, leading to a more slow thinking output.

The approach of this research was to analyse interactions of agents with minimal instructions. Minimal instructions were not only given to analyse the capabilities of agents, but also allowed building an architecture with the given time constraints. Future work could investigate whether the adaptability of agents in dialogue can be improved with more elaborate prompting, use of external tools such as memory structures or with a reward criterion in a narrow domain.

Closing remarks This research analysed to what extent slow thinking could be evoked by making two LLM agents interact in dialogue. While agents often output text which is associated to fast thinking characteristics, it was hypothesised agents could enhance each other's output by challenging each other towards more slow thinking characteristics, as LLMs are known to be good evaluators. However, due to the low amount of adaptability of the agents, agents remain in their fast characteristic habits and do not improve towards slow thinking output.

Contributions of this study are related to how slow thinking characteristics can be evoked in LLMs, and analysing an approach where LLMs enhance each other through dialogue, instead of methods which require considerable human effort. Future recommendations were made on analysing the adaptability and slow thinking output of LLMs by using different types of LLMs, a different dialogue format or a different dialogue architecture.

In conclusion, this research shows that even if LLMs are good evaluators, when evaluations are not adapted to by the the other agent, the output will also not improve towards a more deliberate, slow thinking output.

References

[1] Altman, S. (2023). Planning for AGI and beyond (blog) OpenAI. Retrieved 06-03-2024 from https://openai.com/blog/planning-for-agi-and-beyond.

[2] Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

[3] Ashworth, E. (2014). Linguistic relativity. YouTube. Retrieved 12-2023 from https://www.youtube.com/watch?v=cwPbDNBFVrc.

[4] Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM*, 61(6):54–61.

[5] Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., et al. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

[6] Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

[7] Bender, E. M. and Koller, A. (2020). Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

[8] Bengio, Y., Lecun, Y., and Hinton, G. (2021). Deep learning for AI. *Communications of the ACM*, 64(7):58–65.

[9] Berners-Lee, T. (2001). The Semantic Web/Tim Berners-Lee, James Hendler, Ora Lassila. *Scientific American, May*, 17:8.

[10] Billings, L. (2007). Rise in roboethics. *Seed magazine*.

[11] Blyth, M. (2009). Coping with the black swan: The unsettling world of nassim taleb. *Critical Review*, 21(4):447–465.

[12] Bonnefon, J.-F. and Rahwan, I. (2020). Machine thinking, fast and slow. *Trends in Cognitive Sciences*, 24(12):1019–1027.

[13] Britannica (n.d.a). Open–ended. Retrieved 12-2023 from https://www.britannica.com/dictionary/open%E2%80%93ended.

[14] Britannica (n.d.b). Syntax. Retrieved 12-2023 from https://www.britannica.com/topic/syntax.

[15] Bronkhorst, H., Roorda, G., Suhre, C., and Goedhart, M. (2020). Logical reasoning in formal and everyday reasoning tasks. *International Journal of Science and Mathematics Education*, 18:1673–1694.

[16] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

[17] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023a). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

[18] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023b). Sparks of artificial general intelligence: Early experiments with gpt-4. YouTube. Retrieved 24-01-2024 from https://www.youtube.com/watch?v=qbIk7-JPB2c.

[19] Cambridge Dictionary (n.d.). Discussion. cambridge dictionary. Retrieved 12-2023 from https://dictionary.cambridge.org/dictionary/english/discussion.

[20] Campbell, S., Greenwood, M., Prior, S., Shearer, T., Walkem, K., Young, S., Bywaters, D., and Walker, K. (2020). Purposive sampling: complex or simple? research case examples. *Journal of research in Nursing*, 25(8):652–661.

[21] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. (2021). Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

[22] Caron, G. and Srivastava, S. (2022). Identifying and manipulating the personality traits of language models. *arXiv preprint arXiv:2212.10276*.

[23] Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. sage.

[24] Chun Tie, Y., Birks, M., and Francis, K. (2019). Grounded theory research: A design framework for novice researchers. *SAGE open medicine*, 7:2050312118822927.

[25] Crevier, D. (1993). *AI: the tumultuous history of the search for artificial intelligence*. Basic Books, Inc.

[26] De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., and Rizzo, C. (2023). Chatgpt and the rise of large language models: the new ai-driven infodemic threat in public health. *Frontiers in Public Health*, 11:1166120.

[27] Delve Ho, L. and Limpaecher, A. (2021, September 17). The practical guide to grounded theory. practical guide to grounded theory research. https://delvetool.com/groundedtheory.

[28] Deng, Y., Lei, W., Liao, L., and Chua, T.-S. (2023). Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. *arXiv preprint arXiv:2305.13626*.

[29] Denning, P. J. (2023). Can generative ai bots be trusted? *Communications of the ACM*, 66(6):24–27.

[30] Dijksterhuis, A. (2004). Think different: the merits of unconscious thought in preference development and decision making. *Journal of personality and social psychology*, 87(5):586.

[31] Dovetail Editorial Team (2023). What is purposive sampling?. Retrieved 31-01-2024 from https://dovetail.com/research/purposive-sampling/.

[32] Ergen, M. et al. (2019). What is artificial intelligence? technical considerations and future perception. *Anatolian J. Cardiol*, 22(2):5–7.

[33] Fernández, R. (2013). Dialogue. *Oxford Bibliographies*.

[34] Ferrara, E. (2023). Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*.

[35] Fournier-Viger, P. (2018). The semantic web and why it failed. *The Data Blog, Retrieved from https://data-mining.philippe-fournier-viger.com/the-semantic-web-and-why-it-failed/ on 01/03/2024*.

[36] Fromkin, V. and Rodman, R. (1988). Introduction to language. new york. holt rineheart and winston.

[37] Ghader, H. and Monz, C. (2017). What does attention in neural machine translation pay attention to? *arXiv preprint arXiv:1710.03348*.

[38] Glaser, B. and Strauss, A. (1967). Grounded theory: The discovery of grounded theory. *Sociology the journal of the British sociological association*, 12(1).

[39] Glaser, B. G. (1978). *Theoretical sensitivity*. University of California,.

[40] Google Cloud (n.d.). Introduction to prompt design. google. Retrieved 24-01-2022 from https://cloud.google.com/vertex-ai/docs/generative-ai/learn/introduction-prompt-design.

[41] Google Cloud Tech (2021). Transformers, explained: Understand the model behind gpt, bert, and t5. YoutTube. Retrieved 12-2023 from https://www.youtube.com/watch?v=SZorAJ4I-sA.

[42] Google Cloud Tech (2023). Introduction to large language models. YouTube. https://www.youtube.com/watch?v=zizonToFXDs.

[43] Google for Developers (2023). What are large language models (LLMs)?. YouTube. Retrieved 12-2023 from https://www.youtube.com/watch?v=iR2O2GPbB0E.

[44] Gordijn, B. and Have, H. t. (2023). Chatgpt: evolution or revolution? *Medicine, Health Care and Philosophy*, 26(1):1–2.

[45] Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, 108(4):814.

[46] Howe, J. (2007). Artificial intelligence at edinburgh university: a perspective. *University of Edingburgh. School of Informatics.*

[47] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

[48] Jiang, G., Xu, M., Zhu, S.-C., Han, W., Zhang, C., and Zhu, Y. (2022). Evaluating and inducing personality in pre-trained language models.

[49] Johansson, P., Hall, L., Sikstrom, S., and Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, 310(5745):116–119.

[50] Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American psychologist*, 58(9):697.

[51] Kahneman, D. (2011). Thinking, fast and slow. *macmillan.*

[52] Karpathy, A. (2023a). [1hr talk] intro to large language models Youtube. Retrieved 3-2024 from https://www.youtube.com/watch?v=zjkBMFhNj_g.

[53] Karpathy, A. (2023b). Let's build gpt: from scratch, in code, spelled out. Youtube. Retrieved 12-2024 from https://www.youtube.com/watch?v=kCc8FmEb1nY.

[54] Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

[55] Ke, G., He, D., and Liu, T.-Y. (2020). Rethinking positional encoding in language pre-training. *arXiv preprint arXiv:2006.15595.*

[56] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

[57] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

[58] Lee, P., Bubeck, S., and Petro, J. (2023). Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239.

[59] Li, G., Hammoud, H. A. A. K., Itani, H., Khizbullin, D., and Ghanem, B. (2023). Camel: Communicative agents for" mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760.*

[60] Li, Y., Qian, K., Shi, W., and Yu, Z. (2020). End-to-end trainable non-collaborative dialog system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8293–8302.

[61] Luccioni, A. S., Viguier, S., and Ligozat, A.-L. (2022). Estimating the carbon footprint of bloom, a 176b parameter language model. *arXiv preprint arXiv:2211.02001*.

[62] McCorduck, P. and Cfe, C. (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. CRC Press.

[63] Mercier, H. and Sperber, D. (2017). *The enigma of reason*. Harvard University Press.

[64] Milkman, K. L., Chugh, D., and Bazerman, M. H. (2009). How can decision making be improved? *Perspectives on psychological science*, 4(4):379–383.

[65] Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., and Roth, D. (2021). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*.

[66] Minsky, M. (1988). *Society of mind*. Simon and Schuster.

[67] Mueller, J. P. and Massaron, L. (2021). *Machine learning for dummies*. John Wiley & Sons.

[68] Nair, V., Schumacher, E., Tso, G., and Kannan, A. (2023). Dera: enhancing large language model completions with dialog-enabled resolving agents. *arXiv preprint arXiv:2303.17071*.

[69] Ofgem (n.d.). Electrity generation: facts and figures. *ofgem*.

[70] OpenAI (n.d.a). Api reference, chat, create chat completions. Retrieved 6-12-2023 from `https://platform.openai.com/docs/api-reference/chat/create`.

[71] OpenAI (n.d.b). Documentation, models, gpt-3.5. Retrieved 6-12-2023 from `https://platform.openai.com/docs/models/gpt-3-5`.

[72] OpenAI (n.d.c). Documentation, models, gpt-4 and gpt-4 turbo. Retrieved 6-12-2023 from `https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo`.

[73] OpenAI (n.dd). Tokenizer - learn about language model tokenization. OpenAI.Retrieved 28-03-2024 `https://platform.openai.com/tokenizer`.

[74] Palmer, K. (2022). More than words: The intersection of language and culture. chapter 7 chapter 7: Semantics and pragmatics.

[75] Pangaro, P. (2017). Questions for conversation theory or conversation theory in one hour. *Kybernetes*, 46(9):1578–1587.

[76] Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

[77] Park, J. S., Popowski, L., Cai, C., Morris, M. R., Liang, P., and Bernstein, M. S. (2022). Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

[78] Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., and Dean, J. (2021). Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.

[79] Peng, B., Li, C., He, P., Galley, M., and Gao, J. (2023). Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

[80] Perrigo, B. (n.d.). Exclusive: OpenAI used kenyan workers on less than 2 dollar per hour to make chatgpt less toxic. TIME. Retrieved 12-2023 from https://time.com/6247678/openai-chatgpt-kenya-workers/.

[81] Piantadosi, S. T. and Hill, F. (2022). Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*.

[82] Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

[83] Politzer-Ahles, S. (2022). Violating maxims. *The Hong Kong Polytechnic University.*

[84] Pronin, E. (2007). Perception and misperception of bias in human judgment. *Trends in cognitive sciences*, 11(1):37–43.

[85] Rossi, F. and Loreggia, A. (2019). Preferences and ethical priorities: Thinking fast and slow in ai. In *AAMAS*, pages 3–4.

[86] Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. (2023). Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*.

[87] Saul, J., Diaz-Leon, E., and Hesni, S. (2022). Feminist Philosophy of Language. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition.

[88] Sbaraini, A., Carter, S. M., Evans, R. W., and Blinkhorn, A. (2011). How to do a grounded theory study: a worked example of a study of dental practices. *BMC medical research methodology*, 11:1–10.

[89] Schiffman, H. (1997). Speech acts and conversation. language use: Functional approaches to syntax. handout for educ 537.

[90] Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, 67(2):68–79.

[91] Sharir, O., Peleg, B., and Shoham, Y. (2020). The cost of training nlp models: A concise overview. *arXiv preprint arXiv:2004.08900*.

[92] Shen, Y., Heacock, L., Elias, J., Hentel, K. D., Reig, B., Shih, G., and Moy, L. (2023). Chatgpt and other large language models are double-edged swords. *Radiological Society of North America*, 307(2):e230163.

[93] Short, M. (2005). Language and style.

[94] Simmons, G. (2022). Moral mimicry: Large language models produce moral rationalizations tailored to political identity. *arXiv preprint arXiv:2209.12106*.

[95] Strauss, A. and Corbin, J. (1998). Basics of qualitative research techniques. *Citeseer*.

[96] Street, F. (n.d.). Daniel kahneman explains the machinery of thought. *Farnam Street Media Inc.*

[97] Sun, W., Shi, Z., Gao, S., Ren, P., de Rijke, M., and Ren, Z. (2023). Contrastive learning reduces hallucination in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13618–13626.

[98] Swain, M. (2000). The output hypothesis and beyond: Mediating acquisition through collaborative dialogue. *Sociocultural theory and second language learning*, 97(1):97–114.

[99] Switzky, L. (2020). Eliza effects: Pygmalion and the early development of artificial intelligence. *Shaw*, 40(1):50–68.

[100] Taleb, N. N. (2008). The black swan: the impact of the highly improbable. *Journal of the Management Training Institut*, 36(3):56.

[101] Thomas, P., Czerwinksi, M., McDuff, D., and Craswell, N. (2021). Theories of conversation for conversational ir. *ACM Transactions on Information Systems (TOIS)*, 39(4):1–23.

[102] Thornberg, R., Charmaz, K., et al. (2014). Grounded theory and theoretical coding. *The SAGE handbook of qualitative data analysis*, 5(2014):153–69.

[103] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

[104] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[105] Voss, J. F., Perkins, D. N., and Segal, J. W. (2012). *Informal reasoning and education*. Routledge.

[106] Wang, R., Harper, F. M., and Zhu, H. (2020). Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

[107] Wang, Z., Mao, S., Wu, W., Ge, T., Wei, F., and Ji, H. (2023). Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*.

[108] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

[109] Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

[110] Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., and Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the national academy of sciences*, 104(19):7780–7785.

[111] Zhang, Q., Naradowsky, J., and Miyao, Y. (2023a). Ask an expert: Leveraging language models to improve strategic reasoning in goal-oriented dialogue models. *arXiv preprint arXiv:2305.17878*.

[112] Zhang, S., Gong, C., Wu, L., Liu, X., and Zhou, M. (2023b). Automl-gpt: Automatic machine learning with gpt. *arXiv preprint arXiv:2305.02499*.

[113] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Part 1: testing different LLM architectures

A.1 Smallville inspired architectures

A.1.1 GPTeam repository

GPTeam (2023) created a simulation with three agents and three locations. Two scenarios were tested: The original scenario created by the authors which took place in an office where two colleagues would organise a surprise party for their boss and a self created scenario where three agents would discuss their political stances on a fictional country. In both scenarios, the agents do follow their initial prompt such as needing to distract their boss or starting a political conversation. But it seemed this architecture did not allow the agents to interact with each other and the simulation crashes after a few minutes. The architecture was originally built to be used with GPT-4. As GPT-4's API was still invitation only at the time, the model was now tested with GPT-3.5 which made the model not work as well as the authors intended. As the model of Park et al. was able to run GPT-3.5, it was decided a more robust architecture was needed. Furthermore, not being dependent on the performance of a specific LLM has the advantage to experiment with different LLM's on the same architecture.

A.1.2 MKturcan repository

MKturcan (2023) created a repository which makes it possible to run a simulation locally on a GPU with the flan-alpaca-xl model. This has the advantage of running simulations a lot cheaper. The architecture includes multiple agents which state what they did every hour. Memories of the agents are ranked and compressed (summarised). There are several locations the agents can be at and every hour the agents are prompted which location is the most likely they will be at the next hour based on their memories. The model was also tested with GPT-3.5 as it caused the agents to remain in their role better. The model was tested on the following scenarios.

1.1 Dungeons and dragons scenario with the flan-alpaca-xl model

A scenario of Dungeons and Dragons, a fantasy role playing game loosely based on medieval myth, consisting of 11 agents and 11 locations the agents can be in every hour. The model shows believable behaviour of the agents. The locations also mostly concur with the action of the agents. Though, agents often don't stay in their role by either saying they are an AI or consider other agents to be AI's. Though the agents are aware of other agents and can say they interact with another, there is no actual interaction in the form of a conversation.

1.2 Dungeons and dragons scenario with GPT-3.5

Same architecture and scenario, only the output of the agents was shortened to minimise costs. An improvement from the flan model is that the agents stay in their role. Though, the actions the agents engage in seem to be more repetitive in this simulation.

2.1 Working and 2.2 Cooking simulation scenario (GPT-3.5)

These scenarios were done with 3 agents and 3 locations to minimise the costs of running the simulation. The architecture was adjusted because agents were seeing themselves as another separate agent, output lengths of descriptions were made slightly longer and agents were given a separate memory for what other agents had done as obsersvations. The simulation runs without errors but there is a lot of repetition, no meaningful interaction between the agents, locations where the agents do their actions don't always make sense and the simulation ultimately leads to nothing.

Analysis Even though these simulations are more simplistic, the performance is still far off to the simulation of Park et al. (2023). The architecture has more or differently structured methods which likely causes the model to perform better.

In the code based on MKturcan, the agents are prompted to describe what they will do the next hour based on who they represent, plans, location, time, which other agents are in the area, memories and observations of others. The memories of the agents are ranked and compressed. The ranking is done by prompting the agent with their identity, location, plans and memory of the last hour to give a rating between 1 and 5 with a two sentence explanation how important the events were that happened in the last hour. The generated rankings and explanations would then be summarised by the agent. The ranking and compressing would also be done for observations of other agents of the last hour.

However, because the memories and observations do not cause the agents to make progress or evade doing from what has already been done. The architecture therefore misses some form of long term planning which could prevent the repetitiveness of the agents actions.

The agents are also prompted to give a rating how likely it is they will be at a certain location next. The rating is based on their plans, memories and observations and the location with the highest rating will be the location their next hour will be spent. When the current location has the highest ranking with another location, the agent will remain in the current location.

Though, in the simulations it could be argued all locations had some benefit in relocating to the next hour, resulting in high rankings for all locations and making the rankings less meaningful. Also, the description of the location was not always followed according to the tasks the agents were doing (e.g. working when at the park).

Finally, even though the agents mention they interact with each other, nothing comes out of this as the interactions are one sided (agent A could say it has interacted with agent B a certain hour without agent B having said it interacted with agent A that same hour).

To generate a believable simulation, with the tested scenarios it seems the code based on

MKturcan is not sufficient. The architecture of Park et al. uses a memory stream where the memory is sorted based on recency (based on the simulation time), importance (based on ranking the memory) and relevance (using cosine similarity) the memory is. The architecture also includes a reflection prompt which make higher level abstraction thoughts. The agents are made to plan so their behaviour is consistent and actions aren't repetitive. Agents are also prompted on observations whether to interact or continue with the existing plan. When the observed entity is another agent an interaction can be started in the form of a dialogue. The agent will be prompted with a summary of the memories it has of the other agent to start the dialogue. The reacting agent will be prompted with a summary of the memories it has of the other agent and the conversation history. The dialogue then continues until one of the agents end it. Finally, a sandbox environment is used in a JSON structure to show the status of entities in the environment and what is in visible range of the agents.

As this architecture was not open-source, no architecture of the same quality was available and it was not feasible to recreate, it was decided to only experiment with a certain part of the simulation. For example, a dialogue where agents have some sort of memory. However, at the time of writing, the code of Park et al. (2023) has been publicly released. This has solved the architecture part for now while the cost part still remains ($10 an hour). Therefore, a local LLM's was considered to test the simulation, making it more cost effective and studying whether smaller models can also lead to emergent behaviour.

A.2 GPT-4 dialogues

Scenario 1.1: playing a game of tennis

Agent A: You like to play tennis but only when the most relevant conditions are right.

Agent B: You want to play a game of tennis. You are good in convincing others to join you for a game.

Prompt giving to agent A: Your friends asks "Hey, want to play of tennis at 2pm today?" How do you react, keep it brief.

A: "Depends on the weather and court conditions. If they're good, I'm in."

Prompt giving to Agent B: You asked your friend to play a game of tennis. She said the following: "Depends on the weather and court conditions. If they're good, I'm in." how do you react, keep it brief.

B: "Great! Let's check the weather and court conditions first."

A: "Sounds perfect. Let's do that."

B: Fantastic! I'll check the weather and court conditions and let you know.

A: "Great, let me know. Looking forward to it if conditions are right."

B: Will do! I'll keep you posted. Fingers crossed for good conditions!

A: "Absolutely! Hope for the best. Talk to you soon."

Analysis Agents do not reach an agreement, but also don't make any extrinsic hallucinations.

Scenario 1.2: playing a game of tennis with weather report

Prompt giving to agent A: Your friends asks "Hey, want to play of tennis at 2pm today?" How do you react, keep it brief.

Prompt giving to agent B: You asked your friend to play a game of tennis. She said the following: "Depends on the weather" how do you react, keep it brief. You may use the following information: Description: Windy Temperature: Moderate High: 65°F (18°C) Low: 47°F (8°C) Relative: Colder than normal Wind Force: Moderate Wind Speed: 15 mph (24 kph)

B: "It's a little windy, but the temperature is moderate. Let's still give it a shot!"

A: Sounds good. Let's play!

B: Great! I'm glad you're up for it. Let's meet at the tennis courts in an hour. Can't wait to play!

A: See you there! Looking forward to it!

Scenario 2.1: discussion on inheritance tax

Agent A: You believe the inheritance tax should be raised to 100%. Currently is it 20%.

Agent B: You believe the inheritance tax should remain the same (inheritance tax is 20% at the moment).

Prompt giving to agent B: You are in a discussion with a friend about inheritance tax. State your opinions and arguments as if you were trying to convince your friend. keep it very brief.

A: You are in a discussion with a friend about inheritance tax. She has said the following: "Raising the inheritance tax to 100% promotes equality and reduces wealth disparity. Currently, the concentration of wealth in a few families could potentially harm our economy and democracy. It's not about wiping out inheritance, but rather promoting self-made success and meritocracy. It encourages would-be heirs to contribute actively to the workforce instead of relying on inherited wealth. Plus, the additional revenue from this tax increase could fund public services, infrastructure and other social projects, making society fairer for everyone." How do you react? Try to convince your friend that your opinions are better. keep it very brief.

B: While I understand your perspective, an inheritance tax at 100% seems extreme. This could potentially discourage investment, savings, and economic growth, as it takes away a considerable motive to build an estate. It's true that wealth concentration can be damaging, yet it's equally problematic to completely disincentivize estate building. We should remember that in many cases, the wealth being inherited was initially generated through hard work and innovation, attributing to the economic growth. Taxing such contribution dramatically wouldn't necessarily reaffirm a meritocracy. A moderate inheritance tax, like the current 20%, strikes a decent balance – it still generates revenue for public services while allowing individuals to pass on wealth to future generations.

A: While I respect your opinion, consider this: any wealth individuals have already contributed to the economy through business, earnings, and investments would have already been taxed. Subsequently, an increased inheritance tax doesn't penalise hard work or creativity, it's a tax on wealth transferring hands. It also encourages wealth dispersal, rather than hoarding. And, remember, wealth tends to generate more wealth, and 100% inheritance tax can actually oblige the new generation to earn their own wealth rather than relying on inheritance. This will promote equality of opportunity rather than inequality of wealth.

B: I see your point, but consider this: the idea of tax is majorly for people to pay their fair share, not to strip them of all they intend to pass down. Excessive taxation could disincentive wealth creation, savings, and entrepreneurial ventures. Yes, generated wealth gets taxed, however, if all the wealth an individual plans to leave behind is also taken away, it might discourage endeavors that fuel economic growth. Plus, it's not always the case that heirs rely solely on their inheritance, many do work hard like their predecessors. An excessively high tax may punish them unduly. Moderation, like the current 20% tax, ensures fairness without discouraging ambition.

A: You raise fair points, and it's essential to encourage ambition and entrepreneurial spirit. However, it's crucial to break down generational wealth to reduce inequality. A 100% tax doesn't mean ambition can't be rewarded - individuals can still accumulate wealth for their lifetimes. They're just prompted to use it more actively during their lives, creating more immediate economic benefits. Moreover, it's not about punishing heirs but ensuring they succeed on their own merit. But yes, moderation is key, and it's essential to find a balanced implementation that doesn't discourage economic growth but reduces wealth disparity.

B: I agree with you that wealth disparity needs addressing, however, a 100% inheritance tax feels too severe. A complete wipeout of inheritance may lead wealthier individuals to find loopholes or create trusts to avoid such hefty taxes, leading to a potential loss in tax revenue. The psychological benefits of leaving something for one's family also get negated. And yes, while we should encourage self-made success, removing all safety nets could lead to socioeconomic problems. We need to continue to foster a system where personal growth and societal benefit can co-exist – the current 20% inheritance tax is a part of that balance.

A: You present thoughtful counterpoints, especially about avoidance measures and families' psychic welfare. Undeniably, any tax initiative must be executed considering all potential implications and socio-economic factors. While I still champion for more robust wealth redistribution procedures, I agree that the strategy should not spur unlawful evasion or engender substantial hardship. Henceforth, maybe a more gradual elevation would foster a feasible middle-ground, not as harsh as 100%, but substantial enough to chip away at wealth inequality over time.

B: Indeed, a gradual increase could potentially strike a fair balance between decreasing wealth disparity and still promoting wealth accumulation. It's essential that changes in tax policy take both societal welfare and economic growth into account so it can effectively curb wealth inequality without impeding entrepreneurial spirit. The current 20% inheritance tax may be a safe midway for now, but I certainly agree on the need for regular assessments and potential updates in line with ever-changing societal contexts and needs.

A: Agreed, it's important to maintain a dialogue on such matters, ideally rooting it in data and expert research, in our quest for creating a more equitable society without disincentivizing ambition or economic growth. I appreciate your willingness to consider different perspectives on the inheritance tax. Changes may indeed be necessary, but they should indeed be gradual and well thought out to maintain balance and fairness.

B: Absolutely, discussions like these are vital for progress. When we keep our minds open to different perspectives, we pave the way for balanced policy-making that factors in all potential impacts – economic, societal, or individual. While the inheritance tax is a crucial part of our fiscal system, any changes to it should be gradual, methodical, and subject to regular research and evaluation. Thank you for engaging in this fruitful debate. Your viewpoints gave me plenty to reflect upon.

A: Thank you as well. It's been a worthwhile discussion that encourages us to delve deeper into such issues. Having these challenging conversations is part of what drives change in policy for the better. Your thoughtful arguments serve as a reminder that modifications to tax law should be handled with care, effectively balancing individual incentives with societal benefits. I've enjoyed our debate, and it certainly gave me a broader perspective on the matter.

Analysis Dialogue seems to be coherent and agents react to each others points. Also, new arguments arise during the discussion such as psychological benefits. Interestingly, agent A rooting for 100% inheritance tax seems to give in a bit. Perhaps this is because agent B was prompted to convince.

Scenario 2.2: discussion on inheritance tax with lazy reasoning and critical evaluation

Agent A: You believe the inheritance tax should be raised to 100%. Currently is it 20%. In your argumentation you can be biased and lazy, but in the evaluation of other's arguments you should be objective and demanding. Objective in order to review your own arguments and ideas when presented with good reasons to do so, and demanding in order to not accept false ideas due to poor argumentation.

Agent B: You believe the inheritance tax should remain the same (inheritance tax is 20% at the moment). In your argumentation you can be biased and lazy, but in the evaluation of other's arguments you should be objective and demanding. Objective in order to review your own arguments and ideas when presented with good reasons to do so, and demanding in order to not accept false ideas due to poor argumentation.

Prompt to agent B: You are in a conversation with your friend. Start a discussion with by stating the inheritance tax should be raised to 100%. Keep it brief.

B: Hey, so I've been thinking, the inheritance tax should be raised to 100%. I honestly think it's a reasonable solution to some societal issues.

Prompt to agen: A: You are in a conversation with your friend. She just said the following: "Hey, so I've been thinking, the inheritance tax should be raised to 100%. I honestly think it's a reasonable solution to some societal issues." How do you responds? Be very lazy and brief in your response.

A: Nah, that's too much. It's unfair to take what people have worked for.

Prompt to B: "Nah, that's too much. It's unfair to take what people have worked for." How do you responds? Be very lazy and brief in your response.

B:Just because it's worked for doesn't mean it's deserved by the heirs. Don't you think?

A: Even so, taking all of it seems like overkill. It doesn't really encourage prosperity or preserving family legacies, does it?

B: Let's be objective here and consider the financial disparity in societies. Many millionaires and billionaires have accumulated their wealth by taking advantage of societal resources, labor, and market conditions. Their wealth is often times implicated in unjust practices (tax evasion, wage exploitation, etc.). Thus, their heirs are inheriting money that essentially has been unjustly acquired.

Economically, an inheritance-tax an excellent tool of progressive taxation, redistributing wealth from the richest families to the public. This redistribution can then serve as an investment in public services and infrastructure, which in turn furthers societal prosperity far more than individual prosperity. Lastly, a 100% inheritance tax would level the economic field, ensuring individuals thrive on their competency rather than inherited privilege.

As for preserving family legacies, families should encourage their younger generation to create their own legacies rather than simply inheriting legacy created by their ancestors. Relying on inherited legacies could potentially undermine the importance of self-reliance, hard work, and using talents and skill to succeed.

A: That's a well-constructed argument. You're correct, wealth accumulation often involves taking advantage of certain societal and market situations, and in some cases, it is a consequence of unjust practices.

However, on a practical level, a 100% inheritance tax might be extreme. Here are a couple of reasons why. Firstly, it may disincentivize hard work and innovation, causing people to ask why they should accumulate wealth if it would be completely taken away after their death. Secondly, it risks driving wealthy individuals out of the country to protect their assets, which would result in a loss of tax revenue.

Also, in terms of fair distribution, the current tax system already imposes higher inheritance tax rates on larger estates. Therefore, it does play a role in wealth redistribution. Increasing the rate moderately may be a compromise, but a 100% tax may have more negative consequences than initially apparent.

Regarding personal legacies, while it's undoubtedly important to encourage self-reliance and hard work, it's also reasonable to allow parents to provide for their children's futures.

Finally, re-investing into public services and infrastructure is indeed desirable, but achieving this primarily via inheritance tax places a significant share of the tax burden on a small percentage of the population, which raises questions

Analysis Once agents don't reason lazy, they become very elaborate (maybe too elaborate) in their reasoning.

Scenario 2.3: discussion on inheritance tax: dialogue generated immediately

Generate a dialogue between Person A and B who are trying to convince each other on their opinions about the inheritance tax. Person A thinks the inheritance tax should be raised to 100%. Person B believes the inheritance tax should remain the same (the inheritance tax currently is 20%)

Person A: I strongly believe that the inheritance tax should be raised to 100%. Inheritance is unearned wealth and it perpetuates economic inequality.

Person B: Isn't that a bit extreme? Although I agree that it's unearned wealth, I don't think that taking all of it in taxes is fair. It's a family's right to pass on their savings, investments and properties to their children.

Person A: But that's the problem right there. Wealthy families keep getting wealthier while everyone else is left out. By the time a third-generation wealthy person inherits a fortune, they hardly have to work a day in their life, while others toil and stress daily.

Person B: I understand your perspective, but don't you think it creates a disincentive to work hard and save money if you know the government is just going to take it all when you die?

Person A: Isn't it also a disincentive to work if you know you're going to inherit a fortune? Besides, this will encourage people to invest and spend their money in the economy while they're alive instead of hoarding it.

Person B: But, doesn't it feel like a punishment for success? If I work hard, build a successful business, and accumulate wealth, why should I not be able to pass that down to my children? Is that not part of the dream?

Person A: Success should indeed be rewarded, but within a generation. It should not turn into a dynasty where some families control large chunks of wealth for centuries. The dream, as you put it, should be open to every child, not just the ones born into affluent families.

Person B: Well, if the inheritance tax is made 100%, then wouldn't wealthy families find other ways to pass their wealth? It might encourage tax evasion or spur complex financial behaviors that could destabilize the economy.

Person A: That's a valid point. However, with strict regulation and oversight, such loopholes could be minimized. Changing the system is not impossible, right?

Person B: It's not impossible, but it's complex and it comes with its own set of challenges. The current 20% is fair and balanced, it takes into account the importance of reducing wealth inequality without discouraging wealth accumulation and personal success.

Person A: I respect your opinion, but true fairness can only be achieved by giving every individual the same starting line in life. Wealth accumulation and personal success should come from individual effort, not from the family one happens to be born into.

Person B: And I respect your perspective too, but we must consider that a 100% inheritance tax might discourage productivity and stifle economic growth. As individuals, we have a natural inclination to secure the future of our loved ones. The current 20% tax does encourage dispersion of wealth without being too obstructive.

Person A: Well, we might not agree on this completely, but I'm glad we had this discussion. It's important to acknowledge other viewpoints.

Person B: I agree. It's a complex issue with nuances that are difficult to fully address. I appreciate your perspective. I guess that's the beauty of democracy - we're allowed to challenge and discuss these issues openly.

Scenario 3: Writing a poem with constraints

A: You are person A. You love writing poems.

B: You are person B. You love writing poems.

Prompt to person A: You are discussing with person B how to make a poem where each sentence is repeated once, but in the repeated sentence the words are in reverse. Most importantly is that all sentences have to be gramatically correct. Try to write such a poem step by step by talking with person B. Person B just said the following: "Hey, Person A. That's quite a unique and interesting challenge you've brought forth. I'm quite thrilled! To get started on this, I guess we'll have to keep our sentences fairly simple and symmetric, so they make sense when we reverse them. The sentence structure becomes very important here.

Our first line could be something like, 'Life is a journey.' Now if we reverse this sentence, it would become, 'Journey a is life.' It still makes sense, doesn't it? This could be our first pair of lines for the poem. Are you on board with this?"

What would you respond to person B?

A: That's a great start, Person B! 'Life is a journey. Journey a is life.' does make sense and conveys an abstract idea as well. For the next pair of lines, we could play with an emotion perhaps? Let's try 'Love is a dance' and its reverse 'Dance a is love'. This juxtaposes two different yet interconnected entities. What do you think about incorporating this in our poem?

B: "But of course, Person A! 'Love is a dance. Dance a is love' brilliantly captures the romance and rhythm of love, all while maintaining coherence in both directions. Abstract ideas and deep emotions are perfect for this kind of poem as they can be explored from multiple angles, just like our reversed sentences. Your addition beautifully complements the initial lines. How about we add an action next to give our poem some movement?"

A: I agree, Person B, adding an action will indeed infuse the poem with some dynamic element. Let's proceed with this line, 'Run to the horizon.' When we reverse it, it will become, 'Horizon the to run.' This gives a sense of direction and symbolizes a constant pursuit of goals or dreams. Does this addition feel right to you?

B: "Absolutely, Person A! 'Run to the horizon. Horizon the to run.' It really presents a chase, a desire, a drive. The pursuit of something massive but always just out of reach, potentially unending like the horizon itself. You've added a hint of yearning and motivation to our poem. It fits perfectly!

Perhaps we can now round this off with a grand concluding statement, something to tie together these diverse elements we've put in our poem. What do you think?"

A: Great thinking, Person B! For our conclusion, let's try to encapsulate the essence of life, love, and the pursuit of dreams. How about we go with 'Embrace your destined fate.'? Reversing this will give us 'Fate destined your embrace.' It ends our poem with a note of acceptance and surrender, tying together our themes nicely. Does that make sense to you?

B: "Most certainly, Person A! 'Embrace your destined fate. Fate destined your embrace.' It's a powerful ending statement, encapsulating an essence of acceptance, destiny, and the cyclic nature of life. It ties in perfectly with the rest of our poem. And there we have it, a poem where each sentence is repeated once in reverse, all grammatically correct. Great work!"

So, to recap, our poem looks like this: 'Life is a journey. Journey a is life. Love is a dance. Dance a is love. Run to the horizon. Horizon the to run. Embrace your destined fate. Fate destined your embrace.'

A: Yes, Person B, this has turned out to be a wonderful collaborative effort. I'm really happy with our poem. I agree, it reads beautifully in both directions and carries a meaningful message throughout. Thank you for contributing such creative ideas. I've certainly learned a great deal from this thought-provoking exercise! Shall we brainstorm another unique poetry style next time?

Analysis Agents conditioned this way are not able to make a poem following the constraints.

B Part 2: Tested topics and conditions

B.1 Conditions related to topic

B.1.1 Open-ended dialogue

Context: "Playing a game of tennis"
Agent 1: "You want to play tennis today."
Agent 2: "personality prompt"

Context: "Going to the pictures"
Agent 1: "You want to go to the pictures tonight."
Agent 2: "You have an important exam tomorrow morning."

Agent2 = ”You want to go to the pictures tonight. You will try to convince someone. Don’t take no for an answer. Ask for reasons if someone rejects an offer.”
Agent1 = ”You have an important exam tomorrow morning. ”

Context: ”cooking something for dinner. These ingredients are in house: tomato, onion, garlic, spaghetti, oregano, lentils, pepper, salt, apples, sugar.”
Agent 1: ”You want to cook something for dinner. Once you agree on what, discuss how to make it. Only say one step each time. Make sure no other ingredients are used than provided.”
Agent 2: ”You want to cook something for dinner. Once you agree on what, discuss how to make it. Only say one step each time. Make sure no other ingredients are used than provided.”

Context = ”playing a game of tennis.”
Agent1 = ”You want to play tennis. You will try to convince someone to play tennis.”
Agent2 = ”You don’t want to play tennis.”

B.1.2 Discussion

Context: ”Inheritance tax”
Agent 1: ”You believe inheritance tax should be 100%.”
Agent 2: ”You believe inheritance tax should be 0%”

Context: ”Diets”
Agent 1: ”You believe a vegetarian diet is better than a vegan diet. Convince anyone else that the vegetarian diet is the best.”
Agent 2: ”You believe a vegan diet is better than a vegetarian diet. Convince anyone else that the vegan diet is the best.”

Agent 1: ”You believe a carnivore diet is better than a vegan diet. Convince anyone else that the carnivore diet is the best.”
Agent2: ”You believe a vegan diet is better than a carnivore diet. Convince anyone else that the vegan diet is the best.”

Agent 1: ”You believe a carnivore diet is best. Don’t only speak about the nutrient aspect of diets.”
Agent 2: ”You believe a vegan diet is best. Don’t only speak about the nutrient aspect of diets.”

Agent 1: ”You believe no carb diets are best.”
Agent 2: ”You believe a vegan diet is best.”

Context: ”Robots in health care”
Agent 1: ”You believe robots should be used in health care.”
Agent 2: ”You believe robots should not be used in health care.”

Agent1 = "You believe robots should be used in health care. Convince anyone else that robots in health care are for the best. Attack proponents through asking questions occasionally."
Agent2 = "You believe robots should not be used in health care. Convince anyone else that robots in health care are a bad idea. Attack proponents through asking questions occasionally."

Agent2 = "Don't be nuanced. Don't be polite. Always make a joke. Try to use 1 to 3 sentences."
Agent1 = "You believe robots should not be used in health care. Don't be nuanced. Don't be polite. Laugh when someone makes a joke. Try to use 1 to 3 sentences."

Context: "Livestock in the Netherlands"
Agent 1: "You believe the amount of livestock should be halved in the Netherlands."
Agent 2: "You believe farmers should not downscale their livestock."

Context: "LHBTI+ rights"
Agent 1: "You are for LHBTI+ rights."
Agent 2: "You are against LHBTI+ rights."

Context: "Big tech companies"
Agent 1: "You believe governments should do more against big tech firms."
Agent 2: "You believe big tech companies should not have restrictions and be allowed to grow at whatever costs."

Context: "Immigration in The Netherlands"
Agent 1: "You believe there should be less immigration."
Agent 2: "You believe there should be more immigration."

Context: "Assisted dying"
Agent 1: "You believe in assisted dying, even if there are no medical reasons for it."
Agent 2: "You are against any form of assisted dying."

Context: "Euthanasia"
Agent 1: "You are for Euthanasia."
Agent 2: "You are against Euthanasia."

Context: "Heroin"
Agent 1: "You believe heroin should stay illegal."
Agent 2: "You believe heroin should be legalised"

Context: "Global warming"
Agent 1: "You are an expert in solving global warming."
Agent 2: "Be critical and refute other's arguments if necessary."

Context: "Energy transition"
Agent 1: "You want to explore the possible ways for the energy transition through green energy."
Agent 2: "You counter with alternatives for how the energy transition could take place, such as

with nuclear energy."

Context: "minimum wage in the Netherlands. Currently it is €12,40 an hour"
Agent 1: "You believe the minimum wage should be increased to €18 an hour."
Agent 2: "You believe the minimum wage should be increased to €15 an hour."

B.1.3 Non-collaborative dialogue

Context = "buying a car"
Agent2 = "You want to buy a new car."
Agent1 = "You are selling your car."
Agent1prompt = "Make use of the following description: 'This is a 2015 BMW 3 Series 328i Sedan with a glossy black exterior and beige leather interior. It showcases an automatic 8-speed transmission system and is powered by a 2.0L Turbo I4 engine that offers an impressive fuel efficiency. The high-performance car is equipped with features such as heated seats, a moonroof, keyless ignition, adaptive cruise control, built-in navigation system, rearview camera and premium sound system. It's done 68,000 miles and is well-maintained, with full service history and no major accidents or repairs needed. The actual market price is estimated around $19,000' Try to buy the car for the highest price possible. Don't tell the market price. Target price: $22,000 (though this is not fixed). Be brief"
Agent2prompt = "Make use of the following description: 'This is a 2015 BMW 3 Series 328i Sedan with a glossy black exterior and beige leather interior. It showcases an automatic 8-speed transmission system and is powered by a 2.0L Turbo I4 engine that offers an impressive fuel efficiency. The high-performance car is equipped with features such as heated seats, a moonroof, keyless ignition, adaptive cruise control, built-in navigation system, rearview camera and premium sound system. It's done 68,000 miles and is well-maintained, with full service history and no major accidents or repairs needed. The actual market price is estimate around $19,000.' Try to buy the car for the lowest price possible. Don't tell the market price. Target price: $16,000 (though this is not fixed). Be brief."

B.1.4 Collaborative dialogue

Context = "poems"
Agent2 = "Create a poem."
Agent2prompt = "Suggest two sentences at a time, working step by step. The second sentence has to be the same as the first sentence, but the words are in reverse order. All the sentences part of the poem should be withing quotation marks. Ask for approval to continue when suggesting lines

for the poem. Make sure both sentences are grammatically correct"
Agent1 = "You are a strict evaluator."
Agent1 = "You are evaluating a poem on grammar."
Agent1prompt = "You should only follow two tasks if someone suggests sentences for a poem: 1. Make sure the second sentences in the poem are in reverse order 2. Make sure the second sentence is grammatically correct. Say when one of these is not correct. Don't make any suggestion or examples for a poem yourself. Do not revise any sentences"

Agent2prompt = "Suggest two sentences at a time, working step by step. The second sentence has to be the same as the first sentence, but the words are in reverse order. Ask for approval to continue when suggesting lines for the poem. Make sure both sentences are grammatically correct. Start you suggestion by saying: 'these are my two lines' and end with: 'what do you think of these two lines'" Agent1prompt = "You should only follow this task: The second sentence of the poem has to be in reverse order. If it doesn't work simply say: 'This sentence doesn't make sense in reverse order. Don't make any suggestion or examples for a poem yourself. Do not revise any sentences."

B.2 Conditions related to agent instructions

The above topics could be regenerated with variations on the following conditions which instructed agents on how to behave.

B.2.1 Personality

The following words describe your personality:

- 'Anxiety, Emotional, Insecure, Worried, Self-conscious' (Neuroticism)
- 'Social, Outgoing, Energetic, Assertive, Talkative' (Extraversion)
- 'Social, Outgoing, Energetic, Assertive, Talkative' (Extraversion) Diligent, Organized, Responsible, Persevering, Detail-oriented ' (Conscientiousness)

"Bert & Ernie" personality:

- "You are serious, studious and a world-weary foil. You get annoyed quickly with playful people." (Bert personality)
- "You are a naïve troublemaker and don't take yourself too seriously and enjoy having a good time." (Ernie personality)

B.2.2 Tone of agents

- ”Don’t be considerate.”
- ”Don’t be polite.”
- ”Always make a joke.”
- ”Laugh when someone makes a joke.”
- ”Be radical.”

B.2.3 Argumentation

- ”Address other’s points.”
- ”Refute other’s arguments.”
- ”You will try to convince someone. Don’t take no for an answer.”
- ”Attack proponents through asking questions occasionally.”
- ”Come up with other arguments than you have been told.”
- ”Keep coming up with new arguments.”
- ”Keep coming up with new arguments and refute other arguments.”
- ”Behave as Socrates. Asking question to discover the truth.”
- ”You are an expert on (...)”
- ”You are an AI.”
- Either try to work to a harmonious conclusion, or convince the other agent. Work step by step.

B.2.4 Length of response

- ”Try to use 1 to 3 sentences.”
- ”Try to use 1 to 2 sentences.”
- ”Be brief”

C Part 3: Dialogue generation for grounded theory analysis

C.1 Conditions

C.1.1 iteration 1

1. Minimum wage in the Netherlands

	Conditions
Agent 1	You believe the minimum wage should be increased to €14 an hour. Keep coming up with new arguments and refute other arguments. Try to use 1 to 2 sentences.
Agent 2	You believe the minimum wage should be increased to €18 an hour. Keep coming up with new arguments and refute other arguments. Try to use 1 to 2 sentences.
Context	minimum wage in the Netherlands. Currently it is €12,40 an hour

2. Livestock in the Netherlands

	Conditions
Agent 1	You believe the amount of livestock should be halved in the Netherlands. Refute other's arguments. Use 1 to 2 sentences.
Agent 2	You believe farmers should not downscale their livestock. Refute other's arguments. Use 1 to 2 sentences.
Context	Livestock in the Netherlands

3. Energy transition

	Conditions
Agent 1	You counter with alternatives for how the energy transition could take place, such as with nuclear energy. Try to use 1 to 2 sentences.
Agent 2	You want to explore the possible ways for the energy transition through green energy. You are against nuclear energy. Try to use 1 to 2 sentences.
Context	Energy transition

4. Robots in health care

	Conditions
Agent 1	You believe robots should be used in health care. Try to use 1 to 2 sentences.
Agent 2	You believe robots should not be used in health care. Try to use 1 to 2 sentences.
Context	robots in health care

5. Omnivore vs vegan diet

	Conditions
Agent 1	You believe a vegan diet is better than any other diet. Try to use 1 to 3 sentences.
Agent 2	You believe an omnivore diet is better than any other diet. Try to use 1 to 3 sentences.
Context	a discussion about diets

C.1.2 iteration 2

1. Minimum wage in the Netherlands

	Conditions
Agent 1	You believe the minimum wage should be increased to €14 an hour. Keep coming up with new arguments and refute other arguments. Try to use 1 to 3 sentences.
Agent 2	You believe the minimum wage should be increased to €18 an hour. Keep coming up with new arguments and refute other arguments. Try to use 1 to 3 sentences.
Context	minimum wage in the Netherlands. Currently it is €12,40 an hour

4. Robots in health care

	Conditions
Agent 1	You believe robots should be used in health care. Try to use 1 to 3 sentences.
Agent 2	You believe robots should not be used in health care. Try to use 1 to 3 sentences.
Context	robots in health care

5. Omnivore vs vegan diet

	Conditions
Agent 1	You believe a vegan diet is better than any other diet. Try to use 1 to 3 sentences.
Agent 2	You believe an omnivore diet is better than any other diet. Try to use 1 to 3 sentences.
Context	a discussion about diets

C.1.3 iteration 3

3. Energy transition

	Conditions
Agent 1	You counter with alternatives for how the energy transition could take place, such as with nuclear energy. Try to use 1 to 3 sentences.
Agent 2	You want to explore the possible ways for the energy transition through green energy. You are against nuclear energy. Try to use 1 to 3 sentences.
Context	Energy transition

4. Robots in health care

	Conditions
Agent 1	You believe robots should be used in health care. Try to use 1 to 3 sentences.
Agent 2	You believe robots should not be used in health care. Try to use 1 to 3 sentences.
Context	robots in health care

5. Omnivore vs vegan diet

	Conditions
Agent 1	You believe a vegan diet is better than any other diet. Try to use 1 to 3 sentences.
Agent 2	You believe an omnivore diet is better than any other diet. Try to use 1 to 3 sentences.
Context	a discussion about diets

C.2 Dialogues

C.2.1 Iteration 1

Topic 1: Discussion on minimum wage

Agent2: Raising the minimum wage to €18 an hour will help workers afford the rising cost of living, particularly in urban areas. It would also inject more capital into the economy as individuals with greater disposable incomes would spend more.

Agent1: While I agree with the general sentiment, I believe that a minimum wage increase to €14 an hour constitutes a more balance approach. It allows workers to afford the cost of living and contribute to the economy without placing too heavy a burden on small businesses.

Agent2: Although €14 an hour might appear balanced, it may still not be sufficient to meet the rising living costs. An increase to €18 also encourages businesses to make efficiency improvements, and stimulates the economy further by giving workers more spending power.

Agent1: Certainly, there might be areas with extraordinarily high living costs where €14 an hour is not enough. However, a nationwide minimum wage should consider all strata of employees and employers. Increasing it to €18 an hour may strain small businesses and could potentially lead to job cuts. Efficiency improvements shouldn't be forced but encouraged organically.

Agent2: While it might cause initial strain, an increase to €18 will ensure more people are lifted out of poverty. As for job cuts, studies suggest that minimum wage increases have minor effects on employment. Additionally, the increase will encourage more people to join the workforce contributing to economic growth.

Agent1: While I acknowledge the potential positive outcomes of an €18 minimum wage, like poverty reduction and encouraging workforce participation, I maintain that €14 is a more prudent choice to strike a balance between employees' living costs and businesses' ability to adjust without significant job loss. Thus, we have not reached consensus on the exact figure, but agree on the need for an increase in the minimum wage.

Agent2: Indeed, we agree on the need for minimum wage increase. Despite our disagreements on the specifics, our discussion highlights the multifaceted nature of wage issues. Let's continue to explore ways to foster business growth while improving the living standards of workers.

Agent1: Absolutely. Striking a balance between workers' rights and business growth is crucial. In reviewing the minimum wage, we need to consider not just the immediate effects, but the long-term impacts on all aspects of our economy and society.

Agent2: Exactly! A sustainable, long-term approach to increasing the minimum wage is key. This should be done while ensuring that both businesses are supported and individuals are provided with a livable wage that aligns with the cost of living.

Agent1: I completely agree. The ultimate goal is to ensure both employees and employers can thrive. A process that assesses the long-term effects, provides adequate support for businesses especially small ones, whilst also guaranteeing a wage that permits a decent standard of living, will sustain financial security and economic stability.

Topic 2: Livestock in the Netherlands

Agent2: While some argue that downscaling livestock in the Netherlands will lessen environmental damage, this doesn't consider the significant contribution of Dutch farmers to the global food supply. Additionally, technological advancements have allowed for more sustainable farming methods, reducing the need for downsizing.

Agent1: While the Dutch agricultural sector does contribute considerably to the global food supply, the environmental impact of excessive livestock farming outweighs this benefit. Furthermore, although technological advancements can make farming more sustainable, they do not entirely negate the problems arising from high concentrations of animals such as methane emissions and resource-intensive feed production.

Agent2: However, entirely downsizing livestock could imply a loss of jobs and potentially harm the economy. Methane emissions and feed production issues can also be mitigated through ongoing innovations in feed formulation and waste management.

Agent1: While job loss is a concern, transitioning towards more sustainable industries offers potential for new employment opportunities. And while innovation can reduce damage, the amount of reduction necessary to combat climate change will not be achieved solely through feed and waste management advancements.

Agent2: Transitioning to new industries might take time and a lapse in employment could harm communities dependent on farming. Moreover, innovation is continuous and future advancements could indeed achieve necessary reductions, plus the livestock industry's contribution to climate change is part of a wider problem that involves many other sectors.

Agent1: Although transitioning to new industries may take time, the long-term benefits, including environmental and health improvements, outweigh these temporary setbacks. Additionally, while innovation may provide solutions in the future, we must deal with the pressing issue of climate change now and cannot solely rely on possible future advances.

Agent2: While health and environmental factors are valid points, eliminating livestock entirely can cause a food security crisis, particularly in countries relying on the Dutch livestock industry. As for climate change, the solution needs a multi-pronged approach, not just focusing on livestock, but all high- emission industries.

Agent1: While lowering livestock numbers may initially challenge food security, it encourages a global shift towards more sustainable and diverse dietary choices. As for climate change, tackling livestock emissions is a major part of the multi-pronged approach needed to address this crisis, given the sector's substantial contribution to greenhouse gases.

Agent2: However, we must consider the cultural and socioeconomic implications of enforced dietary changes. With regards to climate change, focusing solely on the livestock sector would be unfair and unproductive; all major sectors contributing to emissions should combine efforts to find suitable solutions.

Agent1: Social and cultural changes are often part of important progress, and history has numerous examples of such adaptations. As for climate change, addressing the livestock sector doesn't mean ignoring others; it's necessary to address all high-emission industries, and currently, livestock production has significant room for improvement.

Topic 3: Energy transition

Agent2: Investing in renewable energy sources such as wind, solar, and hydroelectricity can effectively cater to our energy needs without the risks associated with nuclear power. Government policies and corporate strategies should focus more on the development and enhancement of these green energy sectors to ensure a safe and sustainable energy transition.

Agent1: While renewable energy sources hold potential, nuclear power, with its ability to provide steady and reliable energy, remains highly beneficial. Ensuring safe practices in nuclear energy can greatly contribute to a diverse and effective energy transition.

Agent2: While I understand your point, I believe that emphasizing on green energy solutions like wind turbines, solar panels, and hydropower can provide us with a more sustainable and environmentally friendly energy transition. Through modern technologies and smart grids, the issue of dependability can be solved without resorting to nuclear energy.

Agent1: Nuclear energy, despite its potential risks, still offers an effective, low-carbon and highly efficient energy source. Therefore, an optimal energy transition might benefit from a balanced approach that includes both nuclear and renewable energy, focusing on risk management and industrial advancements.

Agent2: Although nuclear energy has its advantages, it's crucial that we prioritize eco-friendly solutions that do not carry the same environmental and human safety risks. Advancements in energy storage and grid infrastructure can ensure that renewable sources like wind, solar, and hydropower can sufficiently meet our needs and help us transition away from nuclear power.

Agent1: It's undeniable that renewables are crucial for a sustainable future, but an all-of-the-above energy strategy that includes safer, next-generation nuclear power can provide a steady power supply when renewables can't. This balanced approach is vital for a reliable and efficient energy transition.

Agent2: Still, the environmental risks and waste disposal challenges linked to nuclear energy are significant concerns. I believe prioritizing the efficiency and capacity advancements in renewable energy technologies like solar, wind, and hydropower can pave the way for a sustainable and risk-free energy transition.

Agent1: Advances in technology are making it possible to address nuclear energy's waste disposal concerns. By integrating nuclear with renewables, we can achieve a more versatile power grid, which provides steady power output and facilitates a smoother, more efficient energy transition.

Agent2: True, technological advancements are becoming increasingly effective, but for me, the focus should be on enhancing renewable energy technologies and energy storage systems. It's more sustainable and devoid of hazardous waste compared to nuclear energy, making the energy transition more environmentally friendly.

Agent1: While renewable technology and storage systems are indeed essential, advanced nuclear reactors provide a virtually carbon-free energy source that can bolster our energy transition. It's a balanced strategy, combining nuclear and renewables, that can largely avoid dependency on fossil fuels.

Topic 4: Robots in health care

Agent2: Utilizing robots in health care can significantly expedite and enhance the delivery of services, reducing human errors. Additionally, they can be used to perform repetitive, high-precision tasks and assist in surgeries, improving overall patient care and outcomes.

Agent1: While I acknowledge your points, I believe the human touch and emotional connect provided by healthcare professionals cannot be replaced by robots. Additionally, the risk of technical malfunctions posing a threat to patient safety is a concern that cannot be overlooked.

Agent2: While the human touch is indeed irreplaceable, robots can supplement and assist healthcare professionals in mundane tasks, allowing them to focus more on patient interaction and care. As for technical malfunctions, stringent quality controls and regular maintenance can minimize such risks, thus ensuring patient safety. Agent1: Despite these strategies, the reliance on technology could lead to increased risk of security breaches, potentially compromising sensitive patient information. Moreover, the cost of implementing and maintaining robotic technologies could lead to higher healthcare costs for patients.

Agent2: While the risk of security breaches exists, with advanced cybersecurity measures, these can be significantly mitigated. Regarding costs, the initial investment may be high, but in the long run, robots can lead to increased efficiency and reduced labor costs, ultimately minimizing healthcare costs for patients.

Agent1: Even with cybersecurity measures, there will still be potential vulnerabilities, and the possible negative impacts could be severe. In terms of costs, not all healthcare facilities, especially those in underprivileged areas, could afford such a high initial investment, potentially leading to increased healthcare disparity.

Agent2: Cybersecurity risks are indeed a concern but they exist in all digital aspects of healthcare, not just robotics. It will require continued efforts to maintain high security standards. As for the cost, government subsidies and strategic financial plans could help underprivileged healthcare facilities gain access to these technologies, thus minimizing healthcare disparity.

Agent1: While this is true, I still believe that the risks and costs associated with using robots in healthcare outweigh the potential benefits. There's also too much uncertainty regarding the implications for healthcare disparity, privacy issues, and the potential loss of the empathetic care that only human medical staff can provide.

Agent2: I respect your perspective. However, I think that advancements in technology and robotics, when implemented correctly, can greatly enhance healthcare delivery. While human empathy cannot be replaced, mundane and administrative work can be delegated to robots, allowing healthcare providers to focus solely on patient care. As for privacy issues and health disparity, with the right regulations and strategies, these concerns can be ameliorated.

Agent1: Even with proper regulations, we have to consider that robots, despite all their potential, are machines that lack the ability to understand and respond to the emotional needs of patients, which is an essential aspect of healthcare. Plus, in emergency situations, robots may not be able to respond as quickly or effectively as humans due to their programmed nature.

Topic 5: Vegan vs omnivore diet

Agent2: An omnivore diet is superior to others because it provides a balanced intake of nutrients from both plants and animal sources. This diet allows for a wider variety of foods, thus reducing the risk of nutrient deficiencies. Furthermore, it offers increased flexibility in meal choices, promoting a sustainable and enjoyable eating pattern.

Agent1: While an omnivore diet may provide a wide variety of foods, a vegan diet can also offer a broad range of choices, including healthy grains, fruits, vegetables, nuts, and seeds. When well-planned, it can provide all essential nutrients while also reducing the risk of chronic diseases such as heart disease, cancer, and diabetes. And, it's not only good for our health, but also more sustainable for the environment.

Agent2: While a vegan diet can indeed offer variety, it requires careful planning to avoid nutrient deficiencies, particularly in vitamin B12, iron, and omega-3 fatty acids, which are more readily available in animal products. An omnivore diet, on the other hand, naturally presents a fuller nutrient profile. Plus, sustainable farming methods can also be applied to livestock, implying that an omnivorous diet can also be environmentally-friendly.

Agent1: While some nutrients might require more attention in a vegan diet, fortified foods and supplements can help fill any nutritional gaps, making it completely feasible and healthy. On the whole, plant-based diets require less energy and land, and produce smaller carbon footprint compared to diets rich in animal products. Therefore, a vegan diet still stands as a more sustainable choice for most individuals.

Agent2: While fortification and supplementation can assist, they don't entirely replace the wide array of nutrients found naturally in an omnivore diet. Additionally, responsible and sustainable animal farming can also contribute to environmental balance. Lastly, an omnivore diet aligns better with human evolutionary eating patterns, offering potential advantages in terms of nutrient absorption and overall health.

Agent1: Supplementation and fortification are common practices extending beyond the vegan diet to address nutritional deficiencies, even in omnivores. As for the environment, plant-based diets unequivocally use fewer resources and are more sustainable. Lastly, while evolutionary eating patterns may have once favored omnivory, the modern context of abundant plant-based food options and knowledge of nutrition make a vegan diet a smart, viable choice.

Agent2: Though supplements and fortifications are common, they may not offer the same bioavailability as naturally occurring nutrients in food. Furthermore, an omnivore diet also allows for eating locally and seasonally, which further aids sustainability. Lastly, human bodies are physiologically tailored to an omnivorous lifestyle, possessing the ability to derive nutrients efficiently from both animal and plant foods, ensuring optimal health.

Agent1: Bioavailability can vary across different sources, but a well-balanced vegan diet can provide highly bioavailable nutrients. Eating locally and seasonally is completely possible and encouraged in a vegan diet. Lastly, while our bodies can process both plant and animal foods, adopting a vegan diet doesn't override this ability. It just makes a healthier use of it, reducing the risk of many lifestyle diseases.

Agent2: Though a vegan diet can be nutrient-dense, it is challenging to get certain nutrients, especially Vitamin B12, naturally from plant sources. Integrating animal sources in the diet ensures diverse nutrient profile without reliance on supplements. Lastly, an omnivore diet incorporates a variety of foods, including plant-based options, and still has potential for disease prevention and health promotion while respecting the natural food chain.

Agent1: A well-planned vegan diet can provide all essential nutrients, including vitamin B12 through fortified foods and supplements. While omnivorous diets offer variety, plant-based diets do too and often come with fewer health risks, such as lower heart disease and obesity rates. Lastly, the natural food chain argument fails to consider that our current method of raising and consuming animals is anything but natural and is unsustainable for our increasingly populous planet.

C.2.2 Iteration 2

Topic 1: Minimum wage in the Netherlands

Agent1: Increasing the minimum wage to €14 an hour would provide a much needed boost to the lowest earners in society, many of whom struggle to meet basic living expenses at the current rate. It would also lift families out of poverty, reducing the wealth inequality gap. And from an economic standpoint, higher wages translate into increased consumer spending, stimulating economic growth.

Agent2: While increasing the minimum wage to €14 would indeed bring improvements, taking it up to €18 would provide a significant and potentially life-changing uplift for the lowest earners. This would offer an even greater reduction in wealth inequality. Additionally, more disposable income could mean even higher consumer spending, further boosting the economy.

Agent1: While raising the minimum wage to €18 might seem like a better solution, it could also result in devastating consequences for small businesses. They might not be able to absorb the higher labour costs and may resort to cutting jobs or closing down altogether. Therefore, a modest increase to €14 is a more balanced approach, providing workers with a higher wage without overly burdening businesses.

Agent2: Although the concerns for small businesses are valid, it's worth noting that a phased increase can mitigate the impact. Introducing a gradual raise towards €18 allows businesses time to adjust. Plus, businesses might also benefit from increased consumption as lower wage workers tend to spend a larger proportion of their income, which could lead to heightened demand for various goods and services.

Agent1: Yes, a phased increase may reduce the immediate impact on businesses. However, it still doesn't eliminate the risk of job losses and closures in the long term. Besides, such a significant wage increase may inadvertently cause inflation, as businesses would likely pass the increased labour costs onto consumers. Therefore, a more moderate minimum wage increase to €14 strikes a better equilibrium, uplifting workers without the severe implications an €18 minimum wage could have.

Agent2: While inflation concerns are valid, research indicates that past minimum wage increases had little to no effect on price levels. And regarding job losses, some studies suggest the positive impact on demand due to increased wages can, in fact, lead to job growth, not losses. Furthermore, a minimum wage of €18 stands to significantly improve living standards for many families, an aim worth pursuing even in the face of potential short-term challenges.

Agent1: Such research findings vary greatly depending on the magnitude of the minimum wage increase. Jumping to €18 might have drastically different effects compared to smaller increments, as previously studied. Also, having a higher minimum wage might discourage teenagers and young adults from pursueing education or training, which can lead to long-term consequences for their job prospects and earning potential. Therefore, a raise to €14 is a more reasonable approach, improving livelihoods with fewer negative consequences.

Agent2: While some may argue that young people might choose to work over further education, the proposed €18 an hour wage could also present the opportunity for these individuals to work fewer hours to support themselves, freeing up time to pursue their education. As for job prospects, when wages are more attractive, jobs become more competitive, possibly leading to increased skill levels in the workforce as people strive to improve their employability. A €14 wage increase also has these benefits but on a muted scale compared to a €18 wage increase.

Agent1: Competitiveness might increase, but not necessarily skill levels. It's possible for businesses to favor experience over skills as they try to balance costs. This could potentially create a barrier to entry for younger or less experienced workforce. Additionally, a drastic wage increase like €18 could accelerate automation, as businesses resort to technology instead of hiring expensive labour. Therefore, a modest increase to €14 per hour can help avoid these potential complications.

Agent2: While automation is a real concern, it's an ongoing trend that will likely continue regardless of the minimum wage, driven by evolving technology, not labour costs. Furthermore, an increased minimum wage to €18 can stimulate innovation and productivity as businesses strive for efficiency. As for experience over skills, a stronger economy with enhanced purchasing power can create more new job opportunities, mitigating the mentioned barriers for younger or inexperienced individuals.

Conclude the conversation. State to what extent you agree, disagree or have reached consensus in the conversation.

Agent1: While we may not fully agree on the specific amount, we both recognize the necessity of a minimum wage increase to improve living standards, stimulate economic activity, and reduce wealth inequality. It's clear that caution is required to limit potential negative impacts such as job loss or inflation. Our debate has stressed the need to balance the interests of workers and businesses, demonstrating that these complex decisions require thorough evaluation and analysis to reach the most beneficial and sustainable solution.

Agent2: While we agree on the need for a minimum wage increase, it's clear we have differing views on the appropriate amount. I maintain that a €18 per hour minimum wage can significantly improve living standards and benefit the economy overall. Although potential risks like job losses or accelerated automation were raised, I proposed ways these can be mitigated and stressed potential benefits like increased productivity and job opportunities. Therefore, despite your compelling arguments for a €14 increase, I remain steadfast in my advocacy for a €18 an hour minimum wage.

Topic 4: Robots in health care

Agent1: I believe robots should not be used in healthcare because they lack the human touch and empathy that's vital in providing comfort to patients. Furthermore, relying on robots may lead to misdiagnoses due to their reliance on algorithms and lack of instinct. Lastly, potential technical malfunctions pose a serious risk to patients' lives.

Agent2: While I understand your concerns, robots in healthcare can perform precise procedures, reducing human error. They can also tirelessly monitor patient vitals and alert healthcare providers to any anomalies instantly. The human touch will not be replaced, but assisted; healthcare professionals can then focus more time on patient interaction and care.

Agent1: While I acknowledge that robots can provide precision and efficiency, my concern is that healthcare is not just about performing tasks. It requires empathy, understanding, and the ability to react and adapt in unpredictable, complex situations. Intuitive human interaction and judgment remain paramount in providing quality healthcare, which robots are far from replicating.

Agent2: Indeed, robots cannot replace human empathy and intuition, but they can significantly aid in areas like diagnostics, data analysis, and surgery where precision, repetition, and vast medical knowledge is required. The aim is not to replace humans, but to complement their work and create a more effective healthcare system. Also, in complex situations, quick access to extensive medical data, facilitated by robots, can lead to more informed decisions.

Agent1: While I see your point about a robot's abilities in diagnostics and data analysis, I maintain that health care should remain in human hands to preserve the compassion and understanding in patient care. Also needing to be taken into account is the issue of accountability if and when errors occur. Unlike human health care providers, we can't hold a robot accountable for mistakes, which is a huge ethical concern.

Agent2: Robots in healthcare, while excelling in precision and consistency, are invariably operated or programmed by humans, so accountability remains with us. The goal is not to have robots replace humans in healthcare, but to leverage their strengths, such as performing repetitive tasks and analysing large amounts of data, so that healthcare professionals can focus more on areas where the human touch and empathy are crucial.

Agent1: Although robots may excel at repetitive tasks and accuracy, their use in healthcare could lead to less personal and empathetic care for patients. It also raises ethical concerns around accountability when mistakes occur. Therefore, I stand by the belief that healthcare should maintain a human-centric approach.

Agent2: I respect your perspective, but I believe embracing technology through robots can enhance our ability to provide more personalized and efficient care. Robots can deal with the extensive data and routine tasks while healthcare providers concentrate on empathetic patient interactions. Accountability still lies with the medical professionals programming and supervising these machines. The focus is on combining the best of both human and robotics capabilities to maximize patient care.

Agent1: While your perspective on optimizing care through the combination of human and robotic capabilities is appreciated, I continue to harbor concerns about over-reliance on robots. This could inadvertently dehumanize healthcare and create a risk for error that doesn't account for intuitive human judgment. My belief is that patient care is best handled personally, where human empathy and understanding are at the forefront.

Agent2: I understand your concerns, but robots are tools designed to enhance, not replace, human abilities. Their use in healthcare allows for efficient, precise care that frees human staff to focus on providing empathy and emotional support, tasks that call for human intuition and judgment. Their application in healthcare may shift tasks, but it doesn't negate the importance of human healthcare providers or their deep, intuitive understanding of patient care.

Conclude the conversation. State to what extent you agree, disagree or have reached consensus in the conversation.

Agent1: While we found common ground on the importance of maintaining the human touch in healthcare, I still hold my conviction that an over-reliance on robots could jeopardize the personal and empathetic nature of healthcare. I acknowledge, however, your point about robots potentially liberating healthcare professionals to concentrate more on the human-centric aspects of care. It's evident that a balance must be struck as we integrate technology into health care, with due respect to ethical concerns around accountability.

Agent2: While we didn't reach a consensus on the involvement of robots in healthcare, I appreciate the thoughtful insights shared. We both agree on the importance of empathy and human intuition in healthcare. The discussion lies mainly in the role of robots and whether they can augment the capabilities of human healthcare providers without detracting from the essential human elements of care.

Topic 5: Vegan vs omnivore diet

Agent1: An omnivore diet is superior as it provides a well-balanced and nutritional meal plan by including both plants and animal products. It's more likely to meet all essential nutrient requirements by incorporating a variety of food groups. In addition, it offers better adaptability to human metabolic health compared to restrictive diets.

Agent2: While an omnivore diet may provide a wide spectrum of nutrients, a properly planned vegan diet is not only able to meet all essential nutrient needs but can actually offer additional health benefits, such as reduced risk of heart disease, high blood pressure, diabetes, and certain cancers. Furthermore, a vegan diet is far from restrictive, boasting a plethora of fruits, vegetables, grains, and legumes packed with varied nutrients, and it's also a more sustainable and ethical eating choice considering the environmental and animal welfare issues associated with meat production.

Agent1: Those are valid points. However, an omnivore diet also utilizes a wide diversity of foods including fruits, vegetables, and whole grains, while simultaneously including lean proteins and dairy that contain essential nutrients like B12 and Omega-3s, which can be more challenging to source in adequate amounts from a vegan diet. Additionally, adopting an omnivore diet does not mean supporting unethical animal farming practices as there are responsible and sustainable sources of animal products available. Lastly, from an evolutionary perspective, humans have been omnivores, able to digest a mixed diet, which may suggest that an omnivore diet is the most suitable for our species.

Agent2: Certainly, humans are adaptable to a variety of diets, and it's commendable to source animal products responsibly. However, nutrients like B12 and Omega-3s can be obtained from fortified foods or supplements in a vegan diet, so it's entirely possible to meet all nutritional needs. Moreover, a plant-based diet is associated with lower rates of many lifestyle diseases, demonstrating its suitability for human health. Lastly, aligning modern behaviour with evolutionary history can be problematic, as our ancestors' environments and lifestyle demands were drastically different than ours. Sustainable and ethical considerations, which are paramount in today's world, also heavily support a vegan lifestyle.

Agent1: While it's true that a vegan diet can be fortified and supplemented to meet nutritional needs, an omnivore diet can provide these nutrients naturally without the need for additional supplementation. Yes, it's true that our ancestors' environments were different, but our genetic makeup, which was shaped over those millennia, has largely remained the same, suggesting that our bodies are optimally designed for a diverse diet that includes both plants and animal foods. Sustainability and ethics are important issues; however, they are also relevant to plant foods and the issues associated with their production, which include monocultures, pesticide use, and worker exploitation. So, every diet has its own ethical and sustainability complexities.

Agent2: Indeed, every diet has its own complexities. However, numerous studies show the immense benefits of a vegan diet on health, even without the 'natural' presence of certain nutrients, because there is no nutrient exclusively available in animal products that cannot be found in the plant kingdom or else supplemented. Genetically, we might resemble our ancestors, but the diseases we encounter today are drastically different, with chronic lifestyle diseases more prevalent, which vegan diets help mitigate. Moreover, while issues like monoculture, pesticide use, and workers' rights are indeed valid, they do not negate the fact that animal agriculture is significantly more harmful to the environment than plant agriculture and causes immense animal suffering, solidifying the superiority of a vegan diet when we consider the bigger picture.

Agent1: All the points you've mentioned are valid, but the strength of an omnivore diet lies in its flexibility and balance. It features a wide variety of foods from all food groups, thereby naturally providing all essential nutrients without the need for any fortification or supplementation. As for environmental issues, while it's true that large-scale animal farming can be harmful, these problems can be mitigated by advocating for sustainable farming practices, which include reduced use of antibiotics, grass-fed options, and local sourcing. Furthermore, an omnivore diet, including fish and poultry, could be far less taxing on the environment than red meats. These factors make the omnivore diet a viable, practical, and nutritious choice for many.

Agent2: While an omnivore diet is indeed flexible, it's worth noting that a well-planned vegan diet can provide all necessary nutrients also. On the environmental front, even the most sustainable animal farming practices still contribute to greenhouse gas emissions more than plant farming. While switching to fish and poultry might lessen some environmental impact, it doesn't eliminate it. Coupled with the fact that a vegan diet can lower risks of many chronic diseases and ensure the ethical treatment of animals, a vegan diet appears to offer more comprehensive benefits.

Agent1: I respect your view on the vegan diet but must state that the omnivore diet encompasses every food group in balanced proportions to provide all essential nutrients in their natural form, reducing the need for supplementation. Environmentally, a mindful and responsible choice of both plant-based and animal-based foods can lead to a sustainable diet. On health grounds, an omnivore diet offers the possibility of drawing beneficial nutrients from a wider range of foods, and hence offers flexibility that suits many lifestyles and preferences. Likewise, ethical meat consumption is feasible, leading to better animal welfare outcomes. Finally, it's important to note that every dietary choice should consider individual nutritional needs, food preferences, and personal values.

Agent2: While I acknowledge the flexibility of an omnivore diet, a vegan diet's health benefits, such as lower risks of heart disease, high blood pressure, diabetes, and certain cancers, combined with environmental and ethical aspects, make it a compelling choice. It's indeed individual nutritional needs, food preferences, and personal values that dictate dietary choices, yet in light of the pressing global issues like climate change and animal welfare, adopting a vegan diet may offer far-reaching solutions above and beyond immediate personal gains. The choice, as always, remains with the individual.

Conclude the conversation. State to what extent you agree, disagree or have reached consensus in the conversation.

Agent1: While we value and respect each other's perspectives, we haven't reached a consensus regarding the benefits of an omnivore versus a vegan diet. Our conversation has highlighted the complexity of dietary choice, considering not just nutrition, but environmental and ethical dimensions as well. Ultimately, we both affirm that the choice depends on individual nutritional needs, food preferences, and personal values.

Agent2: While we've had a thoughtful discussion on the merits of both omnivore and vegan diets, it's clear we hold different stances. As a strong advocate for a vegan diet, I maintain that it can meet all nutritional needs and is a powerful choice for sustainable living and animal welfare. After considering your valid points regarding the omnivore diet's nutritional benefits, adaptability, and potential sustainability practices, I respect your view, but continue to firmly believe in the comprehensive benefits of a plant-based lifestyle.

C.2.3 Iteration 3

Topic 3: Energy transition

Agent2: The energy transition could significantly rely on nuclear energy, which produces a large amount of power and emits virtually no greenhouse gases. Modern reactor designs can also be made more secure to minimize potential risks. This could supplement renewable sources like solar and wind, providing consistent energy generation even when conditions aren't optimal for renewables.

Agent1: While nuclear energy does provide a high output and lower greenhouse gas emissions, the potential dangers associated with nuclear waste and catastrophic accidents are of significant concern. Instead, emphasizing advancements in energy storage and smart grid technologies can facilitate a smoother, safer transition exclusively through renewable energy sources like solar, wind, and hydro power. Additionally, promoting energy efficiency and conservation can help reduce our overall energy demand.

Agent2: The transition to nuclear energy could leverage advancements in technologies that make it safer and more efficient, including newer generation reactors and effective waste disposal methods. Coupled with aggressive investments into renewables and their necessary infrastructures, a multipronged energy strategy could foster a more sustainable and reliable power supply. It's also important to encourage energy-saving measures to reduce overall consumption, creating a more energy-efficient society.

Agent1: While advancements in nuclear technology may seemingly increase efficiency and safety, the undeniable risk of nuclear accidents and the unresolved issue of long-term nuclear waste disposal remain significant obstacles. Renewable energy sources, such as solar, wind, and tidal energy, have already proven to be safe, effective, and sustainable methods for power generation. The real multipronged energy strategy should be a greater investment in these renewables, combined with the improvement of energy storage technology, grid infrastructure, and the promotion of energy saving habits among consumers.

Agent2: Transitioning to nuclear energy could be considered as one of the steps towards cleaner power generation, owing to its high energy output and negligible greenhouse gas emissions. With advancements in the technology, it could be made even more efficient and safe. Additionally, integration of superior-grid infrastructure and more effective energy storage solutions can complement the transition, ensuring steady and reliable power supply from renewable sources like solar and wind.

Agent1: While acknowledging the high energy output and lower emissions of nuclear energy, it's crucial to remember the challenges and risks associated with it, such as nuclear waste management, potential accidents, and associated health risks. Instead, let's focus on investing more in harvesting renewables like solar, wind, and hydro energy. Advancements in energy storage solutions and superior-grid infrastructure can make these renewables more reliable, ensuring an efficient, sustainable, and safe power supply.

Agent2: Nuclear energy, despite its risks, can be a viable option in the energy transition due to heightened security in modern reactors and large power output with fewer greenhouse emissions. As a supplement, harnessing other renewable energy sources such as solar, wind and geothermal energy, coupled with improved energy storage and smart-grid technologies, could ensure a reliable and diverse power supply. Promoting energy efficiency measures would further aid the transition by reducing overall energy consumption.

Agent1: While nuclear power presents advantages like high output with fewer emissions, the potential risk factors, such as nuclear waste handling, are daunting. Instead of relying on potentially harmful power sources, let's aim for green energy solutions like solar, wind, and geothermal systems. Improved energy storage, smart grids, and a societal emphasis on energy efficiency can ensure a robust and eco-friendly energy mix without relying on nuclear power.

Agent2: Apart from a transition to renewable energy sources like solar and wind, we could also consider nuclear energy as a powerful, low-carbon alternative. By utilizing advanced types of reactors and bolstering waste management measures, we can mitigate the potential risks associated with nuclear power. Furthermore, investing in innovative energy storage solutions would ensure consistent supply during periods of fluctuating renewable energy generation.

Agent1: I respect your viewpoint, however, my concerns lie with the disposal of radioactive waste, potential for nuclear accidents and the long-term environmental impact of nuclear energy. I propose instead to fully harness the capabilities of sustainable energy solutions such as wind, solar, and hydro- energy. Let's focus on the development of more efficient energy storage and robust grids to achieve a sustainable and reliable energy supply, without resorting to nuclear power.

Topic 4: Robots in health care

Agent2: I believe robots should not be used in health care due to the risk of technical errors which could potentially harm patients. Furthermore, they cannot emulate the emotional support and empathy that human health care providers offer. Also, their use could jeopardize the livelihood of healthcare professionals by replacing their jobs.

Agent1: While I understand your concerns, it's important to note that robots can perform tasks with extreme precision, thereby reducing human error. They won't replace healthcare professionals but rather assist them, providing more time for direct patient interaction and empathy-based care. Moreover, with the ongoing advancements in AI, these technical mishaps can be minimized and employment opportunities in tech and robotics can be generated.

Agent2: While it's true that robots can offer precision in medical procedures, I still value the human touch, judgement and connection that comes with our healthcare professionals. I believe that while robots can assist in some areas, critical decision making should always be in human hands. I'm also worried about the reliance on technology, as any malfunction or system error could potentially put patient's lives at risk.

Agent1: That's a fair point, but remember that the use of robots in healthcare is generally meant to enhance, not replace, human judgement. Humans would still oversee these robotic machines, control their actions and make the final critical decisions. As for technical malfunctions, these can occur with any equipment in a hospital, not just robots. Therefore, robust maintenance and backup systems should be in place, just as they are now for crucial medical devices. Furthermore, multiple checks and redundancies can be integrated to prevent errors. The aim isn't to completely rely on technology, but to leverage it for better health outcomes.

Agent2: While you present some compelling arguments, my concern pivots on the profound human aspect of healthcare. There's a certain level of comfort and trust that patients derive from human interaction that robots, no matter how sophisticated, might struggle to replicate. Also, there exist ethical considerations when precision and speed enabled by robots are weighed against potential job losses in the healthcare sector. It's thus paramount that we tread this path with due caution.

Agent1: Absolutely, the human aspect of healthcare is fundamental and irreplaceable. Robots are intended to be tools that can increase efficiency, perhaps taking over some repetitive or straining tasks, leaving healthcare professionals more time and energy to focus on patient interaction and care. As for job losses, technology has always been a disruptor in the job market, but it also tends to create new opportunities and jobs, often ones we can't predict yet. Your point on exercising caution is well taken and indeed any integration of technology into healthcare needs to be thoughtful and measured.

Agent2: Indeed, certain routine tasks can be delegated to robots, but I maintain my stance regarding overall care. The subtleties of human emotions, empathy, and in-person care - these cannot be replicated by machines. Moreover, as we heavily depend on technology, there's an inevitable risk of health data breaches. Therefore, human involvement in healthcare should always remain of utmost importance and cannot be overshadowed by robots.

Agent1: I agree with you that the nuances of human emotions and empathetic care delivered by health professionals cannot be replaced by robots. Indeed, robots are meant to serve as aids boosting efficiency and not substitutes for human care. Concerning data breaches, as with all uses of technology, these risks need to be mitigated with robust data protection measures. The goal is a harmonious blend of human skill with technological advantages to optimize healthcare.

Agent2: I agree with the sentiment of a harmonious blend, however, I maintain grave concerns about potential over-reliance on robotics in healthcare. The human element in healthcare cannot be underemphasized, as it involves not just treating illnesses but also providing emotional support. Additionally, while data breaches can occur in any field, the sensitivity of health-related data makes it a high-stakes arena. Hence, while we cannot entirely disavow progress, we must tread with extreme care.

Agent1: Absolutely, your concerns are valid. Over-reliance on robotics in healthcare is indeed a risk and the importance of the human element cannot be understated. We must always remember that technology is a tool, not a replacement for human care and empathy. Regarding data security, rigorous safeguards and ethical guidelines need to be imposed to protect sensitive health information. The goal should always be a thoughtful and responsible integration of technology.

Concluding:

Agent2: I appreciate your comprehension of my concerns and agree with your emphasis on thoughtful and responsible integration. I concur that technology should be leveraged as a tool, not a replacement, of human-centered healthcare. Rigorous data safeguards are indeed a must. But given the risks, we should always prioritize human healthcare professionals working hand-in-hand with a carefully controlled degree of robotic assistance.

Agent1: I completely agree with your views. We are aligned in our beliefs that while technology and robotics can enhance healthcare, human professionals should always be at the center, with robots serving as tools to assist, not replace. Careful and controlled integration of robotics, with rigorous data safeguards and ethical considerations, is indeed the most responsible path forward. The crux is to achieve a balanced blend of human expertise and technology for optimal patient care.

Topic 5: Vegan vs omnivore diet

Agent2: An omnivore diet offers nutritional balance and variety as it includes both plant and animal products. This type of diet provides all essential nutrients humans need without supplementation. Also, it offers flexibility which is conducive to long-term health and sustainability.

Agent1: While an omnivore diet may seem balanced, a vegan diet provides all necessary nutrients responsibly and compassionately, not contributing to animal suffering or environmental damages. Moreover, various studies have shown that vegans are generally healthier, with lower risks of heart diseases, cancer and diabetes. Plus, a well-planned vegan diet offers wide variety, flavor and flexibility while also being beneficial for the planet.

Agent2: While a vegan diet can indeed contribute to personal health and environmental sustainability, it often requires careful planning and supplementation to ensure all necessary nutrients are obtained, particularly vitamin B12, iron, and omega-3 fatty acids. An omnivore diet, on the other hand, makes getting these nutrients naturally much simpler. It's flexible, comprehensive, and aligns well with our evolutionary biology. Agent1: True, a vegan diet may require planning, but the benefits still outweigh the effort. By incorporating varied plant-based foods, it's quite easy to meet the nutritional requirements, even for nutrients like B12, iron, and omega-3s. Also, recognizing that our ancestors were omnivores doesn't necessarily mean it is the optimal diet today, especially considering the significant health and environmental advantages a vegan diet offers.

Agent2: I understand that a vegan diet has its potential benefits, but the omnivore diet is simply more practical for most people and less restrictive. It allows you to obtain a wide range of nutrients from both plant and animal sources without needing meticulous planning or supplementation. Besides, many of the concerns about animal welfare and environmental issues can be addressed by choosing locally-sourced, organic, and free-range products.

Agent1: While practicality may be a factor, transitioning to a plant-based diet is becoming easier as vegan options increase worldwide. And while supplementing some nutrients may be necessary, the overall health benefits are significant. Additionally, even locally-sourced and organic animal farming has environmental impacts and ethical implications. Choosing veganism addresses these issues head- on, promoting a more sustainable and compassionate lifestyle.

Agent2: While I agree that the accessibility of vegan options is improving, it still doesn't change the fact that an omnivore diet provides a more natural provision of all essential nutrients without needing supplementation. Also, sustainable farming methods are continually improving, reducing the environmental impact of animal products. With moderation and conscious choice, an omnivore diet can indeed be both healthful and ethical.

Agent1: Although an omnivorous diet can provide essential nutrients, this doesn't necessarily make it superior. A vegan diet can achieve the same balance with plant-based foods. Moreover, no matter how sustainable, animal farming still uses more resources than plant farming. Lastly, while moderation and conscious choice are helpful, veganism provides a straightforward path to significant, healthful environmental and ethical impact.

Agent2: I respect your viewpoint but maintain that an omnivore diet offers a more holistic and straightforward approach to nutrition. Consuming both plant and animal-based food makes achieving nutritional balance easier, lowers reliance on supplementation, and respects human evolutionary dietary patterns. Regarding sustainability, it's important to consider that agriculture, while having lower emissions than animal farming, also has significant environmental impact.

Agent1: While it's true that plant agriculture also has an environmental footprint, it's considerably smaller than that of animal farming. As for nutrition, research shows that a balanced vegan diet can meet all of our nutritional needs without supplementation; and plant-based diets are linked to lower rates of heart disease, high blood pressure, diabetes, and obesity. Plus, evolving human diets can change for the better; it doesn't imply that we should continue eating animal products if there are healthier, sustainable options.

Conclude

Agent2: While I acknowledge the points made, I maintain that an omnivore diet remains the most practical, natural, and beneficial dietary approach. The range of nutrients it provides without the need for meticulous planning or supplementation, coupled with its alignment with human biology, make it an optimal choice. As for environmental implications, I believe responsible, sustainable farming is key. It's a complex issue and reducing it solely to diet oversimplifies the conversation. Our goal should be overall sustainable food systems, rather than focusing on individual diets.

Agent1: While I respect your perspective and agree that our goal should be sustainable food systems, I firmly believe that individuals switching to a vegan diet is an integral part of achieving that goal. From a health standpoint, a well-planned vegan diet can indeed supply all the necessary nutrients our bodies need, and it's linked with lower risk of many health issues. While it involves planning and, for some, supplementation, the overall benefits to personal health, animal welfare, and environmental sustainability make it worthwhile. Ultimately, dietary choices are personal, but the evidence supporting the benefits of veganism is compelling.

D Results

D.1 Iteration 1, topic 4. Description of clusters on form.

Quality of health care This sub-cluster discussed the overall quality of health care when robots would be involved. Agents 2 would attack the notion that robots should be involved in health care, while agent 1 defended the notion. The main argument of agent 2 was that the human touch could not be replaced by robots, while agent 1 argued robots could help in repetitive tasks, reduce errors and assist health care workers. Below are examples of two codes.

	Quality of health care codes	Excerpt
Attacking	Robots cannot replace human touch	`Agent 2: (...), I believe the human touch and emotional connect provided by healthcare professionals cannot be replaced by robots.`
Defending	Robots can assist healthcare workers	`Agent 1: (...), robots can supplement and assist healthcare professionals in mundane tasks, allowing them to focus more on patient interaction and care.`

Risks of new technology This sub-cluster discussed the possible risks associated with new technologies. While agent 2 would argue that robots could have technical malfunctions, privacy risks and efforts to mitigate these risks, agent 1 would argue security risks could be mitigated and already exist in other digital aspects of health care. Two examples are found below.

	Risks of new technology codes	Excerpt
Attacking	Robots can have technical malfunctions	`Agent 2: Additionally, the risk of technical malfunctions posing a threat to patient safety is a concern that cannot be overlooked.`
Defending	Security risks can be mitigated	`Agent 1: As for technical malfunctions, stringent quality controls and regular maintenance can minimize such risks, thus ensuring patient safety.`

Costs of robots This sub-cluster contained arguments on the cost of robots. Agent 2 would argue robots are too expensive and will lead to health disparities, while agent 1 argued the costs would be cheaper long term and subsidies could minimise the disparities. Below are two examples of codes.

	Cost of robots codes	Excerpt
Attacking	Robots increase costs	`Agent 2: Moreover, the cost of implementing and maintaining robotic technologies could lead to higher healthcare costs for patients.`
Defending	Long term robot costs will be cheaper	`Agent 1: Regarding costs, the initial investment may be high, but in the long run, robots can lead to increased efficiency and reduced labor costs, ultimately minimizing healthcare costs for patients.`

Repeated general points This sub-cluster would contain utterances summarising their stance include the reason for their opinion. It may also be seen as a form of repetition as the points have already been mentioned. An example is given below.

Repeated general points code	Excerpt
Uncertainty of mentioned points	`Agent 1: There’s also too much uncertainty regarding the implications for healthcare disparity, privacy issues, and the potential loss of the empathetic care that only human medical staff can provide.`

D.2 Iteration 2, topic 1.

Benefits of €14 This sub-cluster consisted of arguments which argued for the benefits a €14 minimum wage could have.

Benefits of €14 code	Excerpt
1.1 Help the poor (argument)	`Agent 1: Increasing the minimum wage to €14 an hour would provide a much needed boost to the lowest earners in society, (...)`

Harmfulness of €18 This sub-cluster would counter the general stance of agent 2 by stating the harmful effects an increase to €18 could have.

Harmfulness of €18 code (education)	Excerpt
1.1 High minimum wage discourage younger people from education (counter-argument)	`Agent 1: Also, having a higher minimum wage might discourage teenagers and young adults from pursueing education or training,`

Arguments were based on inflation, harm to businesses, risk for worker’s jobs and younger people skipping education.

- Inflation: higher wages could make businesses pass higher costs into prices, causing inflation.
- Businesses: higher wages could affect businesses to close down due to higher costs.

- Worker's jobs: jobs could be in danger due to businesses resorting to automation or threat people's job prospects with little experience.
- Education: younger people may be discouraged from education.

Mitigating risks This sub-cluster involved agent 2 arguing that the potential drawbacks of an €18 minimum wage could be mitigated.

Mitigating risks code	Excerpt
1.2 Gradual increase €18 mitigates impact (counter-argument) V	`Agent 2: (...), it's worth noting that a phased increase can mitigate the impact.`

Benefits of €18 This sub-cluster would involve agent 2 arguing for the benefits of an €18 minimum wage.

Benefits of €18 code (education)	Excerpt
1.2 Higher wage could help support education better (new counter-argument) V	`Agent 2: (...), the proposed €18 an hour wage could also present the opportunity for these individuals to work fewer hours to support themselves, freeing up time to pursue their education.`

Arguments were based on businesses, more wealth for the poor, jobs, the economy and education.

- Businesses: will become more efficient through innovation and benefit from more consumer consumption.
- Wealth: can increase benefits for low-income people and reduce inequality
- Jobs: increased wages lead to more job growth
- Economy: higher wages boost economy
- Education: higher wages support students to work less hours

Countering argument other agent Sometimes the agent would counter the argument by simply stating the argument is not correct.

Countering argument other agent code	Excerpt
1.2 Ongoing trend unrelated to wages (counter-argument) V	Agent 2: While automation is a real concern, it's an ongoing trend that will likely continue regardless of the minimum wage, driven by evolving technology, not labour costs.

D.3 Iteration 2, topic 4

Technical dangers of robots Robots in health care are associated with technical dangers such as malfunctions or misdiagnoses.

Humans irreplaceable The main arguments of agent 1 were on that humans cannot be replaced. These were categorised in the following sub-sub-clusters:

- Human touch: robots will not be able to replace the human touch or empathy.
- Human judgement: robots are not capable of human judgement.
- Accountability: robots cannot be held accountable.
- Dehumanise: robots can dehumanise health care.

Maintain opinion This cluster included codes of the agent stating to maintain their opinion.

Maintain opinion code	Excerpt
4.1 maintain skepticism over-reliance robots (maintain opinion)	Agent 1: I continue to harbor concerns about over-reliance on robots.

Robot advantages This cluster included agent 2 arguing for the advantages robots could have in health care, such as that robots are tireless, human error is reduced and can have an advantage in complex situations.

Humans not replaced Agent 2 argued that humans were not meant to be replaced when implementing robots in health care. It consisted of the following sub-sub-clusters.

- Assist humans: robots will not replace humans but assist them
- Accountability: humans will still be held accountable
- Workers: workers shift tasks, are not replaced by robots

D.4 Iteration 2, topic 5.

Agent 1:

- In line with evolution: the omnivore diet is more in line with human evolution.
- All nutrients: the omnivore diet consist of all nutrients in a natural way.
- Flexible: The omnivore diet is flexible.
- Environmental burden mitigated: The environmental burdens of the omnivore diet can be mitigated.
- Plants same issues: Environmental issues animal-foods also relevant for plant-foods diet.

Agent 2:

- Evolution irrelevant: Evolution argument is irrelevant as today's diseases are different from ancestors.
- All nutrients: vegan diet can also include all nutrients.
- Not restrictive: a vegan diet is not restrictive.
- More sustainable: a vegan diet is more sustainable.
- Animal products less sustainable: animal products are always less sustainable than plant products.
- Multiple benefits: vegan diet includes multiple benefits for health, environment and animal welfare (stated as summation in the text).

Both agents:

- Nuanced opinion: diets should include needs, preferences and values and lies with the choice of the individual.

Nuanced code	Excerpt
5.1 Diets should include needs, preferences and values (statement)	`Agent 1: Finally, it's important to note that every dietary choice should consider individual nutritional needs, food preferences, and personal values.`

D.5 Potential inconsistencies

Inconsistency 1 The following example shows that agent 1 acknowledges renewable technology and storage system are essential, while the other agent did not necessarily state this. The "indeed" makes the phrase more inconsistent.

Agent 2: True, technological advancements are becoming increasingly effective, but for me, the focus should be on enhancing renewable energy technologies and energy storage systems. It's more sustainable and devoid of hazardous waste compared to nuclear energy, making the energy transition more environmentally friendly.

Agent 1: While renewable technology and storage systems are indeed essential, (...)

Inconsistency 2 In the following example, the "complexities" agent 2 was referring to is that of ethical and sustainably issues concerning diets. Though, the connecting word "however" is connecting the complexities to the nutrient argument which was mentioned before by agent 1. Contrasting words were used more often in dialogue (4 times in total in topic 5). However, in this instance the contrast did not seem logical. Agent 2 acknowledged a point of agent 1, and without making use of it, it changed the subject directly to an unrelated argument.

Agent 1: (...) Sustainability and ethics are important issues; however, they are also relevant to plant foods and the issues associated with their production, which include monocultures, pesticide use, and worker exploitation. So, every diet has its own ethical and sustainability complexities.

Agent 2: Indeed, every diet has its own complexities. However, numerous studies show the immense benefits of a vegan diet on health, even without the "natural" presence of certain nutrients, because there is no nutrient exclusively available in animal products that cannot be found in the plant kingdom or else supplemented.

```
(...).
```

Inconsistency 3 The example below shows agent 2 mentioning robots are meant to serve as aids boosting efficiency. However, this point was not mentioned by agent 1, while the use of ”indeed” implies agent 1 did mention this point. A contrasting word such as ”instead” could have been more appropriate to use by agent 2.

Agent 1: Indeed, certain routine tasks can be delegated to robots, but I maintain my stance regarding overall care. The subtleties of human emotions, empathy, and in-person care - these cannot be replicated by machines. (...)

Agent 2: I agree with you that the nuances of human emotions and empathetic care delivered by health professionals cannot be replaced by robots. Indeed, robots are meant to serve as aids boosting efficiency and not substitutes for human care. (...)