UTRECHT UNIVERSITY

Department of Information and Computing Science

Artificial Intelligence master thesis

Exploring the use of large language models to improve one-to-one communication training scenarios

First examiner:
Johan T. Jeuring

Candidate:
Jens L.L. Hartkamp

Second examiner:
Albert Gatt

External supervisor:
Jordy van Dortmont

September 11, 2024

Abstract

This paper explores the automation of communication training scenario writing by using AI generated statements. Two different research questions were explored, the first attempts to improve the quality of the training scenarios by replacing so-called non-functioning "Distractor" options. These options are pitfalls to make a user think more about the best practice response in a certain scenario. Distractors were first evaluated where non-functioning distractors were identified and replaced by AI generated statements. The original and generated distractors were implemented in communication training scenarios which were used in a bachelor course. The differences in how often these options were chosen were analyzed using a Mann Whitney U test. There were no significant differences found, although the outcome was very close to the 0.05 significance cutoff with a p-value of 0.057. For the second research question we attempted to generate statements based on the desired parameter settings. The output was evaluated by experts who concluded the results are promising, but not ready for automation without human evaluation. The expert grades of the AI generated options were significantly worse than the rating of the human written distractors meaning more work has to be done before fully automated parameter based answering option generated is viable.

Contents

1. Introduction

Communication is fundamental to human existence. Without it, the world would not be able to function in the way it currently functions. Teachers would not be able to teach, politicians would not be able to express their ideas and views, and relationships would be nearly impossible to maintain. In history there have been many examples of the importance of communication. Even in high stakes scenarios such as the aviation business, Molesworth and Estival [21] showed 20% of pilots still experience miscommunication on a weekly basis.
As important as communication is, humans are not always good at communicating. Words all have a certain meaning, but this meaning can be interpreted in multiple ways. Non-verbal communication also influences the way a message is received. Facial expression, as well as body movement, posture and eye-contact can heavily influence what a person is trying to convey [41, 38]. All of these factors combined can have a lot of influence on the way a message is perceived and can sometimes completely change interpretation from person to person.
The impact of bad communication is often not that high and is easily resolved. Simply asking whether you understood somebody correctly can iron out a lot of miscommunication. However, in some cases the stakes are higher. Especially when there is a high emotional load on a message, bad communication can cause damage [35, 2, 3, 18].

DialogueTrainer, originating from the platform "Communicate!" [12, 17], is a company that develops simulations for communication skills training. They provide virtual scenarios in which a user can communicate with a virtual agent. Every scenario has one or more predetermined goal(s) that a user should try to achieve. For example, a scenario in which a user is a doctor talking to a patient, the goal of the doctor might be to bring bad news in a clear and emphatic way. A user gets a number of dialogue options to choose from. After they choose an option, the agent shows an emotion, and replies. An example is shown in Figure 1.1. The dialogue options chosen by a user lead to a score for each parameter in the scenario. In our example those parameters are clarity and empathy. Apart from a score, DialogueTrainer also provides textual feedback af-

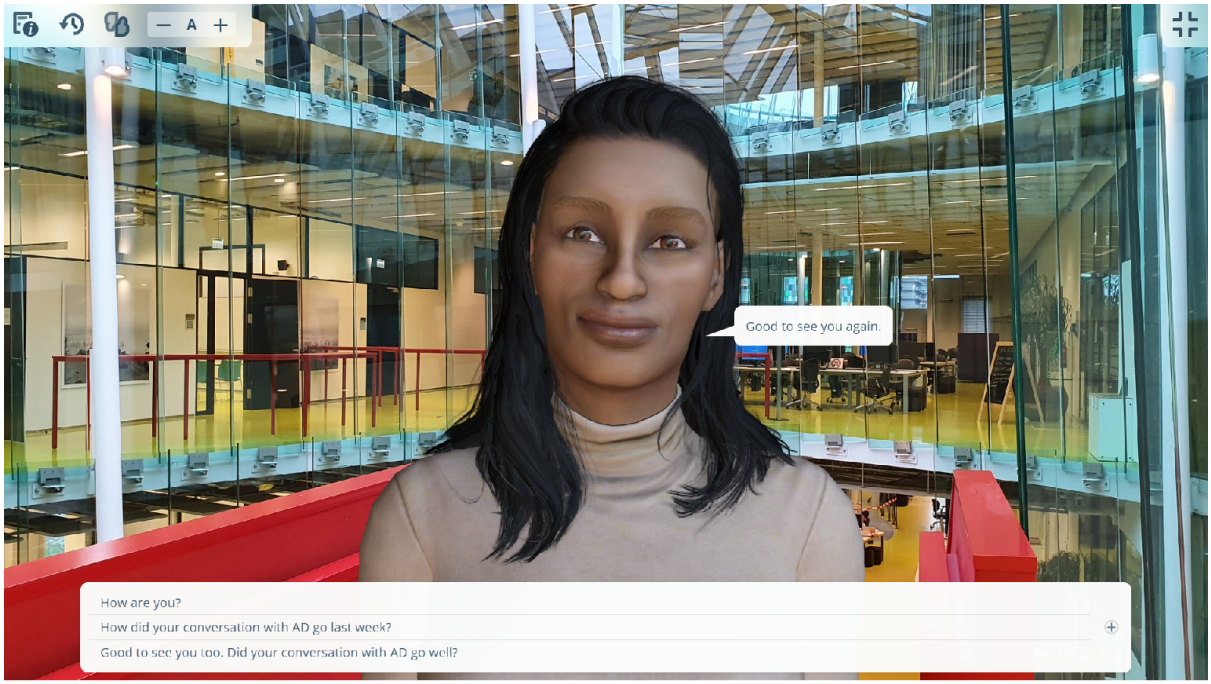

Figure 1.1: Example of a DialogueTrainer scenario with a prompt and three answering options

ter choosing certain options as shown in Figure 1.2. Moreover, more in depth feedback is given after finishing a scenario. This feedback can be constructive, when a user has improvements to make, or reinforcing when a user is already doing a good job.

DialogueTrainer is a popular application with positive reviews, an honours award rewarded by CES [1], and thousands of users playing scenarios every year. But what can we say about the quality of the scenarios? Do scenarios resemble real-life conversations? How do they support learning? We will view scenarios as a sequence of multiple-choice questions (MCQs). Does the answering data satisfy quality criteria from classical test theory? What would such quality criteria look like?
A high quality scenario is a scenario in which a user who mastered the skills practiced in the scenario chooses the best options, but a learner regularly chooses a sub optimal option. First we need to establish how to determine the quality of a scenario. Different approaches to determining the quality of a scenario are: asking experts to analyse the scenarios [13], do user tests with questionnaires on user experience [16] or take a more data-driven approach by looking at the data of user behavior when playing a scenario [10].

[1] https://www.ces.tech/innovation-awards/honorees.aspx

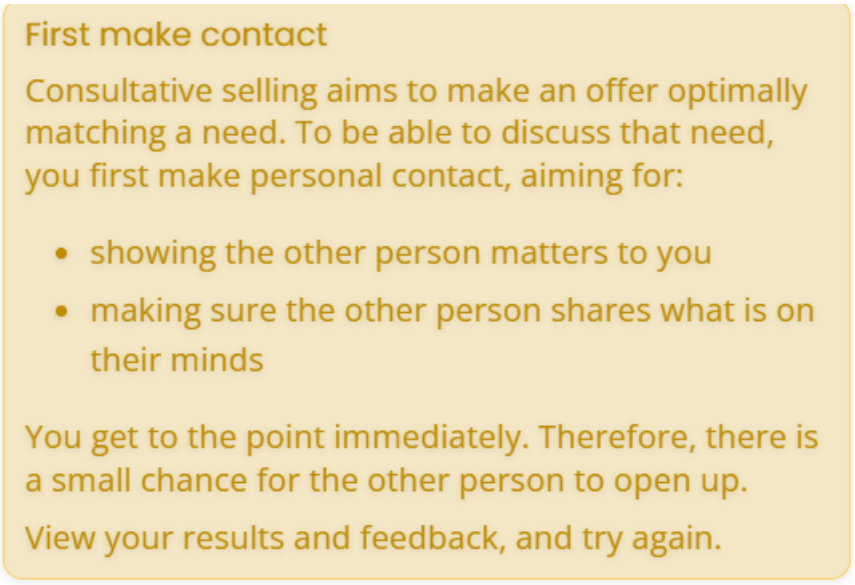

Figure 1.2: Example of feedback in a DialogueTrainer scenario

The goal of this paper is to develop an approach to *improving* the quality of DialogueTrainer scenarios. We will develop a data-driven approach to identify where to improve these scenarios. On top of that expert feedback will be used to asses and ensure high quality communication training scenarios. Within data-driven approaches there are still many options: sentiment analysis, regression analysis and factor analysis just to name a few. The quality of the multiple-choice questions used in educational measurement depends on their difficulty, discrimination, and distractor efficiency [7]. As the answering style in DialogueTrainer communication training scenarios is similar to multiple choice questions, literature on multiple choice questions will be used for this research. Research has already been done on difficulty and discrimination within DialogueTrainer scenarios [10]. This thesis will attempt to further improve the quality of DialogueTrainer scenarios, by focusing on distractor efficiency.

In a good quality scenario, we expect that the optimal answer is chosen most often, and the sub-optimal answers less frequently, but approximately equally often when compared to the other sub-optimal answers. Each MCQ in a DialogueTrainer scenario has a best practise answer, and multiple sub-optimal answers. The sub-optimal answer options will be referred to as "distractor options" from this point forward as this is a commonly used term in related work [30, 37, 20, 7, 9]. We will use the term “choice frequency” [9, 34] to refer to the distribution of users choosing answering options. A distractor with a low choice frequency is a distractor that is not often chosen com-

pared to other distractors from a given multiple choice question. Multiple choice tests are often used to differentiate between different levels of knowledge. In order for multiple choice tests to achieve this, distractor options need to be good enough to deceive those who do not understand the material well enough to answer the question [34, 29]. Distractors with a choice frequency of 5% or less are considered non-functional, meaning these do not fulfill this role of properly testing knowledge [34, 40, 9, 30]. This research will develop an approach to improving the quality of a DialogueTrainer communication training scenario by trying to replace distractor options with a low choice frequency, by distractor options with a higher choice frequency.

Lastly, DialogueTrainer is experimenting with automation of scenario creation. Currently, the process of creating a scenario takes about 20 hours on average. One of the challenges when creating these scenarios is to create quality distractor options. Automation of (part of) this process could save a lot of resources in the long run. This research will attempt to contribute to the automation process by using machine learning models to generate distractor options. To achieve this we will utilize information about the scenario, the question and the existing distractors for generation.
We also want to experiment with generation based on the desired parameter scores. This would mean that instead of only using information from the question, the desired impact of the answer is also taken into account. For example when the generated output would contain three empathetic and clear replies, adding an option with a very low score on empathy and clarity could result in more variety of generated distractors. If this method proves to be effective, this would make the process of scenario writing a lot more efficient as the scenario writer only has to think of the desired outcome of the distractor option. Moreover, distractor generation based on desired parameter settings would allow the scenario writer more influence on the generation process in the future as machine learning is used more and more on the work floor.

For every problem we want to address, we will use machine learning to generate distractors as it is an explicit goal of this research to investigate the application of automatic generation methods. To achieve this automation, large language models (LLMs) will be used. LLMs have taken the world by storm as they seem to be capable of answering any question. OpenAI has set a new record by achieving 100 million users

within only 2 months with their launch of ChatGPT[2], the most popular LLM currently available. Large language models are models that have been trained on an immense amount of data. In the case of ChatGPT, 570 GB of data, or 300 billion words, were used to train the algorithm[3]. Most of this data consists of text in the form of online articles and books. Models like ChatGPT are specifically trained to excel at dialogue and question answering. This sparked the idea of using LLMs to help DialogueTrainer achieve its goals. At its core LLMs are very simple. They predict the next word based on the words typed so far. The model retrieves context from the input, after which the model will output an answer accordingly. One might expect the results from next word prediction to be quite bad as it is a very simple model that does not have a grasp of context. However, as the size of the training data increases, next word prediction has proven to be a very powerful tool. ChatGPT is remarkably good at answering questions, writing emails, and pretty much any other task that uses natural language [5, 24]. This lead us to believe large language model could help us when automating communication training scenario creation.

This research will explore the use of LLMs to improve one-to-one communication training scenarios provided by DialogueTrainer. To do so, we investigate two main research questions:

RQ1: "Can large language models be used to generate alternative distractor options to achieve higher choice frequency of those distractor options in DialogueTrainer communication training scenarios?"

RQ2: "Can large language models be used to generate dialogue options based on the desired parameter scores in DialogueTrainer communication training scenarios"

[2]https://www.thehindubusinessline.com/info-tech/social-media/chatgpts-popularity-tops-globally-100mn-users-in-2-months/article66470565.ece

[3]https://www.sciencefocus.com/future-technology/gpt-3

2. Literature Review

In this section related work will be explored. First we will explore related work on communication, followed by ways of training communication. Next we analyse related work on distractor options to find out what makes a distractor option good according to the literature. Lastly, we explore similar research from the past in which distractor option generation was attempted.

2.1 Communication & Communication Training

Communication has shown to be important in many different scenarios. Especially in conversations where stakes are high, for example in the medical field, effective communication has had a positive impact on not only patient satisfaction, but also symptom relief and physiological outcomes of care [2, 3]. On top of that studies show a reduction in conflict between doctor and patient as well as positive effects on mental health for the patient. [31, 26] As effective communication has shown such a positive impact, it is important to find out what makes communication effective.

Effective communication consists of three parts: non-verbal, paraverbal and verbal communication [26]. However, training a user in non-verbal and paraverbal communication through a computer simulation is near impossible as it is hard to analyse these attributes of communication without human assistance. Therefore, we will mainly focus on verbal communication.

When performing verbal communication in high stakes scenarios there are a number of things to focus on. Lets assume a scenario in which a boss has to bring bad news to an employee. Firstly, the person initiating the conversation, in this case the boss, should make sure the receiver of bad news knows they are in a safe environment to talk about the high stake matter. One can achieve this by ensuring you know the name of the person you communicate with, greeting the person first. This is done to avoid starting with the loaded topic right away. When a safe environment is established research shows it is important to be as clear and straight to the point as possible. To achieve this goal it is important to show you are empathetic and understanding of the circumstances. The receiver of bad news should feel as though their feelings are valued and understood [14, 26, 11].

Next, let us look at training communication skills. Meta analysis by Reith-Hall et al. shows communication skill acquisition is possible through training [27]. Furthermore, meta analysis on communication training by Berkhof et al. shows role-play, feedback, and small group discussions to be the most effective ways of training communication skills. More passive approaches such as oral presentations about communication or written information on communication have relatively little effect on communication skills [4]. This goes hand in hand with the finding that making mistakes helps to learn effectively [1, 33].
When checking whether the methods used by DialogueTrainer are in line with the methods recommended by literature, we noticed DialogueTrainer applies all parts of effective communication. DialogueTrainer provides a virtual agent with movement and facial expressions to simulate non-verbal communication. On top of that, DialogueTrainer provides audio to simulate paraverbal communication attributes such as tone, pitch and volume. Text is used to allow a user to interact and converse with the agent mimicking verbal communication. Furthermore, the approach used by DialogueTrainer, a role-play simulation, has shown to be relatively effective in training communication skills. In this research we will attempt to make the scenarios more difficult by using automation to improve distractor options, this should help users to learn more effectively.

2.2 Distractor Options

As the main research question of our research is focused on improving the quality of distractor options, we need to investigate what makes a distractor option good or bad. The main consensus from literature is that distractor options should be topical [9, 7]. If the distractor options are related to the same topic as the question, people are more likely to choose these distractor options. Furthermore, the distractor options should be plausible. Lack of plausibility has shown to be the main cause of non-functional distractor options, as an implausible answer option is almost always discarded by an examinee [30, 37, 9, 7].
Next, Gierl et al. [9] state distractor options should have an equal choice frequency as this shows an equal quality of distractor options. Combined with the non-functional distractor threshold this ensures quality distractors, as long as the distractor with the lowest choice frequency is replaced when resolving unbalance in choice frequency between distractors. Lastly, good quality distractor options should be equal to the correct option in structural features such as complexity, length, formatting and grammar [19,

20].

When it comes to the amount of options given, most researchers claim two distractor options is optimal as this increases efficiency for both multiple choice developers and users [28, 39, 8]. However, this claim is mainly based on teachers who have to think of the options themselves. Designing more options requires more time, and hence more resources. On top of that, Rahma et al. show a reduction in the amount of distractor options can reduce the difficulty. However, it should be noted that the reliability of this study was low and more research would be necessary to reach this conclusion [25].

In our research time spent thinking of distractors will not be relevant as the distractor options are generated by a large language model, meaning this limitation will not be relevant in our research.

2.3 Generating Distractor Options

Multiple researchers have tried to generate distractor options with varying success [42, 32, 23, 22]. Qui et al. [23] propose a sequence-to-sequence based model "EDGE" to generate distractors. Their model consists of three key components: the encoding module, the enriching module, and the question reforming module. They first employ an encoding module to extract the contextual semantic representations. This module converts words and sentences into vectors. Next, the enriching module use attention mechanisms to further improve the semantic representations of the question and its answer. More specifically they merged the reading passage information into the question and answer by multiplying the reading passage vector, and question/answer vectors using a scaled dot product and a fusion kernel based on the work by Chen et al. [6]. Finally, the question was reformed in the question reforming module such that any information related to the answer that is included in the question was removed before feeding the information into the decoder. This is done to ensure the generated content are distractor, and therefore incorrect, options. Furthermore, the distance between the generated distractors and the correct answer was measured to verify a significant difference.

Apart from incorrectness, the plausibility of distractors is very important. To ensure a plausible answer, the distractor generator uses the semantic representation of the reformulated question to initialize the generation process. This way EDGE should produce distractors with similar semantic representation. In the results EDGE showed a better performance than the state of the art competition, especially the fluency and

coherence was convincing. However, the ability to distract from the correct answer was still quite weak with an average grade of 5.5 out of 10 given by human experts.

Susanti et al. [32] attempted to generate distractors using a corpus based approach. First they select target word from a reading passage. The answering options are all synonyms to this target word, with one best practise answer and a few distractors. Next, their method consists of three steps: they select a collection of candidate distractors from the corpus. This is done by searching for words that share the same hypernym. For example: dog, is a hypernym for Chihuahua, and poodle. The siblings of the target word should share a similar meaning, but also have a certain difference in meaning. Next these candidates are filtered based on differences in the complexity and length of the distractors compared to the correct answer. Furthermore, synonyms (words with a similar meaning) and antonyms (words with an opposite meaning) of the correct answer are filtered out. Lastly, the distractor options are ranked from best to worst. This ranking is done by aggregating word-embedding based semantic similarity and word collocation information of the distractor candidates with respect to the target word, the reading passage, and the correct answer. Susanti et al., showed that it is possible to generate distractors using their model, however the quality of these distractors seemed low as according to human experts 39 distractors were problematic in a 45 question test, meaning there was at least one low quality distractor in 86% of the questions.

Lastly, Offerijns et al. used GPT2 to generate distractor options [22]. This is perhaps the most interesting related work as the approach of generation is very similar to the approach we will use. Firstly, the GPT2 model was finetuned using the RACE dataset [15]. The training data included context to help replicate stylistic features, the question to make sure the generated distractor options would be relevant to the question and the answer to the question to ensure the generated output would not include the correct answer. Furthermore, the generated distractor options were penalized for similarity to create syntactically dissimilar distractor options. Lastly, the generated distractors were filtered on similarity to avoid duplicated output.
Based on the tests run by the authors this model outperformed all state of the art competition. This is mainly based on automatic semantic evaluation of the distractor options. The human evaluation was also positive, but lacked statistical power as it was very limited in scale. Interestingly, the training was performed using a small scale (117 million parameters) and a medium scale model (355 million parameters) where

the medium sized model only performed marginally better than the small model. This could indicate that the amount of parameters does not significantly improve the performance of the model. However, it is also possible that the scale size difference simply is not large enough.

3. Method

In this section the methods will be explained. To answer the first research question we will compare choice frequencies of human written and AI generated distractors. Participants will be sampled from two courses at Utrecht University that already use DialogueTrainer scenarios. The second research question is answered by generating answer options based on desired parameter settings. Communication experts from DialogueTrainer will analyse the quality of these answer options. The information on participants and sampling can be found in section 3.1. The generation process is explained in section 3.2 and the procedure can be found in section 3.3.

3.1 Participants

The research includes 521 participants, all of which are students from the faculty of social sciences at Utrecht University. These students are sampled by asking them to participate in our study at the start of the course "Professionele Gespreksvoering" which is given in the third (365 students) and fourth quarter(156 students) of the 2023-2024 academic year. Participants are informed about the research and asked for consent using the consent form and information form found in appendix 6.
The "Ethics and Privacy Quick Scan of the Utrecht University Research Institute of Information and Computing Sciences" classified this research as low-risk with no further ethics review or privacy assessment required. This ethics check can be found in appendix 6.

3.2 Model training & distractor generation

We first select a large language model that we will use to generate distractor options for our research. Since DialogueTrainer does not want to share its intellectual property with LLM providers, we chose for an open source model to allow for local training. This means state of the art closed models such as GPT4 and PaLM 2 L could not be used. We opted for a version of the Mistral 2 model that was pre-trained to better facilitate chat interactions. This training data consists of publicly available datasets provided by huggingface. Mistral 2 (trained with 7 billion parameters) outperforms other state of the art open source models such as Llama 2 (13 billion parameters) [36].

This is remarkable as the performance remains better on all benchmarks used, while the number of parameters used is almost half when compared to Llama 2(13B). According to their benchmarks Mistral 2 even outperforms the 34 billion parameter version of Llama 2 in most cases. Mistral 2 seems like a good option for this research as it requires relatively little computing power while performing well on text based interactions.

3.2.1 Finetuning distractor generation

To further fine-tune the model we create a dataset using 73 communication training scenarios provided by DialogueTrainer. This dataset is consists of scenarios containing a number of attributes to give the model as much context as possible. When creating the prompts used to finetune the model, the general description of the scenario was included to give a setting. The question or sentence that should be replied to was included to ensure the generated response is topical. The best practise option was included to make sure the generated distractors are not too similar to the correct answer. Two ancestor statements prior to the moment of generation are included in an attempt to give the model more context. The complete scenario was not included as we were limited by a maximum number of characters for each prompt. On top of that having too much information could make the prompts too convoluted. This means the model has a disadvantage compared to human scenario writers as the model does not know the entire conversation prior to the point of generation. Using all this information the model should output a distractor option with associated feedback. In the training data the desired output is also given with the hope that the model will be able to replicate similar output after the finetuning process. Lastly, the learning rate and number of epochs are determined based on results of experimentation with large language models within DialogueTrainer. The model is trained on a local computer with the following specs: AMD Ryzen 5 3600 6-Core Processor, 64 GB RAM, GPU: NVIDIA GeForce RTX 4060 Ti (16GB), the finetuning process took approximately 48 hours. After the model was finetuned using the aforementioned dataset, the model is given the same input from new scenarios. Using this information the model generates distractor options. Based on the quality of the output, the finetuning process could be repeated with slightly different parameters to increase performance.

3.3 Procedure

3.3.1 RQ 1: Generating alternative distractors

To answer the first research question we compare the choice frequency of non-functioning distractor options in two DialogueTrainer scenarios. We use the data of the participants from the first group to determine which distractor options are non-functioning distractor options [40]. Any option with a choice frequency lower than 5% is labeled as non-functioning. The participants in the second group play scenarios where these non-functioning distractors are replaced by generated ones. On top of that distractors that substantially differ in choice frequency will be replaced by generated distractors similar to the research by Gierl et al. [9]. For example if a question has 3 answers, with a distribution of answer A: 70% B: 24% and C: 6%, then option C will be replaced with a generated distractor even though option C is not a non-functional distractor.
As these scenarios were in Dutch, but the model was trained in English, we first translate the scenario from Dutch to English using DeepL[1]. After translating the scenario with the selected distractors, the model was given the general context of the scenario, the sentence that should be replied to, the ancestor statements and the best practise answer.
With this information the model gives five different candidate distractor options, each with the associated feedback as output. These candidate distractor options are then translated back to Dutch and evaluated by experts to ensure the quality is high enough to be used in a scenario. From evaluation it became clear the feedback was not up to DialogueTrainer standards. We decided to ask experts to rewrite the feedback as the main goal of this research was to generate distractors. The participants in the second course play the DialogueTrainer scenario in which the selected distractor options are replaced with the generated distractor options. The total score is the score given to a player after completing a DialogueTrainer scenario. To test for differences between the groups, this total score will be compared for scenarios that have not been changed. If no significant differences between the groups are found we conclude that any differences in choice frequencies of discriminated distractors are likely caused by changes made for this research. The choice frequencies of the generated distractor options are compared to the original human-written distractor options used in the first course.

[1] https://www.deepl.com/nl/translator

The results are compared using a Mann Whitney U test.

3.3.2 RQ 2: Parameter based distractor generation

To answer the second research question the model is first fine tuned similar to the fine tuning for research question one. This means using the general context of the scenario, the statement that should be replied to and the ancestor statements. On top of that we include the parameter descriptions and changes for each answer option in the fine tuning process. The parameter change was reformatted from numerical to a textual scale. Negative values were changed to "negative", positive values were changed to "positive" and unchanged parameters or parameter changes of zero were changed to "unchanged". This was done as the model is trained on a lot of text and tends to perform better using text when compared to numbers. After fine tuning we attempt to generate distractor options given the question and the desired parameter settings. These answering options are then examined by communication experts and scenario writers from DialogueTrainer to evaluate the quality of the output. To evaluate the output we asked experts to rate 20 different generated answering options as well as the original options. These values are then compared using a Wilcoxon Signed-Rank test to see if there are significant differences. In case there are no significant differences, this indicates the generated options are good enough not to be distinguishable from human-written answering options.

4. Results

This chapter shows the results. The prompt experimentation will be explained in Section 4.1. Results related to distractor generation and replacement can be found under Section 4.2. Results of the parameter based generation can be found under Section 4.3

4.1 Prompt experimentation for distractor replacement

After the initial training, experts evaluated the model's output. Their findings indicated that the model's performance was insufficient for an experiment involving students, who use the DialogueTrainer scenarios as course material. This meant further fine-tuning was necessary. The primary issue was the model's inability to effectively connect newly generated distractors with the subsequent statements. Additionally, the feedback generated was often overly general or inaccurate.

To address these issues, a second round of fine-tuning was conducted. This included incorporating six additional ancestor statements with their associated feedback, bringing the total to eight ancestor statements per prompt. The aim was to provide the model with more examples of high-quality feedback. Furthermore, two descendant statements were added to each prompt to give the model more context for the conversation following the generated statement. For an example of the final prompt refer to Appendix 6

These adjustments resulted in improved output from the model when it comes to the connection between generated distractors and the rest of the conversation. Although in some cases the generated options still did not fit well within the conversation's context, most cases fit a lot better. This allowed the DialogueTrainer expert and the course coordinator to select one suitable distractor from the five generated candidate distractors without needing to adjust the scenario. However, the adjustments were also aimed at targeting the issues regarding the feedback. We expected the additional ancestor statements with associated feedback to improve the feedback quality outputted by the model, however after revision it became clear that the output was still below the DialogueTrainer standards. Therefore, a DialogueTrainer expert was asked to rewrite the feedback as needed.

The parameter based generation was not as strictly revised as DialogueTrainer cus-

tomers would not use this output, meaning there was more room for experimentation. Since the increase in ancestor statements showed promising results these were also used in training for the parameter based generation. An example of the final prompt can be found under Appendix 6. Examples of the output are provided in Table 4.1 to give some insight on the output of the model. In these examples we provided a "Good" and "Bad" example for both research questions. We decided to leave the context of the prompt out as giving this all of this information would cover multiple pages, for an example of a prompt refer to Appendix 6. The "good" examples are labeled as such when the original and generated statements fit well into the context of the scenario. The "bad" examples are often unrelated to the context of the scenario, as shown in the second row, or extremely general to the point it is unclear whether the model understood the context, like shown in row four. Whether these examples are actually end up performing well as distractor options should follow from our statistical data.

Table 4.1: Examples of Good and Bad Responses

	Original	**Generated**
Good example RQ1 (Distractors)	Vind je Utrecht een fijne studentenstad?	Hoe zou je de Utrecht beoordelen als studentenstad?
Bad example RQ1 (Distractors)	Wat is je woonsituatie?	Wat vind je van de kanalen?
Good example RQ2 (Parameters)	No, I think we can get started right away.	We can work together to find a solution that works.
Bad example RQ2 (Parameters)	Unfortunately, a request like this is not automatically a priority.	I understand that this is very important to you.

4.2 RQ1: Distractor generation

This research was performed using 40 cases from two scenarios with 365 participants in the control group and 156 participants in the test group. We first test for differences between the groups. Next, we show the choice frequencies before and after replacing non-functioning distractors. Lastly, we show whether we found significant differences between the generated and non-generated statements.

RQ1: Group Differences

The following statistical tests were conducted to test for differences between the control (students Q3) and test (students Q4) groups. We perform these tests to ensure that we can explain the differences found in the data using the changes we made between

the groups, instead of group differences themselves causing different responses. These tests compare the total scores given after completing a DialogueTrainer scenario. For these statistics we used a scenario that was left unchanged in both groups.

Shapiro-Wilk Test

The Shapiro-Wilk test was used to check the normality of the data for both groups.

Table 4.2: Shapiro-Wilk Test Results

Group	Statistic	p-value
Students Q3	0.9678	1.61×10^{-8}
Students Q4	0.9201	3.15×10^{-10}

Both p-values are less than 0.05, indicating that the normality is rejected in the data in both groups. From this we conclude we should use non-parametic tests to test for differences between the groups.

Skewness and Kurtosis

Table 4.3: Skewness and Kurtosis

Statistic	Skewness	Kurtosis
Students Q3	-0.3652	1.6251
Students Q4	-1.0638	2.7699

The first group has a skewness of -0.3652, indicating it is slightly left-skewed, while the second group has a skewness of -1.0638, indicating a more pronounced left skew. The kurtosis values for both groups (1.6251 for the control and 2.7699 for the test group) suggest the data does not follow the normal bell-curve. This is further evidence that parametric tests are not applicable for this data, hence non-parametric tests will be used.

Mann-Whitney U Test

The Mann-Whitney U test was used to compare the medians of the total scores of the both groups.

The p-value is greater than 0.05, from this information we conclude the null-hypothesis is not rejected. This means there were no significant differences found between the medians of the total scores of students in Q3 and Q4 when analysing our data. In other

Table 4.4: Mann-Whitney U Test

Statistic	Value
U	59316.0
p-value	0.3123

words, significant differences identified in our tests are less likely to be attributed to group differences and are more likely caused by the changes we implemented.

RQ1: Statistical Test Results

Choice frequencies

Firstly, we provide the choice frequencies for each distractor option before and after replacement in Table 4.5. In this table we see the case number indicating a distractor is replaced, the original percentage, which is how often the first group chose the distractor, and the generated percentage, which is how often the generated distractor was chosen by the second group after replacing the original. Non-functioning distractors, distractors where the original percentage is under 5%, were always replaced. Options where the original percentage is higher than 5 and the difference between neighbouring distractors (distractor options distracting from the same best-practise answer) is 20% or bigger were also replaced. Any distractor option in the table where the original percentage is bigger than five, is the distractor with the lowest choice frequency relative to its neighbours and a difference of 20% or bigger when compared to the neighbouring distractors.

Normality

As we want to compare average choice frequencies, which are denoted in percentages, between the two groups it is difficult to test for normality. Percentages are bounded between 0 and 100%. Data in normality tests can theoretically extend infinitely in both directions, the bounded nature of percentages makes it unlikely to follow a normal distribution. On top of that our data contains non-functioning distractor options, meaning these data points cluster between zero and five percent with some additional outliers caused by the options which were selected due to a big difference in choice frequency.
We have chosen not to test for normality using the total scores given after playing a DialogueTrainer scenario. This idea of using the total scores for our normality tests was

Table 4.5: Choice Frequencies for Original and Generated Distractors

Case Number	Original %	Generated %	Case Number	Original %	Generated %
1	10.78	8.33	21	4.32	3.90
2	16.00	13.33	22	1.87	4.08
3	12.57	13.58	23	5.23	7.54
4	1.34	3.16	24	4.05	5.60
5	5.63	10.30	25	2.49	4.80
6	3.42	10.14	26	4.79	6.45
7	13.50	20.00	27	4.00	16.67
8	11.11	57.14	28	10.12	8.93
9	9.09	25.00	29	17.90	5.42
10	6.25	3.63	30	1.23	3.13
11	10.00	37.50	31	4.67	4.08
12	13.29	4.10	32	5.93	3.13
13	10.29	4.88	33	4.97	7.03
14	6.89	8.33	34	2.56	11.11
15	2.75	14.11	35	3.77	3.13
16	4.20	3.17	36	0.31	3.17
17	4.26	7.81	37	2.90	1.88
18	3.97	17.78	38	4.32	6.25
19	1.09	1.28	39	6.78	14.06
20	4.92	8.64	40	1.04	5.81

disregarded as this would mean all unaffected distractor options would influence the normality tests. This could create a normal distribution or refute a normal distribution due to noise in the data. Moreover, these total scores are not the choice frequencies that we want to compare. In theory the total scores could be normally distributed while the choice frequencies are not and vice versa.

As we can not guarantee normally distributed data using traditional normality tests we decided not to report these normality tests and use non-parametric tests for this research. Based on the nature of our data we do not expect the data to be normally distributed. Normally distributed data is unbounded while our choice frequency percentages are bounded between 0 and 100%. Moreover, our data is likely clustered between zero and five percent due our criteria when selecting distractors for replacement. This clustering causes the data to be skewed which is another indicator to use non-parametric tests. Lastly, a normal distribution assumes that most data points are clustered around the mean, with equal variability on either side. In our case this is not expected due to the cases where the choice frequency was higher than five, but differed more than 20 percent from neighbouring distractor options. These options cause outliers on one side of the mean which further reinforces our idea that our data is not

normally distributed and non-parametric tests are most appropriate for this research.

Mann-Whitney U Test

Since we compare two independent groups where we do not expect normality, we use the The Mann-Whitney U test to compare the choice frequencies of original and generated distractors for significant differences in our data.

Table 4.6: Mann-Whitney U Test Results

Statistic	Value
U	602.0
P-value	0.0574

The p-value is slightly above 0.05, suggesting that the difference between the two groups is not statistically significant. However, since this is such a borderline case we will provide a few more tests to hopefully help determine whether the differences are significant.

Wilcoxon Signed-Rank Test

The Wilcoxon Signed-Rank test is used to compare the choice frequencies of two related groups. In our case we do not consider the groups to be related, however the argument could be made that this test is meaningful as the demographic of the groups (bachelor students from Utrecht University) is very similar. That being said the results from the Mann-Whitney U test should be taken as the most conclusive while the Wilcoxon Signed-Rank test is added for additional insight.

Table 4.7: Wilcoxon Signed-Rank Test Results

Statistic	Value
W-statistic	208.0
P-value	0.0058

The p-value is less than 0.05, indicating a significant difference between the two groups, where the generated distractors have a significantly higher choice frequency.

T-test

Even though normality is likely rejected in the data we wanted to include these results as the Mann Whitney-U test is so close to the cutoff of 0.05. Similar to the Wilcoxon

Signed-Rank test these results should be taken with a grain of salt as it is not the most appropriate test for this research.

Table 4.8: T-test Results

Statistic	Value
T-statistic	-2.1531
P-value	0.0360

The p-value is less than 0.05, indicating a significant difference between the two groups.

4.3 RQ2: Parameter based generation

The results regarding the parameter based generation can be found in Tables 4.9 & 4.10. The results from both experts are not combined as expert 2 had prior information on the scenario, thus leaving room for bias. This can also be seen in the grades given to the original statements. Therefor I decided to statistically analyse the results of expert one, while only mentioning the results retrieved from experts two.

4.3.1 Expert Evaluation

In the Tables 4.9 & 4.10 we find the evaluation of the original and generated statements as graded by two DialogueTrainer experts. These values range from zero to five where zero is the lowest, and five is the highest possible score. Scores were determined based on how well the statement fit the scenario and parameter settings, and whether the wording was up to DialogueTrainer standards.

Table 4.9: Expert one: Grades for the original and generated statements

Original	2	5	3	4	4	4	4	3	5	5	4	4	4	3
Generated	4	2	4	2	4	3	3	5	2	3	3	1	1	4

Table 4.10: Expert two: Grades for the original and generated statements

Original	5	5	5	5	5	5	4	4	4.5	5	5	5	3	5
Generated	2	4	2.5	4	4	5	4	3	4.5	2	5	2.5	2	5

Table 4.11: Wilcoxon Signed-Rank Test Results

Statistic	Value
W-statistic	21.0
P-value	0.0830

Non-Parametric Test Results

Wilcoxon Signed-Rank Test

The p-value is greater than 0.05, suggesting that the difference between the original and generated statements is not statistically significant.

Mann-Whitney U Test

As the Signed-Rank test showed results on the border of significance a Mann-Whitney U test was done to get more insights in these results. Keep in mind a Mann-Whitney U test is not appropriate here as the groups are related since the grades for both generated and original statements are given by the same person.

Table 4.12: Mann-Whitney U Test Results

Statistic	Value
U-statistic	142.0
P-value	0.0370

The p-value is less than 0.05, indicating a significant difference between the two groups.

5. Discussion & Future Work

In this paper we attempted to use large language models to automate, parts of, the communication training scenario writing process. In the results section we found that our work did not produce significant differences in the tests we ran on distractor replacement, finding a p-value which was just barely higher the threshold alpha value of 0.05. When analyzing the parameter based generation we found significant differences in favour of human-written statements. In this is section we will discuss these results. We will first discuss the prompt engineering and model fine-tuning process. Next, distractor replacement is discussed followed by parameter based generation. Lastly, we discuss opportunities for future work

5.1 Prompt engineering & fine-tuning

In the process of writing and revising a prompt that shows the most promising results a few things are to be noted. Firstly, it is very difficult to decide when a prompt is optimally worded. There are no real ways of testing a prompt other than using that prompt. This means all results found by these prompts have to be compared to each other to find out which works better. This in turn leads to possibly using sub-optimal prompts, for example due to finding a local minimum. While testing and revising the prompt different amounts of text and parameters were used, each with varying results. After some testing we found a prompt that seemed to perform the best. This prompt uses ancestor and descendant statements to help steer the model towards a more specific output.

Specificity was a big challenge while fine-tuning our model since most of the scenarios were specifically tailored towards training some aspect of communication. Large language models are known to be quite general as the model is trained on such a wide variety of data. The prompts used in this paper tended to get quite elaborate with all the information provided. It is possible that due to the sheer length of the prompts, the model could not accurately asses which information is most important. As mentioned, we achieved more specificity using both ancestor, and descendant statements. However, some of the output still seems very general. Apart from the length of our prompts, this could also be caused by the nature of large language models. LLMs are

trained on a huge amount of data which leads to output that is often not very specific, often averaging many different opinions on a topic.

5.2 Non-Functional distractor replacement

The first question we tried to answer in this research was: "Can large language models be used to generate alternative distractor options to achieve higher choice frequency of those distractor options in DialogueTrainer communication training scenarios?". When analysing the results, the answer to this question is still indecisive. Firstly, we tested whether there were differences between the different groups of students. Luckily our tests showed no significant differences with a p-value of 0.31. This means that if we find significant differences between the groups in our test cases, these differences are likely caused by the changes we made.

Next, we addressed normality in our data. We want to analyse choice frequencies which are made up of percentages between 0 and 100. We decided against using the total scores provided after playing a DialogueTrainer scenario to test for normality. This decision was made because there are too many factors which could influence the outcome of this normality test. The total score includes every chosen option meaning the majority of this score is influenced by answering options different from the non-functioning distractors we want to test. Simply using the data points from participants to compare to other participants is not possible as we compare averages of these choice frequency. Therefor, we decided to use a textual explanation to indicate whether or not our data is likely normally distributed without a statistical test to back this up. In our selection process we included some boundaries to define non-functioning distractors. This means most of our data will fall between the zero and five percent interval. Apart from that, there will be a few cases which are above five percent due to a big discrepancy between neighbouring choice frequencies. As explained in the results we do not expect our data to be normally distributed due to the nature of percentages. Percentages are bounded which contradicts properties of data that is normally used in normality tests, such as an even distribution around the mean. Our data is likely skewed due to the selection process of distractors which gives more reason to believe normality is rejected in our data. For these reasons we decided to use non-parametric tests to determine whether there are significant differences in our data.

From the Mann-Whitney U test we get a p-value of 0.057. As this value is very close to the 0.05 cutoff we did not want to disregard the possibility of an effect right

away. Because of this we also provided a Wilcoxon Signed-Rank test. This test should normally be used when there is a relation between groups. As we had similar participants in both groups with both groups containing exclusively bachelor students from the Utrecht University, as well as having such a borderline significant result it was decided to include this test as well. From the Signed-Rank test we find a p-value of 0.006. As mentioned this test is not perfect for the job and should therefor be taken with a grain of salt. However, this does show our method could have potential. Lastly, a t-test was also performed in the hopes to provide some more clarity about our results. As mentioned we rejected normality and thus these results are far from definitive, however it does show significant differences with a p-value of 0.036.

From the tests performed we can conclude the method has potential. However, more research should be provided to make more decisive conclusions. Looking at the choice frequencies, the average choice frequency increased. This increase in average choice frequency by itself is not conclusive as there are a few cases where the choice frequency increases immensely. For example in case 11 in Table 4.5 we see the choice frequency rise from 10 to 37.5 percent. These massive increases are caused by branches in the scenario which a relatively small amount of people reached, for example due to differences in choices earlier on in the scenario. These cases have a relatively big influence on the average choice frequency increase overall. However, the results definitely show potential as,in combination with an increase in average choice frequency, the choice frequency increased in 28 out of 40 cases. It should be noted that this paper makes the assumption that a scenario has a higher quality when the scenario is more difficult as this shows from previous research. However, when concluding the method shows potential, this is based on increases in choice frequency. It is possible that a higher choice frequency does not necessarily mean a higher quality scenario. Future research would be necessary to make more conclusive statements on the topic.

Next, the scenarios used for this research were made in collaboration with the Utrecht University. These scenarios are used by in a bachelor course which is taught twice a year as mentioned in the methods. This was very convenient for this research as we could tests our hypotheses in a realistic scenario with real clients from DialogueTrainer instead of setting up a mock scenario with participants for scientific research. However, this also meant that the scenarios are very specific towards the goals of teaching the students their material. As a result the scenarios are not stereotypical DialogueTrainer scenarios. Instead these scenarios are specifically designed with the

goal of helping students learn certain aspects of communication training, from the psychologist point of view. This means that significant results in this experiment might not translate to DialogueTrainer scenarios which serve a more general purpose and use clients that do not have prior knowledge of psychology. We expect our model to have performed worse due to the specificity of the scenarios. In many cases the model gave an output that was going in the right direction, while remaining too general. It is possible that this was in part caused by the scenarios used. On the contrary, it is possible that our model showed more promising results due to the scenarios used. Maybe the non-functioning distractors were significantly worse than non-functioning distractors in other DialogueTrainer scenarios. We did not find any evidence that suggests this, but more analysis would be required to fully rule this out.

Lastly, the scenarios used were all written in Dutch. To use our model we first had to translate the Dutch prompts. Then use our model to generate English output with these translated prompts, and lastly translate the output back to Dutch to use the output in the scenarios. DeepL was used to perform these translations. Even though this program does a phenomenal job at translating, it is definitely possible that some of the meaning was lost or changed as everything was translated twice. When deciding on the model to use for this paper there were no Dutch open source models available that were up to the standards of state of the art models. However, when these models undoubtedly will be created in the future, this could further improve the results of research similar to this. It would be interesting to see whether using a Dutch-trained model would significantly outperform an English-trained model that uses translation for Dutch output.

5.3 Parameter based generation

The second research question we tried to answer in this paper was: "Can large language models be used to generate dialogue options based on the desired parameter scores in DialogueTrainer communication training scenarios". This research was more exploratory as we chose not to perform experiments using participants due to time constraints. Instead we asked experts from DialogueTrainer to perform an experiment inspired by the Turing test. From these tests we found borderline significant differences, which in this case meant the experts graded the human written statements significantly higher than the generated output. During the experiment it became clear that one of the experts had written the scenario used in the experiment. Therefor it was decided to exclude these results from the statistical tests as this introduced a huge

bias. When analysing the results gathered from expert two, it is quite clear the grades for the human written responses are very high. This makes sense considering this expert had written these statements. However, the generated output still received a decent grade with an average grade 3.54 out of 5. Moreover, in 4 out of 14 cases the expert assessed the generated output as equally good when compared to the statements she wrote. These are promising results as the model only used parameter settings and some general context to generate these statements.

When analysing the results from expert one it is clear the model in not up to the standards set by DialogueTrainer professionals. The model achieved an equal or higher score in 5 out of 14 cases with an average grade of 2.93 out of 5. Although this grade seems promising, the individual cases show the model lacks consistency. Some cases were graded five out of five while others are graded one out of five. If DialogueTrainer wanted to implement this model, it should always produce an output that is at least sufficient, and ideally up to the standards of a human expert.

Another reason the model output has room to improve was the data used. To perform this research a dataset was created using DialogueTrainer scenarios. These scenarios were then transferred into prompts using the parameters, ancestors and the general context of the scenario. The problem with this dataset was the parameter settings. There was very little consistency within the parameter settings. In some cases the scenario was in Dutch while the parameters were in English or vice versa. In some cases the parameters were denoted in words such as "high" and "low" while in other cases the parameters were numerical. There were even cases where it seemed the scenario writers made a mistake while assigning these values. This leads to situations where an option should add or subtract one or two from the parameter value where some options influence the parameter score by significantly higher values instead. For this experiment we did our best to clean the data to a point where it could be used, but we can not guarantee the data to be 100% clean. We would advice DialogueTrainer to set more consistent parameters before using this data to train or fine-tune AI models.

5.4 Future Work

Based on the findings in this paper there are a number of things that could be interesting in future work. Firstly, it would be interesting to see this research performed with different communication training scenarios. Currently the scenarios as well as all the training data is material provided by DialogueTrainer. It would be interesting to

see whether the same methods would result in similar performance when used in a different setting. This would also allow for conclusions to be more robust as in theory it is possible that all these results are specific to DialogueTrainer scenarios.

Secondly, we found one main issue with our output while performing the generation process. The output often lacked specificity in the results. There are a number of possible explanations as to why this has occurred. It is possible that too much data was used in the fine-tuning process resulting in overfitting. DialogueTrainer scenarios use a lot of statements such as: "I can imagine this is difficult for you". These statements are not bad to use in certain situations. However, they should not be used more than a few times according to DialogueTrainer experts. In section 5.1 we explain how adding descendants to the prompt increased the specificity of the output. For future work it would be interesting to further investigate achieving more specific output. For example, one could experiment with using more or less information in the prompt. Using the same information with different wordings of the prompt, or find a way to increase salience of specific parts of the prompt which contain the most important information.

Thirdly, it would be interesting to perform our research again with a number of different large language models. We used Mistral in this paper, which is a mainly English trained model. This lead to the decision to translate the Dutch scenarios to English to be to used in our prompt, and back to Dutch after the output was generated. It would be interesting to see how this method holds up against a model that is trained using Dutch data. Moreover, the world of artificial intelligence and large language models is rapidly evolving. While writing this paper Llama 3 was released showing very promising results. There is no doubt there will be many more models which will perform better and better over time. It would be interesting to choose a number of those models and compare the outputs. For example, we did not use closed source models due to restrictions in the use of DialogueTrainer data. How would a state of the art closed source model compare to a state of the art open source model on this task?

Next, performing the same tests regarding generated distractor options on a bigger scale could rule out some of the indecisiveness found in our conclusions. This research was performed using 40 cases from two scenarios and about 521 participants. Doing a similar research with a bigger number of participants as well as more scenarios with a wider range of topics of the scenarios could be interesting as it is possible that the results found here are possibly noise in the data. Moreover, the demographics of the

participants used in this research are very similar as they are all students from the same Bachelor course. Using random sampling for this research was difficult due to limited resources as well as limited availability with DialogueTrainer being a paid product. Doing the research using random sampling with a big enough sample size was simply not feasible. For future research it would be interesting to perform the same research using random sampling and compare results to those found in our paper.

Lastly, the research question on parameter based design was still quite exploratory in this paper. Especially since one of the DialogueTrainer experts was biased in evaluation due to prior knowledge about the scenario. For future research it would be interesting to let a neutral third party evaluate the generated results to see how they compare to human-written statements. Although we think the results seem very promising, it is very hard to evaluate the results provided as the possibilities of introduced biased are very high. Not only did the human expert have information on the scenario prior to our version of the Turing test. Both experts also knew the researcher conducting the research possibly leading to bias where they give answers which they think is the researchers desired outcome. For future research it is recommended to select third party experts to review and critique the outcomes of the large language model.

6. Conclusion

Effective communication training is a critical component in various professional and educational settings. It creates better understanding, and improves overall performance. The ability to communicate effectively is essential for collaboration, problem-solving, and leadership. Given its importance, continuous improvement in communication training methods is vital.

This thesis explored the potential of large language models in communication training scenarios. The first research question was aimed at determining if these AI-generated distractors could achieve a higher choice frequency of distractors. The second research question was directed at automation based on parameter settings. This would allow scenario writers to have influence on generated content, while still allowing for automation in order to reduce the amount of resources required to write DialogueTrainer communication training scenarios.

This thesis shows results that are around the border of significant results, indicating that LLMs could have the potential to increase the effectiveness of communication training by providing more challenging distractors.

Although the results are promising, they are not yet conclusive. The borderline significance suggests that there is potential but also highlights the need for further research as these results could be mere noise in the data. To obtain more decisive results, it is crucial to expand the study to include a larger and more diverse set of cases. A broader dataset would allow for a more in depth analysis and provide greater statistical power to detect significant effects. On top of that testing on a group that is more varied in demographic would make a stronger case for a significant result.

In summary, while the current findings are encouraging, more extensive research is necessary to fully realize the potential of large language models in enhancing communication training scenarios. The integration of large language models in communication training holds promise. However, it requires more investigation to ensure its efficiency and reliability as currently human evaluation is still required to ensure high quality output.

Bibliography

[1] Merav Ahissar and Shaul Hochstein. "Task difficulty and the specificity of perceptual learning". In: *Nature* 387.6631 (May 1997), pp. 401–406. ISSN: 1476-4687. DOI: 10.1038/387401a0. URL: https://doi.org/10.1038/387401a0.

[2] C. Bachmann et al. "Development and national consensus finding on patient-centred high stakes communication skills assessments for the Swiss Federal Licensing Examination in Medicine". In: *Patient Education and Counseling* 104.7 (2021), pp. 1765–1772. ISSN: 0738-3991. DOI: https://doi.org/10.1016/j.pec.2020.12.003. URL: https://www.sciencedirect.com/science/article/pii/S0738399120306674.

[3] Cadja Bachmann et al. "A European consensus on learning objectives for a core communication curriculum in health care professions". In: *Patient Education and Counseling* 93.1 (2013), pp. 18–26. ISSN: 0738-3991. DOI: https://doi.org/10.1016/j.pec.2012.10.016. URL: https://www.sciencedirect.com/science/article/pii/S0738399112004235.

[4] Marianne Berkhof et al. "Effective training strategies for teaching communication skills to physicians: An overview of systematic reviews". In: *Patient Education and Counseling* 84.2 (2011), pp. 152–162. ISSN: 0738-3991. DOI: https://doi.org/10.1016/j.pec.2010.06.010. URL: https://www.sciencedirect.com/science/article/pii/S0738399110003691.

[5] Som Biswas. "ChatGPT and the Future of Medical Writing". In: *Radiology* 307.2 (2023). PMID: 36728748, e223312. DOI: 10.1148/radiol.223312. eprint: https://doi.org/10.1148/radiol.223312. URL: https://doi.org/10.1148/radiol.223312.

[6] Qian Chen et al. "Neural Natural Language Inference Models Enhanced with External Knowledge". In: Jan. 2018, pp. 2406–2417. DOI: 10.18653/v1/P18-1224.

[7] Juliana D'Sa and Maria Visbal-Dionaldo. "Analysis of Multiple Choice Questions: Item Difficulty, Discrimination Index and Distractor Efficiency". In: *International Journal of Nursing Education* 9 (July 2017), p. 109. DOI: 10.5958/0974-9357.2017.00079.4.

[8] Ana Delgado and Gerardo Prieto. "Further Evidence Favoring Three-Option Items in Multiple-Choice Tests". In: *European Journal of Psychological Assessment* 14 (Jan. 1998), pp. 197–201. DOI: 10.1027/1015-5759.14.3.197.

[9] Mark Gierl et al. "Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review". In: *Review of Educational Research* 87 (Aug. 2017), p. 0034654317726529. DOI: 10.3102/0034654317726529.

[10] Anushree Gupta. "Quality of assessment of sequences of choices in serious game DialogueTrainer". In: Utrecht, the Netherlands, Jan. 2019.

[11] Jennifer Fong Ha and Nancy Longnecker. "Doctor-Patient Communication: A Review". In: *Ochsner Journal* 10.1 (2010), pp. 38–43. ISSN: 1524-5012. eprint:

https://www.ochsnerjournal.org/content/10/1/38.full.pdf. URL: https://www.ochsnerjournal.org/content/10/1/38.

[12] Johan Jeuring et al. "Communicate! — A Serious Game for Communication Skills —". In: 9307 (Nov. 2015), pp. 513–517. DOI: 10.1007/978-3-319-24258-3_49.

[13] Sylvia Kasperink. "Assessing validity of the serious game Communicate!: An argument-based approach to validation". In: Utrecht, the Netherlands, 2017.

[14] Lambrini Kourkouta and Ioanna V Papathanasiou. "Communication in nursing practice". en. In: *Mater Sociomed* 26.1 (Feb. 2014), pp. 65–67.

[15] Guokun Lai et al. "RACE: Large-scale ReAding Comprehension Dataset From Examinations". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 785–794. DOI: 10.18653/v1/D17-1082. URL: https://aclanthology.org/D17-1082.

[16] Raja Lala, Gemma Corbalan, and Johan Jeuring. "Evaluation of Interventions in Blended Learning Using a Communication Skills Serious Game". In: *Games and Learning Alliance*. Ed. by Antonios Liapis et al. Cham: Springer International Publishing, 2019, pp. 322–331. ISBN: 978-3-030-34350-7.

[17] Raja Lala et al. "Scenarios in virtual learning environments for one-to-one communication skills training". In: *International Journal of Educational Technology in Higher Education* 14.1 (May 2017). DOI: 10.1186/s41239-017-0054-1.

[18] Tonje Lundeby, Pål Gulbrandsen, and Arnstein Finset. "The Expanded Four Habits Model—A teachable consultation model for encounters with patients in emotional distress". In: *Patient Education and Counseling* 98.5 (2015), pp. 598–603. ISSN: 0738-3991. DOI: https://doi.org/10.1016/j.pec.2015.01.015. URL: https://www.sciencedirect.com/science/article/pii/S0738399115000464.

[19] Ruslan Mitkov and Le Ha. "Computer-aided generation of multiple-choice tests". In: vol. 2. May 2003, pp. 17–22. DOI: 10.3115/1118894.1118897.

[20] Ruslan Mitkov et al. "Semantic similarity of distractors in multiple-choice tests". In: (Nov. 2009), pp. 49–56. DOI: 10.3115/1705415.1705422.

[21] Brett R.C. Molesworth and Dominique Estival. "Miscommunication in general aviation: The influence of external factors on communication errors". In: *Safety Science* 73 (2015), pp. 73–79. ISSN: 0925-7535. DOI: https://doi.org/10.1016/j.ssci.2014.11.004. URL: https://www.sciencedirect.com/science/article/pii/S0925753514002732.

[22] Jeroen Offerijns, Suzan Verberne, and Tessa Verhoef. *Better Distractions: Transformer-based Distractor Generation and Multiple Choice Question Filtering*. 2020. arXiv: 2010.09598 [cs.CL].

[23] Zhaopeng Qiu, Xian Wu, and Wei Fan. "Automatic Distractor Generation for Multiple Choice Questions in Standard Tests". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by Donia Scott, Nuria Bel, and Chengqing Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 2096–2106. DOI: 10.18653/v1/2020.coling-main.189. URL: https://aclanthology.org/2020.coling-main.189.

[24] Basit Qureshi. *Exploring the Use of ChatGPT as a Tool for Learning and Assessment in Undergraduate Computer Science Curriculum: Opportunities and Challenges*. 2023. arXiv: 2304.11214 [cs.CY].

[25] Nourelhouda A A Rahma et al. "Comparison in the quality of distractors in three and four options type of multiple choice questions". en. In: *Adv Med Educ Pract* 8 (Apr. 2017), pp. 287–291.

[26] Piyush Ranjan, Archana Kumari, and Avinash Chakrawarty. "How can Doctors Improve their Communication Skills?" en. In: *J Clin Diagn Res* 9.3 (Mar. 2015), JE01–4.

[27] Emma Reith-Hall and Paul Montgomery. "Communication skills training for improving the communicative abilities of student social workers". en. In: *Campbell Syst Rev* 19.1 (Feb. 2023), e1309.

[28] Michael Rodriguez. "Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research". In: *Educational Measurement: Issues and Practice* 24 (June 2005), pp. 3–13. DOI: 10.1111/j.1745-3992.2005.00006.x.

[29] Michael C. Rodriguez, Ryan J. Kettler, and Stephen N. Elliott. "Distractor Functioning in Modified Items for Test Accessibility". In: *SAGE Open* 4.4 (2014), p. 2158244014553586. DOI: 10.1177/2158244014553586. eprint: https://doi.org/10.1177/2158244014553586. URL: https://doi.org/10.1177/2158244014553586.

[30] Madiha Sajjad, Samina Iltaf, and REHAN KHAN. "Nonfunctional distractor analysis: An indicator for quality of Multiple choice questions". In: *Pakistan Journal of Medical Sciences* 36 (June 2020). DOI: 10.12669/pjms.36.5.2439.

[31] M A Stewart. "Effective physician-patient communication and health outcomes: a review". en. In: *CMAJ* 152.9 (May 1995), pp. 1423–1433.

[32] Yuni Susanti et al. "Automatic distractor generation for multiple-choice English vocabulary questions". In: *Research and Practice in Technology Enhanced Learning* 13 (Dec. 2018). DOI: 10.1186/s41039-018-0082-z.

[33] John Sweller. "Cognitive load theory, learning difficulty, and instructional design". In: *Learning and Instruction* 4.4 (1994), pp. 295–312. ISSN: 0959-4752. DOI: https://doi.org/10.1016/0959-4752(94)90003-5. URL: https://www.sciencedirect.com/science/article/pii/0959475294900035.

[34] Marie Tarrant, James Ware, and Ahmed M Mohammed. "An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis". In: *BMC Medical Education* 9.1 (2009). DOI: 10.1186/1472-6920-9-40.

[35] Lauren J. Taylor et al. "Using Implementation Science to Adapt a Training Program to Assist Surgeons with High-Stakes Communication". In: *Journal of Surgical Education* 76.1 (2019), pp. 165–173. ISSN: 1931-7204. DOI: https://doi.org/10.1016/j.jsurg.2018.05.015. URL: https://www.sciencedirect.com/science/article/pii/S1931720418300783.

[36] Mistral AI team. *Mistral 7B The best 7B model to date, Apache 2.0*. 2023.

[37] Silvia Testa, Anna Toscano, and Rosalba Rosato. "Distractor Efficiency in an Item Pool for a Statistics Classroom Exam: Assessing Its Relation With Item Cognitive Level Classified According to Bloom's Taxonomy". en. In: *Front Psychol* 9 (Aug. 2018), p. 1585.

[38] Jessica L Tracy, Daniel Randles, and Conor M Steckler. "The nonverbal communication of emotions". In: *Current Opinion in Behavioral Sciences* 3 (2015).

Social behavior, pp. 25–30. ISSN: 2352-1546. DOI: https://doi.org/10.1016/j.cobeha.2015.01.001. URL: https://www.sciencedirect.com/science/article/pii/S235215461500011X.

[39] Rashmi Vyas and Avinash Supe. "Multiple choice questions: A literature review on the optimal number of options". In: *The National medical journal of India* 21 (Nov. 2007), pp. 130–3.

[40] Doctor Wajeeha et al. "Difficulty Index, Discrimination Index and Distractor Efficiency in Multiple Choice Questions". In: *annals of pims* 4 (Jan. 2018), ISSN:1815–2287.

[41] Canan P. Zeki. "The importance of non-verbal communication in classroom management". In: *Procedia - Social and Behavioral Sciences* 1.1 (2009). World Conference on Educational Sciences: New Trends and Issues in Educational Sciences, pp. 1443–1449. ISSN: 1877-0428. DOI: https://doi.org/10.1016/j.sbspro.2009.01.254. URL: https://www.sciencedirect.com/science/article/pii/S1877042809002572.

[42] Xiaorui Zhou, Senlin Luo, and Yunfang Wu. *Co-Attention Hierarchical Network: Generating Coherent Long Distractors for Reading Comprehension*. 2019. arXiv: 1911.08648 [cs.CL].

Appendix A

Prompt Examples

```
r_prompt = f"""<s>[INST]You are an assistant for creating Conversational Scenarios.
In every Conversational Scenario there are playerStatements, computerStatements and situationStatements.
Every Conversational Scenario has a Description which states its objective.
Distractor options are playerStatement options which are sub-optimal replies to a computerStatement.
The purpose of these options is to distract a user from the best practise answer by proposing an
answer which is plausible and similar to, yet different from the optimal answer.
You will be provided the scenario description, previous statements from the scenario until this point, the computerStatement
which you generate a distractor response to along with the asociated emotion, the best practice
playerStatement response, and the decendant computerStatement which is the reply to your generated distractor.
Your job is to generate incorrect “distractor” options, which should be similar to, yet different
from the best practice answer playerStatement's response. The distractor option should be a sensible answering option
when regarding the decendant computerStatement.
Please limit your answer to the playerStatement response with the associated feedback.
You don't need to say hello, give an introduction, or explain your answer.

The Scenario Description is: "{description}"
The previous scenario statements from the conversation until this point are:
{ancestors}
The computerStatement you respond to is:
-computerStatement: {origin_statement}
The emotion associated with this computerStatement is: {computer_emotion}
The best practise answer to this computerStatement is:
-playerStatement: {bestPractise}
The decendant computerStatement response to your generated distractor is:
-computerStatement: {distractorDecendants}
Generate five different distractor playerStatements along with its feedback.
[/INST]
```

Figure 1: Example of distractor generation prompt

```
    r_prompt = f"""<s>[INST]You are an assistant for creating Conversational Scenarios."
In every Conversational Scenario there are playerStatements, computerStatements and situationStatements.
Every Conversational Scenario has a Description which states its objective.
You will be provided the scenario description, the parameter change which the distractor should cause, the ancestor statements of
the conversation untill this point with the asociated, the computerStatement which you generate a distractor response to, and the
emotion asociated with the computer statement.
Your job is to generate a playerStatement response option, which should fit the proposed parameter changes and be a sensible
answering option to the computerStatement.
Please limit your answer to the playerStatement response. You don't need to say hello, give an introduction, or explain your answer.

The Scenario Description is: "{description}"
The Scenario parameter names, descriptions and value changes are:
{parameter_name_desc_change}
The previous scenario statements from the conversation until this point are:
{ancestors}
The computerStatement you respond to is:
{origin_statement}
The emotion associated with this computerStatement is: {computer_emotion}
Generate a playerStatement along with its feedback.
[/INST]
"""
```

Figure 2: Example of feedback-based generation prompt

Appendix B

Consent Form

Universiteit Utrecht

Consentformulier voor deelname aan het onderzoeksproject

“Exploring the use of large language models to improve one-to-one communication training scenarios”

Lees de onderstaande tekst zorgvuldig door. Wanneer de tekst hebt gelezen en het met de content eens bent kan je hieronder toestemmen met het onderzoek.

Ik bevestig dat ik 18 jaar of ouder ben. Ik bevestig dat het onderzoeksproject “Exploring the use of large language models to improve. One-to-one communication training scenarios” aan mij schriftelijk is uitgelegd. Verder ga ik ermee akkoord dat het materiaal dat ik bijdraag wordt gebruikt om inzichten te genereren voor het onderzoeksproject “Exploring the use of large Language models to verbetering one-to-one communication training scenarios”.

Ik begrijp dat persoonlijke gegevens van mij worden verzameld zoals uitgelegd in het informatieblad en dat deze gegevens vertrouwelijk worden behandeld, zodat alleen Jens Hartkamp en Johan Jeuring toegang hebben tot deze gegevens en deze tot mij persoonlijk kunnen herleiden. De gegevens worden bewaard in een met een wachtwoord beveiligd gegevensbestand gedurende maximaal 4 weken. Daarna worden ze volledig geanonimiseerd. Conform de Algemene Verordening Gegevensbescherming (AVG) kan ik op elk moment tijdens deze periode inzage krijgen in mijn persoonsgegevens en kan ik verzoeken deze te verwijderen.

Daarnaast begrijp ik dat mijn deelname aan dit onderzoek vrijwillig is en dat ik me tot 4 weken na de start van de cursus kan terugtrekken uit het onderzoek zonder opgaaf van reden, en dat als ik me in die periode terugtrek alle persoonlijke gegevens die al van mij zijn verzameld, zullen worden gewist.

Ik begrijp dat mijn deelname geen vereiste is voor deze cursus en dat het wel of niet deelnemen geen gevolgen voor mij heeft. Ik geef toestemming om de volledig geanonimiseerde gegevens te gebruiken in toekomstige publicaties en andere wetenschappelijke middelen om de bevindingen van het onderzoeksproject te verspreiden.

Ik ga akkoord om deel te nemen aan het bovenstaande onderzoeksproject over “Exploring the use of large language models to improve one-to-one communication training scenarios”.

Information Form

Universiteit Utrecht

Onderzoek participant informatieformulier

"Exploring the use of large language models to improve one-to-one communication training scenarios"
22-01-2024

1. Introductie

We vragen je toestemming om gebruik te mogen maken van je onderwijsdata voor een wetenschappelijk onderzoek voor een Masterthesis bij de opleiding AI. Het onderzoek wordt uitgevoerd met een communicatietraining applicatie "DialogueTrainer" die in deze cursus wordt gebruikt.

2. Wat is de achtergrond en het doel van dit onderzoek?

In dit onderzoek zullen we proberen de communicatietraining applicatie van DialogueTrainer te verbeteren met behulp van kunstmatige intelligentie. Het doel is om een tool te creëren die tekst kan genereren om te helpen in het proces van het maken van een één-op-één communicatie trainingsscenario's.

3. Wie voert dit onderzoek uit?

Dit onderzoek wordt uitgevoerd door Jens Hartkamp (j.l.l.hartkamp@students.uu.nl) als deel van mijn Masterthesis onder begeleiding van Johan Jeuring (j.t.jeuring@uu.nl). Het onderzoek wordt uitgevoerd vanuit de Universiteit Utrecht in samenwerking met DialogueTrainer.

4. Hoe zal het onderzoek worden uitgevoerd?

In deze cursus oefen je gespreksvaardigheden met virtuele karakters in de DialogueTrainer app. Als je meedoet aan dit onderzoek wordt de data, die bestaat uit multiple-choice keuzes, gebruikt voor het trainen van een model. De oefengesprekken maken deel uit van deze cursus, waardoor er geen extra tijd nodig is om aan het onderzoek deel te nemen. We bieden geen compensatie voor dit onderzoek.

5. Wat doen wij met je data?

Als je akkoord gaat met dit onderzoek, bewaren wij de door jou verstrekte gegevens veilig, het gaat om het e-mailadres dat is gekoppeld aan je ULearning-account, je solis-id en je multiple-choice keuzes in de gesprekken met de virtuele karakters. Wij slaan het e-mailadres en solis-id apart van de rest van je gegevens op en gebruiken een pseudoniem(id) om de twee bestanden te koppelen. Dit wordt gedaan zodat deelnemers zich tot 4 weken na het invullen van dit formulier kunnen afmelden voor het onderzoek, terwijl de gegevens die we gebruiken anoniem blijven. Wij verwijderen het e-mailadres na 4 weken, waardoor je gegevens daarna volledig anoniem zijn.

6. Wat zijn je rechten?

Deelname aan dit onderzoek is vrijwillig. Wij mogen je gegevens voor ons onderzoek alleen verzamelen als je daar toestemming voor geeft. Als je besluit niet deel te nemen, hoef je geen verdere actie te ondernemen. Je hoeft niets te ondertekenen. Ook hoef je niet uit te leggen waarom je niet wilt meedoen. Als je besluit deel te nemen, kun je tot 4 weken na de start van de cursus van gedachten veranderen en je deelname stopzetten door een mail te sturen naar Jens Hartkamp: j.l.l.hartkamp@students.uu.nl. Je gegevens zullen dan niet gebruikt worden voor het onderzoek.

7. Goedkeuring van het onderzoek

Dit onderzoek is goedgekeurd door het Onderzoeksinstituut Informatie- en Informatica op basis van een Ethiek en Privacy Quick Scan. Heb je een klacht over de wijze waarop dit onderzoek wordt uitgevoerd, stuur dan een

e-mail naar: ics-ethics@uu.nl. Heb je klachten of vragen over de verwerking van persoonsgegevens, dan kan je een e-mail sturen naar de Privacy Officer van de Faculteit der Wiskunde en Natuurwetenschappen: privacy-beta@uu.nl. De Privacy Officer kan je ook helpen bij het uitoefenen van de rechten die je hebt op grond van de AVG. Voor alle details over onze wettelijke basis voor het gebruik van persoonsgegevens en de rechten die je heeft over je gegevens, kan je de privacyinformatie van de universiteit raadplegen op www.uu.nl/organisatie/privacy.

8. Meer informatie over dit onderzoek?

Als je vragen of klachten hebt over dit onderzoek, stuur dan een email naar Jens Hartkamp via j.l.l.hartkamp@students.uu.nl of mijn begeleider Johan Jeuring via j.t.jeuring@uu.nl

9. Appendix:

Consent formulier

Appendix C

Ethics Check

Response Summary:

Section 1. Research projects involving human participants

P1. Does your project involve human participants? This includes for example use of observation, (online) surveys, interviews, tests, focus groups, and workshops where human participants provide information or data to inform the research. If you are only using existing data sets or publicly available data (e.g. from Twitter, Reddit) without directly recruiting participants, please answer no.

- Yes

Recruitment

P2. Does your project involve participants younger than 18 years of age?

- No

P3. Does your project involve participants with learning or communication difficulties of a severity that may impact their ability to provide informed consent?

- No

P4. Is your project likely to involve participants engaging in illegal activities?

- No

P5. Does your project involve patients?

- No

P6. Does your project involve participants belonging to a vulnerable group, other than those listed above?

- No

P8. Does your project involve participants with whom you have, or are likely to have, a working or professional relationship: for instance, staff or students of the university, professional colleagues, or clients?

- No

Informed consent

PC1. Do you have set procedures that you will use for obtaining informed consent from all participants, including (where appropriate) parental consent for children or consent from legally authorized representatives? (See suggestions for information sheets and consent forms on the website.)

- Yes

PC2. Will you tell participants that their participation is voluntary?

- Yes

PC3. Will you obtain explicit consent for participation?

- Yes

PC4. Will you obtain explicit consent for any sensor readings, eye tracking, photos, audio, and/or video recordings?

- Not applicable

PC5. Will you tell participants that they may withdraw from the research at any time and for any reason?

- Yes

PC6. Will you give potential participants time to consider participation?

- Yes

PC7. Will you provide participants with an opportunity to ask questions about the research before consenting to take part (e.g. by providing your contact details)?

- Yes

PC8. Does your project involve concealment or deliberate misleading of participants?

- No

Section 2. Data protection, handling, and storage

The General Data Protection Regulation imposes several obligations for the use of **personal data** (defined as any information relating to an identified or identifiable living person) or including the use of personal data in research.

D1. Are you gathering or using personal data (defined as any information relating to an identified or identifiable living person)?

- Yes

High-risk data

DR1. Will you process personal data that would jeopardize the physical health or safety of individuals in the event of a personal data breach?

- No

DR2. Will you combine, compare, or match personal data obtained from multiple sources, in a way that exceeds the reasonable expectations of the people whose data it is?

- No

DR3. Will you use any personal data of children or vulnerable individuals for marketing, profiling, automated decision-making, or to offer online services to them?

- No

DR4. Will you profile individuals on a large scale?

- No

DR5. Will you systematically monitor individuals in a publicly accessible area on a large scale (or use the data of such monitoring)?

- No

DR6. Will you use special category personal data, criminal offense personal data, or other sensitive personal data on a large scale?

- No

DR7. Will you determine an individual's access to a product, service, opportunity, or benefit based on an automated decision or special category personal data?

- No

DR8. Will you systematically and extensively monitor or profile individuals, with significant effects on them?

- No

DR9. Will you use innovative technology to process sensitive personal data?

- No

Data minimization

DM1. Will you collect only personal data that is strictly necessary for the research?
- Yes

DM4. Will you anonymize the data wherever possible?
- Yes

DM5. Will you pseudonymize the data if you are not able to anonymize it, replacing personal details with an identifier, and keeping the key separate from the data set?
- Yes

Using collaborators or contractors that process personal data securely

DC1. Will any organization external to Utrecht University be involved in processing personal data (e.g. for transcription, data analysis, data storage)?
- Yes

DC2. Will this involve data that is not anonymized?
- Yes

DC3. Are they capable of securely handling data?
- Yes

DC4. Has been drawn up in a structured and generally agreed manner who is responsible for what concerning data in the collaboration?
- Yes

DC5. Is a written contract covering this data processing in place for any organization which is not another university in a joint research project?
- Yes

International personal data transfers

DI1. Will any personal data be transferred to another country (including to research collaborators in a joint project)?
- No

Fair use of personal data to recruit participants

DF1. Is personal data used to recruit participants?
- No

Participants' data rights and privacy information

DP1. Will participants be provided with privacy information? (Recommended is to use as part of the information sheet: For details of our legal basis for using personal data and the rights you have over your data please see the University's privacy information at www.uu.nl/en/organisation/privacy.)
- Yes

DP2. Will participants be aware of what their data is used for?

- Yes

DP3. Can participants request that their personal data be deleted?

- Yes

DP4. Can participants request that their personal data be rectified (in case it is incorrect)?

- Yes

DP5. Can participants request access to their personal data?

- Yes

DP6. Can participants request that personal data processing is restricted?

- Yes

DP7. Will participants be subjected to automated decision-making based on their personal data with an impact on them beyond the research study to which they consented?

- No

DP8. Will participants be aware of how long their data is being kept for, who it is being shared with, and any safeguards that apply in case of international sharing?

- Yes

DP9. If data is provided by a third party, are people whose data is in the data set provided with (1) the privacy information and (2) what categories of data you will use?

- Not applicable

Using data that you have not gathered directly from participants

DE1. Will you use any personal data that you have not gathered directly from participants (such as data from an existing data set, data gathered for you by a third party, data scraped from the internet)?

- No

Secure data storage

DS1. Will any data be stored (temporarily or permanently) anywhere other than on password-protected University authorized computers or servers?

- No

DS4. Excluding (1) any international data transfers mentioned above and (2) any sharing of data with collaborators and contractors, will any personal data be stored, collected, or accessed from outside the EU?

- No

Section 3. Research that may cause harm

Research may cause harm to participants, researchers, the university, or society. This includes when technology has dual-use, and you investigate an innocent use, but your results could be used by others in a harmful way. If you are unsure regarding possible harm to the university or society, please discuss your concerns with the Research Support Office.

H1. Does your project give rise to a realistic risk to the national security of any country?

- No

H2. Does your project give rise to a realistic risk of aiding human rights abuses in any country?

- No

H3. Does your project (and its data) give rise to a realistic risk of damaging the University's reputation? (E.g., bad press coverage, public protest.)

- No

H4. Does your project (and in particular its data) give rise to an increased risk of attack (cyber- or otherwise) against the University? (E.g., from pressure groups.)

- No

H5. Is the data likely to contain material that is indecent, offensive, defamatory, threatening, discriminatory, or extremist?

- No

H6. Does your project give rise to a realistic risk of harm to the researchers?

- No

H7. Is there a realistic risk of any participant experiencing physical or psychological harm or discomfort?

- No

H8. Is there a realistic risk of any participant experiencing a detriment to their interests as a result of participation?

- No

H9. Is there a realistic risk of other types of negative externalities?

- No

Section 4. Conflicts of interest

C1. Is there any potential conflict of interest (e.g. between research funder and researchers or participants and researchers) that may potentially affect the research outcome or the dissemination of research findings?

- No

C2. Is there a direct hierarchical relationship between researchers and participants?

- No

Section 5. Your information.

This last section collects data about you and your project so that we can register that you completed the Ethics and Privacy Quick Scan, sent you (and your supervisor/course coordinator) a summary of what you filled out, and follow up where a fuller ethics review and/or privacy assessment is needed. For details of our legal basis for using personal data and the rights you have over your data please see the University's privacy information. Please see the guidance on the ICS Ethics and Privacy website on what happens on submission.

Z0. Which is your main department?

- Information and Computing Science

Z1. Your full name:

Jens Leonard Laurens Hartkamp

Z2. Your email address:

j.l.l.hartkamp@students.uu.nl

Z3. In what context will you conduct this research?

- As a student for my master thesis, supervised by::
  Johan Jeuring

Z5. Master programme for which you are doing the thesis

- Artificial Intelligence

Z6. Email of the course coordinator or supervisor (so that we can inform them that you filled this out and provide them with a summary):

j.t.jeuring@uu.nl

Z7. Email of the moderator (as provided by the coordinator of your thesis project):

coordinator-ai-master@uu.nl

Z8. Title of the research project/study for which you filled out this Quick Scan:

Exploring the use of large language models to improve one-to-one communication training scenarios

Z9. Summary of what you intend to investigate and how you will investigate this (200 words max):

DialgueTrainer is a company which trains users in communication skills. Dialoguetrainer has a structure in which the user gets an assignment, with a few predetermined goals. For example the assignment to bring bad news with the goal to communicate clearly whilst showing empathy. The user will then have a conversation in which they can choose from a number of options after every prompt. Each of these options will have impact on the parameters (empathy and clarity in this case) where there is often one optimal option.
This research will attempt to improve Dialoguetrainer scenarios using Large Language Models(LLMs). The first improvement this research will try to make is rephrasing infrequently chosen distractor options. LLMs will be used to generate alternative distractor options to replace infrequently used distractor options.
Furthermore Dialoguetrainer has experimented with open ended answers and noticed a problem when the answer given does not match any of the predetermined answers. This research will generate additional options using LLMs in an attempt to reduce the amount of “no match” cases.
Lastly this research will attempt to generate new sentences given the desired parameter settings in an attempt to automate the process of scenario writing.

Z10. In case you encountered warnings in the survey, does supervisor already have ethical approval for a research line that fully covers your project?

- Not applicable

Scoring

- Privacy: 0
- Ethics: 0
