UTRECHT UNIVERSITY

Department of Information and Computing Science

Artificial Intelligence Master Thesis

Enhancing Data-to-Text Systems with Neural-Symbolic Methods: An Exploration of Large Language Models as Text Scorers

Daily supervisor:
Bas Testerink

First supervisor:
Tejaswini Deoskar

Second examiner:
Floris Bex

Author:
Yahan Ke
1522639

October 31, 2024

Abstract

Data-to-text generation converts structured data into natural language text, simplifying complex data interpretation and reducing manual effort. Traditional rule-based and neural approaches each offer distinct strengths and weaknesses—rule-based systems ensure data fidelity but often produce rigid text, while neural models generate more natural text but risk deviating from the source data. To address these limitations, a neural-symbolic data-to-text conversational system was proposed, consisting of an information retrieval system, a generative grammar, and a text scorer. This study explores the use of large language models as text scorers, focusing on their ability to align with human judgments when scoring grammar-generated text. A benchmark dataset was created to study human preferences, and several large language models (LLMs) were tested using the sentence-scoring method to obtain model judgments. Experiments revealed that all LLMs struggled with the "likelihood trap", favoring bland responses over informative ones. Prompts, including prompts containing shuffled data, were effective in mitigating this issue, suggesting that the prompt's role is less about conveying accurate information and more about mitigating word frequency effects on sentence scoring. Furthermore, increasing model scale did not consistently improve performance, suggesting that larger models primarily enhance competencies that are not critical for the text-scoring task. The FLAN-T5 model outperformed other tested models, with the 783M-parameter variant achieving near-human performance. Finally, the integration of a basic generative grammar with the LLM text scorer demonstrated the effectiveness of the neural-symbolic approach. LLMs' extensive linguistic knowledge allows for simplification of the grammar design, while the grammar ensures accurate data representation in the generated text.

Contents

Appendix

1. Introduction

Data-to-text(D2T) generation is the process of generating natural language text to convey the information embodied in structured data. This approach offers several advantages, such as making complex information more accessible to a wider audience, reducing the effort needed for manual data interpretation and reporting, and minimizing human error. In law enforcement, officers often need to analyze sensor data, such as location tracking and surveillance feeds, or compile reports based on criminal records. These tasks are typically repetitive and time-consuming, requiring substantial human resources. A D2T system can simplify this work by generating clear and insightful texts directly from the data, saving time and maintaining accuracy. This need for efficiency and accuracy in handling large volumes of structured data is a key motivator for developing the D2T system within the police force.

The D2T generation can be seen as a process of making a series of decisions that guide the creation of text from data. Neural D2T systems make these decisions by learning linguistic patterns from large datasets, while rule-based systems rely on predefined rules to process them. Neural models are capable of generating natural text because they develop strong linguistic capabilities during training. However, neural models tend to hallucinate, which is the behavior of producing text that deviates from the source data [1], [2]. On the other hand, rule-based systems produce texts highly accurate in their representation of data since humans have direct control over the generation process [3]. While this approach ensures data fidelity, creating natural text requires encoding complex linguistic knowledge into the rules, which is costly. Even then, some unexpected cases may be overlooked, leading to rigid text outputs.

The complementary strengths and weaknesses of neural and rule-based approaches motivate the investigation of a neural-symbolic system for data-

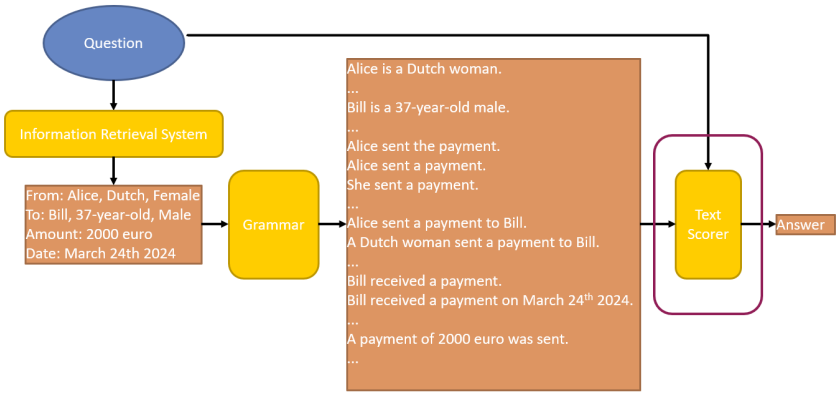

Figure 1.1: The main workflow of the proposed neural-symbolic data-to-text conversational system. This research focuses on the text scorer component, highlighted in the purple frame.

to-text generation. The concept of hybrid natural language generation(NLG) has been around since Knight et al.'s 1995 paper, where they introduced a statistical-symbolic approach that added a statistical component to a rule-based generator [4]. The key idea was to allow the rule-based generator to produce multiple variations of the same idea, while the statistical component would select the most suitable one. Their experiments showed that this combination effectively filled the linguistic gaps of the generator, leading to more natural text. Moreover, by using a statistical filter that could filter out invalid texts, the system could simplify the rule set and combine them more flexibly.

Building on the neural-symbolic NLG concept, and applying it to the police department's D2T needs, the following neural-symbolic data-to-text conversational system is proposed. An illustrative diagram of the system's main workflow can be found in Figure 1.1. The main workflow consists of three steps: first, retrieve a data snippet that contains the information relevant to the user's question; second, apply grammar rules to generate answers describing this data snippet; and third, use a language model as a text scorer to select the candidate answer that best matches human preferences. This design divides responsibilities between components, allowing each to focus on a specific problem. The information retrieval component ensures the correct data is selected, the grammar rules handle the faithfulness of the generated text to the data, and the language model ensures the

linguistic quality of the final output.

This study is an exploratory investigation into the proposed neural-symbolic data-to-text conversational system, with an emphasis on the text scorer component. Specifically, it examines the use of large language models(LLMs) as text scorers. The core objective is to assess whether LLMs can reliably act as proxies for human judgment. The study addresses this by breaking the investigation into three key questions:

How can LLM judgments be obtained? Various methods have been discussed in related work, and this study focuses on the sentence scoring method, which is a widely adopted and straightforward measure of acquiring LLM judgments.

How do humans rate different grammar-generated texts? To explore this, a human preference dataset was constructed, simulating grammar-generated text and gathering human judgments to understand their preferences.

How well do LLM judgments correlate with human preferences? Different LLM architectures and model scales were tested, alongside various methods to improve correlation. Particular attention was given to the influence of model scale, based on the hypothesis that, since factual accuracy is controlled by the grammar component, the small-scale models with basic formal linguistic competence could be sufficient for the task.

This thesis is organized as follows: Chapter 2 reviews related work, starting with an overview of D2T methods. It then examines human text judgments from two angles—acceptability and dialogue context—and includes studies on predicting human preferences, sentence scoring, the "likelihood trap", and the impact of model scale on scoring tasks. Chapter 3 presents the methodology, detailing the approach for capturing LLM preferences, the benchmark dataset for human judgments, and the evaluation metric. Chapter 4 describes experiments that assess model architecture, scale, and strategies for mitigating the likelihood trap through re-ranking models and prompts, concluding with an evaluation of grammar-based generation with LLM scoring. Chapter 5 discusses key interpretations of the experimental

results and proposes directions for future research, and Chapter 6 summarizes the main insights and contributions.

This thesis contributes to the field of AI by exploring innovative methods for enhancing data-to-text systems, an area integral to NLG. The study advances the use of large language models in reflecting human judgment of text, while demonstrating how neural-symbolic AI can be harnessed for tasks requiring factual accuracy. Such advancements underscore AI's potential in transforming data into reliable insights across diverse fields, from law enforcement to finance.

2. Related work

The related work section begins by providing background on the data-to-text (D2T) generation task in Section 2.1, offering an overview of rule-based, neural-based, and neural-symbolic approaches to D2T systems. Following this, Section 2.2 and Section 2.3 explore how humans perceive different types of text. These two sections address the topic from distinct perspectives: acceptability judgments focus on human evaluation of isolated text pieces, while dialogue judgments consider how text is perceived within the context of a conversation. Related studies on predicting these human preferences are also included in the corresponding section. Next, Section 2.4 discusses sentence scoring, a common method for obtaining LLM judgments, and Section 2.5 explores the "likelihood trap", a counter-intuitive phenomenon that can occur when using sentence scoring. Another potential method for obtaining LLM judgments is briefly mentioned in Section 2.6. Finally, the impact of LLM scale on the text-scoring task is discussed in Section 2.7.

2.1 Data-to-Text Generation

Data-to-text generation is the process of generating natural language text to convey the information embodied in structured data. While it was achieved through rule-based models traditionally, neural approaches become common in recent years due to their ability to generate fluent output while reducing the effort of manually designing rules. However, neural models tend to hallucinate, which is the behavior of producing text that deviates from the source data [1], [2]. The tendency to generate plausible yet factual incorrect output has restricted the real-world deployment of data-to-text models due to safety and ethics concerns [5], [6]. For example, in medical applications, a hallucinatory description of patient data could provoke a life-threatening

incident for the patient [5]. Several studies have focused on tackling the fidelity challenge of neural data-to-text generation [1], [7]. However, these methods only mitigate the hallucination and do not guarantee the fidelity of the generated text. Moreover, a recent study from Xu et al. formally defined hallucination and showed that hallucination is inevitable by employing results from learning theory [6].

To acquire more control over the generated output, we revisit the rule-based approach. Rule-based models convert data to natural language by selecting and filling handcrafted templates [8]. Such models are highly robust, allowing humans to have full control over the generated text, which makes them generally accurate in their representation of the data [3]. Despite the high-quality output, the development of the rule set requires extensive time and cost. For complex situations involving multiple domains, the cost of describing data using rules is extremely high, which makes rule-based modeling only applicable to simple scenarios [3]. Rule-based systems also face challenges when they need to be extended to other domains [3].

Rule-based systems, despite generating text with great fidelity, are often criticized for producing rigid responses that lack naturalness and fail to capture the nuances of human expression [9]. On the contrary, neural models, while proving capable of generating more human-like text, inevitably produce text that deviates from the input data. The complementary strengths and weaknesses of neural and rule-based approaches motivate the investigation of a neural-symbolic approach, where the two problems of factual accuracy and naturalness are left to rule-based and neural methods, respectively.

The concept of a hybrid NLG solution dates back to Knight et al.'s 1995 paper [4]. They argued that a rule-based generator requires extensive lexical, grammatical, and conceptual knowledge to produce fluent sentences. However, constructing such a comprehensive knowledge base is practically impossible. To address this, they proposed integrating a statistical language model to bridge this knowledge gap. Their experiments demonstrated that adding the statistical component effectively filled in the linguistic knowl-

edge gap of the generator. Furthermore, they showed that the rule design of the generator could be simplified by delegating lexical choice to the statistical component. This allows the generator to use simpler grammar rules and combine them more freely, though it may produce more invalid texts. The statistical component can then filter out problematic texts from the generator.

The key idea behind the two-level generation in Knight et al.'s study can be summarized as using an over-generating grammar to express the same idea in multiple ways and selecting the optimal one using a statistical component. Despite this innovative approach, Knight et al.'s study is limited in using a bi-gram model as the statistical component. The bi-gram model cannot capture complex linguistic patterns due to its short context. Consequently, only some aspects of lexical choices were delegated to the statistical component. Given the extensive linguistic knowledge acquired by LLMs, the possibility of delegating more text-generation choices to the language model needs to be explored.

2.2 Acceptability Judgment

Acceptability can be regarded as a perception that arises when a speaker encounters a linguistic stimulus [10]. It is a cover term that can be used interchangeably with well-formedness, nativeness, and naturalness in linguistic literature [11].

Chomsky acknowledged that human perceptions of acceptability are not always clear-cut. Some texts are clearly acceptable and others are not [12]. But there is also some middle ground, where texts are perceived as neither entirely acceptable nor unacceptable. Experiments by Lau et al. compared human acceptability rating patterns with binary and gradient judgment patterns [13]. They observed that the distribution of acceptability judgments was similar to the baseline distribution of gradient judgments. Thus supporting the gradient characteristic of acceptability.

Human acceptability judgment is influenced by various factors. Chom-

sky notes that grammaticality can be seen as one of the many factors that influence acceptability [12]. The following studies have also proven that acceptability is sensitive to grammaticality. In Sprouse et al.'s experiments, they observed that 90% of the cases claimed by linguists to be grammatically different also showed differences in acceptability [14]. Other factors widely recognized as having an impact on acceptability include the occurrence of wh-dependencies, the length of dependencies, the relative frequency of lexical items, and the relative frequency of grammatical structures [10].

Multiple studies have focused on predicting human acceptability judgments. The most widely adopted method is to normalize the probability assigned to the target text by a pre-trained language model [13], [15]–[21]. Lau et al.'s empirical experiment shows that there is no correlation between sentence length and human acceptability judgments [13]. Based on their argument that acceptability is independent of sentence length and word frequency, they proposed several normalization methods to mitigate the effect of these factors on sentence probabilities [15], [16]. Experiments demonstrate that normalized sentence probabilities better correlate with human acceptability judgments. In the subsequent studies, Lau et al. have proven the robustness of their approach using different pre-trained language models, different languages, and out-of-domain texts [16], [17]. Following the success of predicting the acceptability of single sentences, Lau et al. studied the acceptability prediction of sentences in a given context [18], [20]. They first examined the impact of context supplements on human acceptability ratings. They observed a compression effect when sentences were rated within contexts. Context increased the acceptability ratings of ill-formed sentences while decreasing the acceptability ratings of well-formed sentences. For acceptability prediction, they continued the previously proposed unsupervised approach (pre-trained language model plus normalization) and once again proved the effectiveness of the method. The performance of bidirectional models was comparable to the estimated human upper bound, which is the estimated correlation between a human participant's ratings and the mean ratings. In Ek et al.'s paper, they investigate the impact of augmenting language models with syntactic and semantic information on acceptability

predictions [19]. Experiments show that enhancing language models with syntactic information can help reduce perplexity[1]. However, both syntactic and semantic information fail to boost the models' performance in acceptability prediction. These results also indicate that lower perplexity does not lead to more accurate prediction in acceptability judgments.

There are attempts to predict acceptability or naturalness in a supervised fashion. Inspired by second language assessment, Tian et al. extract a set of linguistic features to predict human judgments of sentence naturalness [22]. The features include lexical features, such as token-type ratio, parsing features, such as parse tree height, and language model features, such as sentence perplexity. A binary classification model is then trained on the manually labeled dataset to classify sentences as "natural" or "unnatural" based on these linguistic features. Although experiments by Tian et al. show that this automatic evaluation method is highly correlated with human judgments, their model performs poorly in evaluating sentences outside the training domain. To improve model's performance in other domains, supervised training on a labeled set of this particular domain is required. In Warstadt et al.'s study, they trained a classification model to predict binary acceptability. The model contains a sentence encoder and a classification head. The model outperforms Lau et al.'s unsupervised models but falls short compared to human judgments [23]. Despite the encouraging performance of their supervised model, Warstadt et al. move on to obtain acceptability prediction in an unsupervised manner as in Lau et al.'s study [24].

Based on these studies, it can be concluded that unsupervised methods using pre-trained language models tend to be more robust and generalize better than supervised methods, which often suffer from poor domain portability. This observation led us to focus on LLMs in an unsupervised manner, rather than adding an extra classification head or scoring layer. Furthermore, Lau et al.'s finding—specifically, that normalizing

[1]Perplexity is a measure of how well a language model predicts a sample. It is the exponentiation of the average negative log likelihood of a sequence. In simpler terms, lower perplexity indicates that the model is more confident in its predictions, meaning it assigns higher probabilities to the correct words in a sequence.

sentence length and word frequency improves the correlation between sentence probabilities and human acceptability judgments—prompted us to conduct preliminary experiments comparing different normalization techniques. This finding also plays a key role in our subsequent analysis and discussion.

2.3 Dialogue Judgment

The text scorer's task is to select the candidate answer that aligns most closely with human preferences. This requires considering not only the generated text itself but also the context, which is the user's question. The question and the candidate answer together form a dialogue. In the previous subsection, most studies overlooked the role of context, and those that did include it only considered the preceding text in a monologue. Therefore, this subsection reviews studies that focus on human judgment of dialogues.

Paul Grice has proposed the cooperative principle to describe and explain how humans behave in conversations.

Make your contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged.

The cooperative principle can be concluded in simple words: say what you need to say, when you need to say it, and how you need to say it. Grice then lists four maxims to specify the behavior that follows the cooperative principle: quantity, quality, relation, and manner. These maxims can be seen as requirements that people seek to follow themselves and expect the other party in the conversation to follow when forming a response.

The maxim of quantity, where one tries to be as informative as one possibly can, and gives as much information as is needed, and no more.

The maxim of quality, where one tries to be truthful, and does not give information that is false or that is not supported by evidence.

The maxim of relation, where one tries to be relevant, and says things that are pertinent to the discussion.

The maxim of manner, when one tries to be as clear, as brief, and as orderly as one can in what one says, and where one avoids obscurity and ambiguity.

Several studies used Grice's maxims as the evaluation criterion of dialogues. Jacquet et al. investigate the impact of the Gricean Maxims of Quality, Quantity, and Manner on chatbots' humanness [25]. They observe that the violations of the maxim of quality have a significant negative effect on the humanness of the chatbot. Nam et al.'s study evaluates a chatbot's communicative performance based on the maxims violated by the generated utterances [26]. The results show that the chatbot violates the maxim of relation most frequently.

Several studies investigate the effectiveness of using pre-trained language models in dialogue evaluation tasks. Mehri et al.'s work proposed a zero-shot dialogue evaluation framework that can assess fine-grained qualities without any ground-truth reference [27]. Their main idea is that the quality of an utterance can be judged by how likely a pre-trained language model is to generate a follow-up sentence that recognizes a specific quality in it. For example, if a response is interesting, the model is more likely to generate the follow-up utterance "This is an interesting response." Wallbridge et al. conducted an experiment where participants rated the acceptability of dialogue turns [21]. They found that the pre-trained language model's probability of the dialogue turn based on the previous context has a weak but significant correlation with human judgments.

Like the studies discussed in the previous subsection, these studies further support the unsupervised use of LLMs in the text-scoring task. Additionally, they encourage us to focus on the probability of the candidate answer being conditioned on the question.

2.4 Sentence Scoring

Essentially, sentence scoring is using a language model to estimate the likelihood of a sentence [28]. The idea behind this is that a better sentence should be more acceptable to the language model, resulting in a higher likelihood

score [24]. Many studies on hypothesis re-ranking and automatic NLG evaluation are variants developed on the idea of computing text likelihood via language models. Based on summarizing previous research, the common variants of sentence scoring are classified into three categories: different language model architectures, different modifications upon sentence likelihoods, and different usages of prompts.

Different language model architectures Methods for calculating sentence likelihood through language models have evolved along with language modeling. In Lau et al.'s study, they built a variety of language models using unsupervised training, including N-grams, Bayesian Hidden Markov Models, and Recurrent Neural Networks for sentence likelihood [16]. Among them, the RNN model's sentence scoring had the highest correlation with human judgment. After the rise of transformer architecture, recent sentence scoring studies mainly use transformer-based language models. Pretrained language models are also widely adopted, as they score sentences out of the box.

Warstadt et al. obtained sentence likelihood from decoder-only models, such as GPT2 and Transformer XL [24]. For decoder-only models, the sentence likelihood can be estimated using the chain rule. Given a sentence $W = \{w_1, ..., w_{|W|}\}$, where w_i is the i-th token of W, and $|W|$ is the sentence length, the log-likelihood of W can be calculated as:

$$\log P(W) = \sum_{t=1}^{|W|} \log P(w_t | W_{<t}) \tag{2.1}$$

Here, $W_{<t}$ represents the list of tokens previous to token w_t. Decoder-only models can calculate sentence log-likelihood within single inference. However, it can only extract unidirectional information without considering bidirectional context [28].

Several studies have investigated the ability of encoder-only models to estimate sentence likelihood and compared them with decoder-only models [29], [30]. Unlike decoder-only models, which can only capture unidi-

rectional contexts, the encoder-only model can utilize bidirectional contexts through its masked language modeling(MLM) objective. In MLM, a token w_t is substituted with [$MASK$] and predicted using the past and future tokens $W_t = \{w_1, ..., w_{t-1}, w_{t+1}, ..., w_{|W|}\}$. To score the sentence W, one needs to first mask each token w_t and calculate the probability of the token being masked. Then, the sentence likelihood can be estimated by summing the probability of each token. The formula is:

$$\log P(W) = \sum_{t=1}^{|W|} \log P(w_t|W_t) \tag{2.2}$$

Salazar et al.'s experiments show that encoder-only models such as BERT and RoBERTa outperform decoder-only models of similar size such as GPT2 on the Benchmark of Linguistic Minimal Pairs (BLiMP) [29]. BLiMP dataset evaluates language models' abilities to assign a higher score to the more acceptable sentence in a sentence pair [24]. Salazar et al. suggest that encoder-only models perform better because they are able to capture bidirectional information. [29]. For example, for the sentence "*The pamphlets about Winston Churchill have resembled those photographs.*", encoder-only models can use both "*The pamphlets*" and "*resembled those photographs*" as clues for determining whether to use the verb "*have*" or "*has*". Decoder-only models, on the other hand, can only make judgments based on "*The pamphlets about Winston Churchill*", which contains misleading signal "*Winston Churchill*" that would cause the models to lean towards using the verb "*has*". However, encoder-only models requires multiple inferences to estimate sentence likelihood, which leads to expensive computational and time cost [28]. Some works have sought to mitigate this problem using techniques such as distillation and stochastic estimation, but this comes at the expense of performance [29], [31].

Encoder-decoder models are trained in a sequence-to-sequence(seq2seq) fashion. The encoder processes the source sequence while the decoder autoregressively generates the target sequence. This makes encoder-decoder

models well-suited to compute the generation probability of a sequence conditioned on another sequence [32]. Given a source sequence $X = \{x_1, ..., x_{|X|}\}$, the probability of generating sentence W can be calculated as:

$$\log P(W|X) = \sum_{t=1}^{|W|} \log P(W_t|W_{<t}, X) \tag{2.3}$$

Inspired by this, Yuan et al. proposed to use BART, an encoder-decoder based pre-trained model, to evaluate the quality of the generated texts [32]. A benefit of using encoder-decoder model is that the scoring can be augmented by adding prompts. However, the use of prompts does not always improve performance. Yuan et al.'s experiments have shown that some prompts lead to lower performance. Therefore, the use of prompts and the content of prompts need to be investigated before applying them to scoring tasks.

To combine the advantages of encoder-only models and decoder-only models while mitigating their drawbacks, Song et al. proposed a novel sentence-scoring model called Transcormer [28]. Transcormer exploits a sliding language modeling approach and employs a triple-stream self-attention mechanism. These innovations allow the model to estimate sentence scores in bidirectional contexts with only one forward pass. Compared to the decoder-only model, Transcormer improves sentence scoring performance without significantly increasing computational cost. Compared to the encoder-only model, Transcormer improves inference efficiency with comparable performance.

Different modifications upon sentence likelihood In many cases, sentence likelihoods can be used directly as sentence scores. However, applying modifications upon sentence likelihoods can yield sentence scores that align with the scoring objective better. In automatic speech recognition, interpolation is often used to combine the sentence scores of the acoustic model with the scores of the language model [30]. In neural machine translation, it is common to introduce length normalization to the scoring func-

tion. Otherwise, the pure sentence likelihood will tend to prefer shorter sentences over longer ones. This is because each token in the sentence adds a negative log-likelihood, yielding a lower score for sentences with more tokens [33]. Lau et al. proposed normalization methods to mitigate the influence of word frequency and sentence length on sentence likelihood, as these factors do not affect human judgments of sentence acceptability. Their experiments demonstrated that normalized sentence likelihood has a stronger correlation with human acceptability judgments [16]. In the following study, Xie et al. proposed to calculate the likelihood difference between an original text and its perturbed version. They argue that perturbation can function as a normalization factor to modulate the effects of word frequency and text length when estimating sequence likelihood [34].

Different usages of prompts Prompting involves adding brief phrases to inputs or outputs and guiding pre-trained models to perform specific tasks. In sentence scoring, prompts can be used to define the background and criteria for scoring. For instance, adding the prompt "Is the text natural?" before the text can direct the model to focus on naturalness scoring. Yuan et al.'s experiments with BART demonstrated that prompts enhance the alignment between model judgments and human assessments [32]. Advances in LLMs now support extensive context windows, enabling stating detailed evaluation criteria in prompts. In their study, Fu et al. introduced a tailored evaluation protocol for varying evaluation tasks [35]. While their approach extends the lengths of prompts, the essence remains sentence scoring, where sentence scores are conditional possibilities conditioned on extensively long prompts.

2.5 Likelihood Trap

A common premise of all sentence scoring studies is that the higher the likelihood that a language model assigns to a piece of text, the higher the quality of that piece of text. However, it is debatable if this intuitive premise holds. Researchers have long reported the existence of the likelihood trap: the counter-intuitive empirical observation that high-likelihood text tends

to be bland, incoherent, and repetitive. Li et al. observed that neural conversation models often assign a higher likelihood to generic responses like "I don't know" than more informative alternatives [36]. They ascribed this behavior to the more frequent occurrences of generic responses in the training set. More informative responses sometimes contain specific information that never appeared in the training set, thus receiving a lower likelihood than bland but safe responses that appear more frequently in the training data. Holtzman et al. demonstrated that human-written text does not maximize the text likelihood [37]. They showed that LLM can generate texts with a much higher likelihood than human-written texts, but these texts are less diverse and more repetitive. Therefore, they argued that high-quality text does not necessarily have a high likelihood, but rather has a likelihood that is close to the likelihood of human-written text. Zhang et al. quantify the relationship between text likelihood and text quality [38]. They sampled a list of context-continuation texts with different model likelihoods and collected human ratings of these texts. They illustrate that the text with the highest quality is not the most likely. Text quality is positively related to text likelihood until an inflection point where it then becomes negatively related.

These findings make researchers aware that though using likelihood maximization as a training objective leads to highly capable language models, using it as a generation objective leads to degeneration. Different sampling and re-ranking methods were proposed to avoid the generation of likely yet generic texts. However, current automatic text evaluation frameworks (e.g., BARTScore and GPTScore) do not take into account the presence of likelihood traps. While they attempt to structure the text evaluation task as a text generation task, they fail to include other essential text generation processes [32], [35]. These frameworks judge text quality solely based on text likelihood, which might result in bland, incoherent texts being perceived as being of higher quality than informative texts. Therefore, in this study, the re-ranking process was tested as a way to improve the quality of the language model's judgments.

In the series of studies on DialoGPT, researchers have proposed two dif-

ferent re-rank models. The Maximum Mutual Information(MMI) scoring function is a pre-trained backward model that predicts the probability of the source input from the given response [36], [39]. Intuitively, maximizing the backward model likelihood penalizes bland hypotheses. Since a bland hypothesis can work as a response to many possible queries, the probability of each specific query conditioned on this bland hypothesis will be low. The human evaluation demonstrates that the inclusion of MMI can significantly improve relevance, informativeness, and human likeness. Dialogue Ranking Pre-trained Transformers(DialoRPT) is a re-ranking model trained on 133M pairs of human feedback data using a contrastive learning approach [40]. Rather than scoring each dialogue individually, the training objective is to maximize the scores of the positive samples while minimizing the scores of the negative samples. Experiments show that DialoRPT has a higher human preference correlation than DialoGPT and MMI.

2.6 Comparative Assessment

A recent study from Liusie et al. compared two options of using LLM to evaluate NLG output: sentence scoring and comparative assessment. Comparative assessment uses relative comparisons between pairs of candidates [41]. Liusie et al. were motivated by the insight that humans often find it more intuitive to compare two options rather than scoring each one independently. Experiments show that for moderately sized LLMs, comparative assessment outperforms absolute scoring. Comparative assessment can also achieve results comparable with state-of-the-art methods. However, comparing the full set is a $O(N^2)$ task, which is computationally prohibitive for large N. To address this issue, Liusie et al. proposed an efficient LLM pairwise assessment framework [42]. Using this efficient approach, the score prediction generated based on a small set of comparisons can achieve similar performance to the full set of comparisons. Another issue with comparative assessment is the positional bias. Liusie et al.'s experiments have shown that most LLMs, especially the larger ones, favor text at a certain position.

2.7 The Scale of LLM

Current LLMs, trained on trillions of tokens and billions of parameters, have not only demonstrated near-human or even beyond-human performance in traditional NLP tasks but have also shown the ability to perform complex tasks such as code generation and mathematical reasoning. However, in the proposed system, the LLM is assigned to only one task - acting as a proxy for human judgment. Deploying a large-scale model for this simple task may lead to under-utilization of the model's capabilities, which prolongs the reasoning time, creates an excessive carbon footprint, and affects the system's accessibility to resource-constrained devices. However, scaling down the model may also lead to a decrease in model performance, as previous research has shown that scaling up improves model performance for a vast majority of tasks [43], [44]. This motivates the investigation into the appropriate size of language models needed for the system.

Mahowald et al. note that the coupling of language and thinking in everyday life leads to confusion between language and thinking in LLM assessment [45]. A common fallacy associated with the language-thinking relationship is that a model that is good at language must also be good at thinking. When a language model generates coherent logical text, it is often assumed that it also possesses relevant knowledge and reasoning skills. Another fallacy is that if a model is not good at thinking, e.g., if it is unable to demonstrate an understanding of world knowledge, then it must also be a poor model of language. To mitigate the conflation of language and thinking, they suggest that when evaluating LLMs, linguistic competence should be separated into formal linguistic competence (understanding of the syntax and semantics of the language) and functional linguistic competence (the ability to use the language in the real world). After evaluating the two competencies separately, Mahowald et al. outline that LLMs have largely acquired formal language competence, but leave many gaps in functional language competence. Unlike formal language competence, which can be acquired with simple next-word-prediction training, to acquire functional language competence, LLMs often require augmentation of other models and

special fine-tuning. LLMs without these additions tend to lack robustness and generality in functional language competence. Mahowald et al. also emphasizes that obtaining functional language competence enhancements often requires several orders of magnitude more training costs than obtaining formal language competence enhancements. Since today's LLMs often possess certain functional language capabilities, we speculate that much of the scale of these models may not contribute to formal language competence gains, but rather be used for functional language competence gains. Given that in the proposed system the factual accuracy of the generated text is taken care of by the grammar, we argue that the language models we use only need to have basic formal language competence.

The scale of a model is composed of three key factors: the parameter size, the training dataset size, and the amount of computation used for training [43]. Warstadt et al.'s study highlights that training set size, compared to parameter size and model architecture, has the greatest impact on a model's grammatical competence [24]. Experiments showed that LSTM and transformer-based models trained on the same training set performed similarly on grammatical tasks. The performance of GPT-2 models with different parameter sizes is also not significantly different. In contrast, the change in the quantity of training data incurs a significant change in the model's grammatical competence. Warstadt et al. also speculate that there is a linear relationship between the model's grammar competence gain and the logarithm of the training set size. In a subsequent study, Zhang et al. tracked changes in language models' different capabilities as the training set's size increased [46]. They found that language models require only 10M to 100M words to learn common syntactic and semantic features. Most of the progress in syntactic learning occurs before 10M words of training, while slight growth in semantic learning can still be observed after 100M words of training. Overall, there is little difference in the linguistic knowledge the 100M and 30B models possessed. Although Zhang et al.'s experiments show that training on billions of words can significantly improve a model's factual knowledge and thus dramatically improve the performance of downstream NLU tasks, a model's factual knowledge (which can be cate-

gorized as functional linguistic competence in Mohawald's criterion) is not the most important concern for us. In another study, Huebner et al. trained a RoBERTa model with a 5M word corpus that mimicked the linguistic input received by children during language development. This RoBERTa model had 15 times fewer parameters and 6000 times less training data than the standard RoBERTa model but shows comparable grammatical knowledge to that of the standard RoBERTa model [47].

These studies inspired us to focus more on small-scale language models that were trained to acquire formal language competence rather than large-scale language models that provide functional language capabilities.

3. Methodology

The Ethics and Privacy Quick Scan of the Utrecht University Research Institute of Information and Computing Sciences was conducted (see B). It classified this research as low-risk with no fuller ethics review or privacy assessment required.

This study investigates the use of LLMs as the text-scoring component within a neuro-symbolic data-to-text conversational system. The role of the text scorer is to guide the system in selecting the answer that most closely aligns with human preferences for a given question. To accomplish this, the text scorer's judgments should mirror those of humans, meaning that when people prefer one candidate's answer over another, the text scorer should also prioritize that answer. This section first outlines the method used to capture LLMs' preferences, followed by a description of the dataset used as a benchmark for human preferences. Finally, the metric used to evaluate the model's performance in the text-scoring task is discussed.

3.1 Sentence Scoring

A straightforward way to capture LLMs' preferences is to use its language modeling capabilities to compute answer probabilities conditioned on a question. The idea behind this is that a higher conditional probability for an answer indicates that the model finds the answer more acceptable. Current LLMs' architectures can be categorized into three main types, i.e., decoder-only, encoder-only, and encoder-decoder. For different model architectures, the methods of calculating conditional probabilities are different.

Let $Q = \{q_1, q_2, ..., q_{|Q|}\}$ represent the sequence of tokens in the question, $A = \{a_1, ..., a_{|A|}\}$ represent the sequence of tokens in the answer. Define the concatenation of Q and A as $W = \{w_1, ..., w_{|Q|}, w_{|Q|+1}, ..., w_{|Q|+|A|}\}$. $\cdot_i$ denotes the i-th token of the respective sequence, and $|\cdot|$ denotes the length

of the respective sequence.

In decoder-only models, the probability of a token is computed based on its left context. Given the word sequence W, the conditional probability of the answer A given the question Q can be calculated using the chain rule of probability as follows:

$$P(A|Q) = \prod_{i=|Q|+1}^{|Q|+|A|} P(w_i|W_{<i}) \tag{3.1}$$

Here, $W_{<i}$ represents the tokens that precede w_i.

For encoder-only models, the probability of a token is derived based on both its left and right context. To obtain the probability of a specific token, the token needs to be substituted with a special [$MASK$] token. Therefore, to compute A's conditional probability, each answer token in W needs to be masked once. The formula is:

$$P(A|Q) = \prod_{i=|Q|+1}^{|Q|+|A|} P(w_i|W_{<i}, W_{>i}) \tag{3.2}$$

Here, $W_{>i}$ represents the tokens that follows w_i. It is worth noting that this approach provides an estimate of the conditional probability rather than an exact value.

Encoder-decoder models function in a sequence-to-sequence manner, treating the question Q as the source sequence and the answer A as the target sequence. For a token in the target sequence, its probability is calculated based on the entire source sequence and tokens that precede it in the target sequence. A's conditional probability is calculated as:

$$P(A|Q) = \prod_{i=1}^{|A|} P(a_i|A_{<i}, Q) \tag{3.3}$$

The above approach can also be extended to support the use of prompt messages. A prompt message providing additional context or guidance can be treated similarly to the question, acting as a conditioning factor for the answer. The conditional probability of the answer given the prompt can be calculated in the same manner as $P(A|Q)$.

The conditional probability of an answer is influenced by its length. Specifically, longer answers tend to have lower probabilities than shorter ones, regardless of their quality, due to the multiplicative nature of the chain rule. To mitigate this bias, a length normalization technique should be applied. Several normalization methods were compared in a preliminary experiment(see A). The best-performing method was found to be the log probability averaged by the sentence length. Therefore, in subsequent experiments, the following normalized conditional probability is used as sentence score:

$$Score(A|Q) = \frac{\log P(A|Q)}{|A|} \tag{3.4}$$

3.2 Dataset

3.2.1 Dataset Design

The primary objective of dataset construction is to simulate possible outputs of the generative grammar and gather human preferences for these outputs. To accurately mimic the texts that such a grammar might produce, several critical decision branches likely encountered during grammar derivation were carefully considered. A general derivation process is illustrated in Figure 3.1, where five main text generation choices were identified based on common decision branches in the derivation tree.

A transaction scenario representing the data a user might query about was created. Five wh-questions related to this transaction were formulated, reflecting the types of queries users might ask. For each question, 20 candidate answers—representing texts derived from the grammar—were gener-

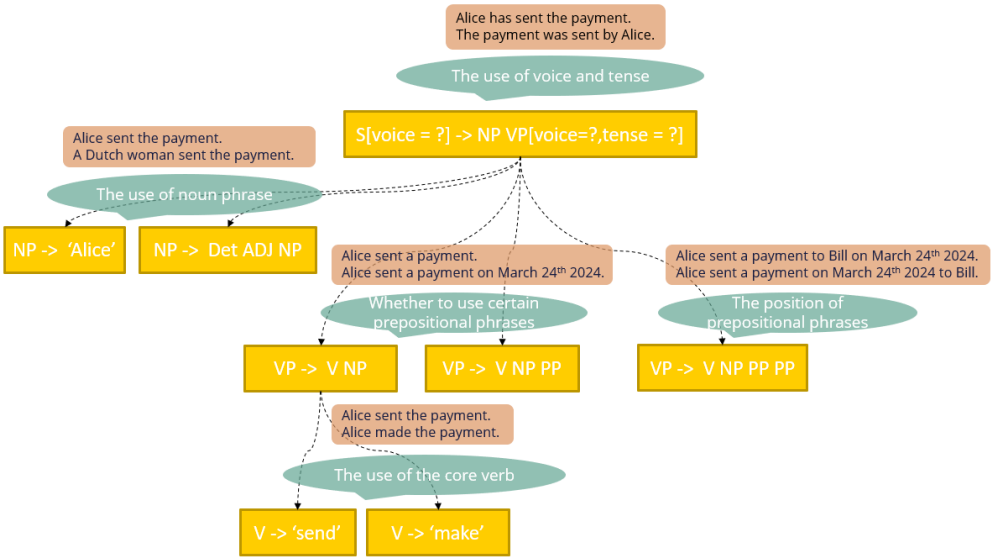

Figure 3.1: An indicative grammar derivation process. Text generation choices are represented in the green bubbles, placed near the decision branches that lead to those choices. Example sentences for each type of choice are shown in the orange boxes.

ated. These candidate answers fall into five categories, each corresponding to one of the identified text generation choices. Within each category, the answers vary based on the specific decisions associated with that text generation choice.

```
From: Alice, Dutch, Female
To: Bill, 37-year-old, Male
Amount: 2000 euro
Date: March 24th 2024

Question:
1. When did Alice send the payment to Bill?
2. Who sent Bill 2000 euro on March 24th 2024?
3. How much money did Alice send to Bill on March 24th 2024?
4. Who received a payment from Alice on March 24th 2024?
5. What is the connection between Alice and Bill?
```

3.2.2 Data Collection

Five surveys were composed and distributed to collect human preference data. Each survey contains a question, and all candidate answers to that question. Participants were asked to rate each candidate answer according to their preferences on a 5-point Likert scale with the following options: unacceptable, poor, passable, good, and perfect. To prompt the participants to capture their nuanced preferences, a context was weaved at the beginning of the survey. The context guided participants to imagine that their organization developed a series of chatbots, which are designed to enable human interactions with structured data through natural language. The candidate answers are responses they got from different chatbots when inquiring about a transaction. The transaction data available to the chatbots are also transparent to the participants. To mitigate bias caused by the order of answers, answers were randomly shuffled for each participant. Additionally, participants were asked to describe their English proficiency level. A check question containing false transaction information was also set in each survey to test whether participants had sufficient English proficiency and took the survey seriously.

When designing the survey, different response formats were considered, specifically 'ranking' versus 'absolute rating'. The subsequent experiments require human rankings of candidate answers as a benchmark. Although a 'ranking' format would provide straightforward data for these experiments, asking participants to rank 20 candidate answers at once could be overwhelming and confusing. Breaking the ranking task into smaller subgroups was also ineffective, as it would not yield a comprehensive ranking of the entire set of answers. Alternatively, the 'absolute rating' format has been successfully employed in several previous studies investigating human preferences in natural language tasks [13], [16], demonstrating its suitability for survey use. While this format does not directly produce ranking data, rankings can be derived by sorting the average participant ratings for each candidate answer.

It is important to note that specific criteria for each rating option were

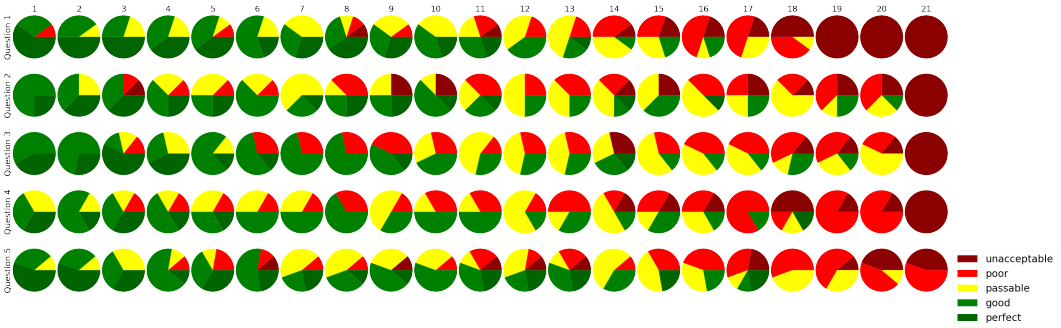

Figure 3.2: Pie charts depicting the distribution of participants' ratings for each Q&A pair. Each row represents the rating distributions for all candidate answers to a given question, with answers ordered from left to right in descending order of average ratings. The questions are listed in Section 3.2.1.

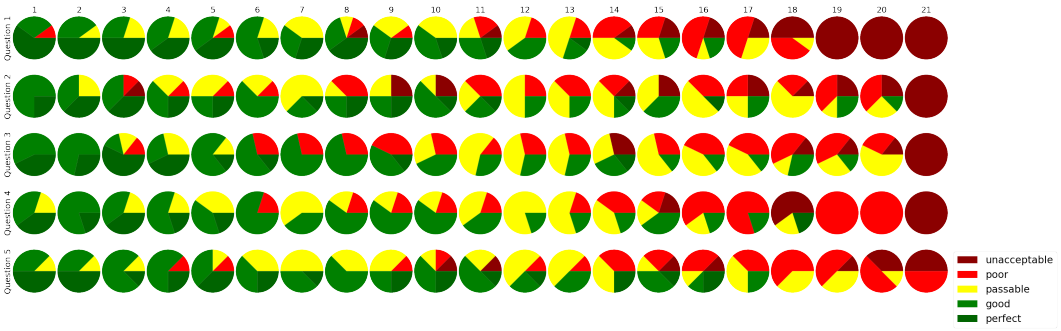

Figure 3.3: Pie charts depicting the distribution of participants' ratings for each Q&A pair, after excluding problematic participants. Each row represents the rating distributions for all candidate answers to a given question, with answers ordered from left to right in descending order of average ratings. The questions are listed in Section 3.2.1.

not provided in the survey. Providing such criteria would have transformed the survey into an annotation task, where participants follow a predefined evaluation protocol to label each response. This goes against the purpose of collecting human preferences.

3.2.3 Dataset Overview

A total of 73 responses were collected. An initial data cleaning was conducted. Participants who did not complete the survey, those who reported limited English proficiency, and those who failed the check question were excluded. 40 responses remained for analysis.

As shown in Figure 3.2, the ratings among participants did not exhibit strong agreement. There are two possible explanations for this lack of agree-

ment: First, human preferences for the candidate answers may generally be uniform, but disagreement arose due to the presence of problematic participants. Second, human preferences for certain candidate answers may inherently vary, making consensus difficult to achieve. An outlier detection was done to further recognize problematic participants. For each candidate answer, the first and third quartiles (Q1 and Q3) and the interquartile range (IQR) of the ratings were calculated. Any rating that fell below (Q1 - 1.5*IQR) or above (Q3 + 1.5*IQR) was considered an outlier. Each participant was then evaluated based on the number of outlying ratings they provided. Two participants who had a significantly higher number of outlying ratings compared to others were removed from the analysis.

As shown in Figure 3.3, the agreement among participants after removing problematic participants does not significantly improve. Moreover, participants did not fail to agree on the ratings of all candidate answers. For every question, some responses were collectively rated positively or negatively by participants. This suggests that the observed disagreements are not due to problematic participants or differing interpretations of the rating scale among participants. Unlike annotation tasks, where high inter-annotator agreement is necessary to ensure dataset quality, this investigation into preferences recognizes that low agreement on certain responses is a normal phenomenon. It reflects the natural variation in individual preferences.

To establish a benchmark for human preferences, mean aggregation was used instead of majority voting. Mean aggregation provides a better representation of overall human preference, especially in this case where participant agreement is low. For instance, consider two answers: one rated as perfect by all five participants, and another rated as perfect by three participants but bad by two. Majority voting would treat both answers as equally good, while mean aggregation would rank the universally approved answer higher, reflecting a more nuanced understanding of collective preference. Each candidate answer received at least five ratings from different participants. The candidate answers for each question were then ranked based on average participant ratings, which reflect the overall level of hu-

Question	Pearson's ρ	Spearman's ρ	Kendall's τ
When did Alice send the payment to Bill?	0.84	0.81	0.69
Who sent Bill 2000 euro on March 24th 2024?	0.64	0.58	0.48
How much money did Alice send to Bill on March 24th 2024?	0.70	0.67	0.56
Who received a payment from Alice on March 24th 2024?	0.76	0.73	0.63
What is the connection between Alice and Bill?	0.68	0.65	0.53

Table 3.1: Average correlation coefficients between participants' individual preferences and benchmark preferences for each question.

Text generation choices	Standard deviation	Range
the use of voice and tense	0.87	1.82
the use of core verb	0.81	1.71
the use of noun phrases	0.95	2.07
whether to use certain prepositional phrases	1.15	2.33
the position of the used prepositional phrases	0.92	2.00

Table 3.2: Average standard deviation and range of human ratings for each text generation choice.

man preference for the different responses to the same question.

As shown in Table 3.1, the benchmark preference exhibits strong correlations with participants' individual preferences.

3.2.4 Data Analysis

The 20 candidate answers for each question can be divided into five sets, each containing four answers that vary according to a specific text generation choice outlined in Figure 3.1. To analyze the impact of each text generation choice on human preferences, the standard deviation of each participant's ratings within each set of answers was calculated. The average standard deviation for each text generation choice was then computed across all questions and participants, along with the average difference between the maximum and minimum ratings. Results are presented in Table 3.2. A higher standard deviation indicates that the ratings are more dispersed, suggesting that the text generation choice introduces greater variability in human preference. A larger range implies a more substantial impact, with some sentences being rated highly while others poorly. As shown in Table 3.2, the choice of whether to use certain prepositional phrases has the strongest impact on participants' preferences, while the choice of core verbs has the least impact.

Overall, text generation choices that have a large impact on human preferences, such as the use of noun phrases, and the use of certain prepositional phrases or not, affect how much transaction information is conveyed in the text. Text generation choices that have less impact on human preferences, such as the use of core verbs and the use of voice and tense, do not affect the amount of transaction information contained in the text but rather the way in which a set of information is conveyed.

3.3 Metric

The performance of a model in the text scoring task is determined by how closely its preferences of candidate answers align with human preferences. To evaluate this alignment, Kendall's τ, a rank correlation coefficient, is employed. Kendall's τ measures the similarity between two rankings. It is defined as:

$$\tau = \frac{Concordant(pair) - Discordant(pair)}{Total(pair)} \tag{3.5}$$

This coefficient is calculated by counting the number of concordant and discordant pairs in the two rankings and then taking the difference, which is divided by the total number of possible pairs. The value of Kendall's τ ranges from -1 to 1, where 1 represents perfect agreement, -1 indicates perfect disagreement, and 0 suggests no association. For each question, Kendall's τ is computed between the model's ranking of all candidate answers and the benchmark rank based on human preferences. A higher Kendall's τ value indicates a stronger correlation between the model's preferences and human preferences, signifying better model performance.

4. Experiments and Results

This chapter presents a series of experiments evaluating key factors influencing model performance in the text-scoring task. Section 4.1 examines the impact of model architecture (decoder-only, encoder-only, and encoder-decoder), followed by an analysis of model scale in Section 4.2. In Section 4.3, we investigate the "likelihood trap", identifying core issues that reduce alignment between model and human judgments, with Sections 4.4 and 4.5 detailing experiments on mitigating this issue using re-ranking models and prompts. Finally, in Section 4.6, we assess the effectiveness of integrating grammar-based generation with LLM text scoring, validating the neural-symbolic approach's role in improving data-to-text outputs.

4.1 Model Architecture

This experiment was designed to investigate the influence of model architecture on model performance in the text-scoring task. GPT2, RoBERTa, and T5 were selected as baselines, representing decoder-only, encoder-only, and encoder-decoder architectures, respectively. An overview of the models can be found in Table 4.1.

Model	Architecture	Parameter size	Training data
GPT2 [48]	decoder-only	355M	Pre-trained on 40GB of text data
RoBERTa [49]	encoder-only	355M	Pre-trained on 160GB of text data
T5 [50]	encoder-decoder	220M	Pre-trained on 750GB of text data
OPT [51]	decoder-only	125M-13B	Pre-trained on 800GB of text data
FLAN-T5 [52]	encoder-decoder	77M-11.3B	Instruction fine-tuned on T5
MMI [39]	decoder-only	355M	Fine-tuned from GPT2 on 27GB of conversational data, with the objective of predicting source sentences from responses

Table 4.1: Overview of models used in the study. The parameter size indicates the range of model variants evaluated, from smaller to larger versions. MMI stands for Maximum Mutual Information scoring function.

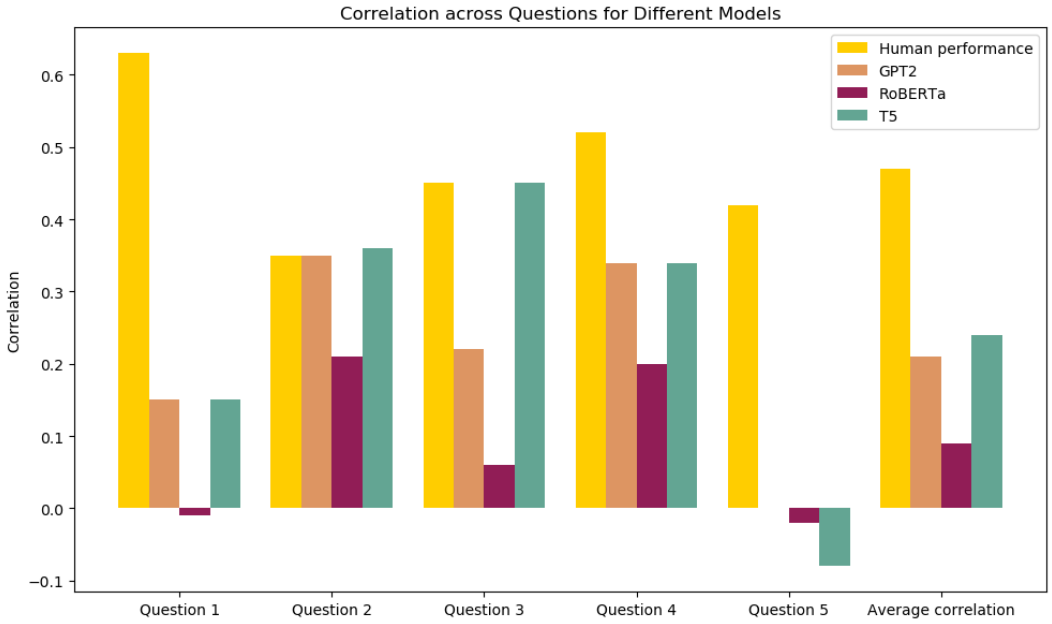

Figure 4.1: Kendall's τ between model and human rankings for each question. "Human Performance" reflects the estimated upper bound for LLMs. The content of Questions 1-5 is detailed in Section 3.2.1.

Given that some candidate answers show low agreement among human participants, it is unreasonable to criticize LLMs for their low correlation with human preferences when humans themselves struggle to reach a consensus. Therefore, estimating human performance is crucial to establish an upper bound for LLM performance. Human performance was estimated using a one-vs-rest approach, where each participant's ratings were compared against the average ratings of the remaining participants. The overall average performance across all participants was used as the estimate of human performance.

As shown in Figure 4.1, RoBERTa's rankings of the candidate answers show a poor correlation with the human benchmark across all questions, often approaching a negligible correlation strength. In contrast, GPT2 and T5 demonstrate better alignment with human rankings, exhibiting moderate correlations for most questions. However, despite their stronger performance relative to RoBERTa, both GPT2 and T5 still fall significantly short of human performance. The models' performance also varies considerably across different questions. Notably, for question 2, both GPT2 and T5 achieve correlations that are close to the estimated human upper bound. In contrast,

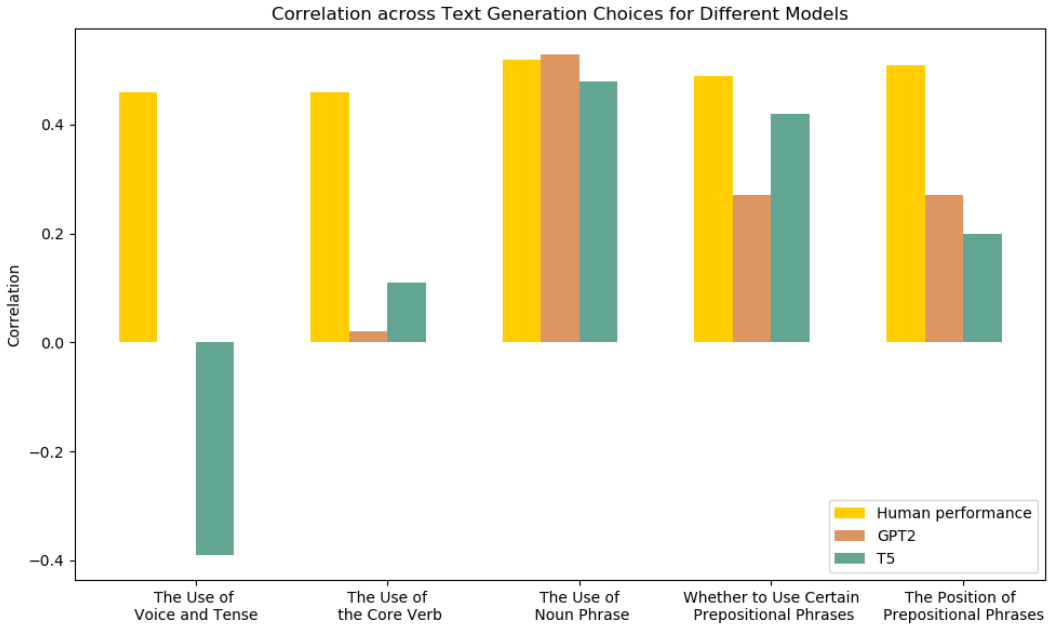

Figure 4.2: Kendall's τ between model and human rankings for each text generation choice. "Human Performance" reflects the estimated upper bound for LLMs.

all models show little correlation with human rankings on question 5.

Given the variation in model scales, it's impossible to isolate the impact of architecture on performance. However, RoBERTa, despite having a similar parameter size to GPT2 and being pre-trained on a larger dataset, exhibits weaker performance. Moreover, its encoder-only architecture, which commonly uses a masked language modeling training objective, requires multiple inferences to compute a sentence score. In contrast, decoder-only models like GPT2 and encoder-decoder models like T5 can compute the score in a single inference. Considering RoBERTa's underperformance despite its considerable scale, along with its higher computational cost compared to GPT2 and T5, we argue that encoder-only models are not suitable for the text-scoring task.

As discussed in Section 3.2.4, different text generation choices exhibit different levels of impact on human preference. This motivates us to investigate models' performances on each text generation choice. As shown in Figure 4.2, models struggled to align with human preferences regarding the use of voice, tense, and core verbs. However, models demonstrated near-

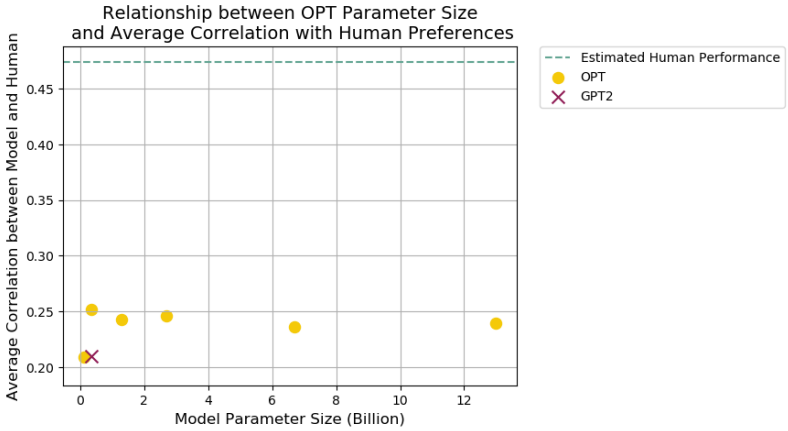

Figure 4.3: The relationship between OPT parameter size and performance. The performance is measured by the average Kendall's τ between model and human rankings across all questions.

human performance in aligning with preferences related to the use of noun phrases. Considering the analysis from Section 3.2.4, it becomes clear that the text generation choices where the models performed poorly are those with a smaller impact on human preferences, while the choices where the models performed well are those with a stronger impact on human preferences.

4.2 Model Scale

In the previous experiment, GPT2 and T5 demonstrated a moderate correlation with human preferences, but substantial gaps remain between their performance and the estimated human performance. This experiment investigates the impact of the model scale on performance in the text-scoring task. OPT and FLAN-T5, which include more variants at different scales, are used for this purpose. These models are considered the successors of GPT2 and T5, respectively. The parameter sizes and training dataset details are shown in Table 4.1.

Parameter size and training set size are key components of the model scale. However, as illustrated in Figure 4.3 and 4.4, increasing the model's parameter size does not consistently lead to better performance in the text-scoring task. For example, the OPT billion-parameter variants have sim-

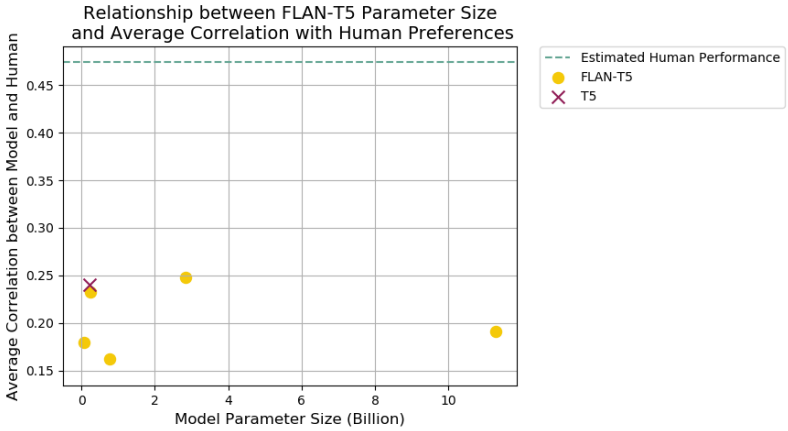

Figure 4.4: The relationship between FLAN-T5 parameter size and performance. The performance is measured by the average Kendall's τ between model and human rankings across all questions.

ilar performance as the smaller variant with 355M parameters, while the FLAN-T5 model with 783M parameters performs worse than its smaller 248M counterpart. Since all variants of each model are trained on the same dataset, it is evident that simply increasing parameter size is not effective in improving performance. Similarly, the OPT variant which has a similar number of parameters as GPT2, does not show a significant performance gain compared to GPT2, even though its training set is more than 20 times larger than GPT2. This suggests that expanding the training set size alone is also ineffective in enhancing model performance in this task.

Overall, regardless of model scale, all models show a weak correlation with human preferences and fall significantly short of estimated human performance. Thus, increasing the model scale—whether through parameter size or training set size—is not an effective solution for improving performance in the text-scoring task.

4.3 Likelihood Trap

The strong performance of LLMs on various natural language processing tasks makes their shortcomings in the text-scoring task surprising. To understand this, a deeper analysis was conducted to explore the differences between model rankings and human benchmark rankings.

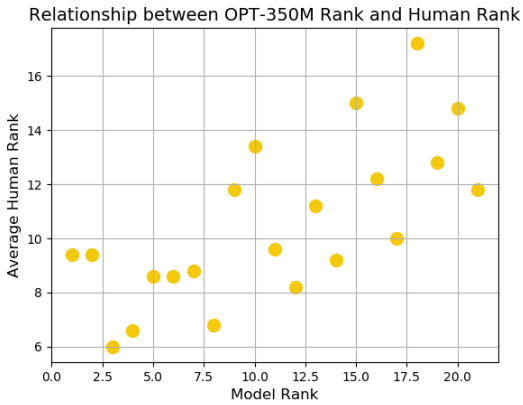

Figure 4.5: The comparison of OPT-350M rankings with average human rankings. The average human ranking represents the mean of human benchmark rankings for candidate answers that have the same model rank.

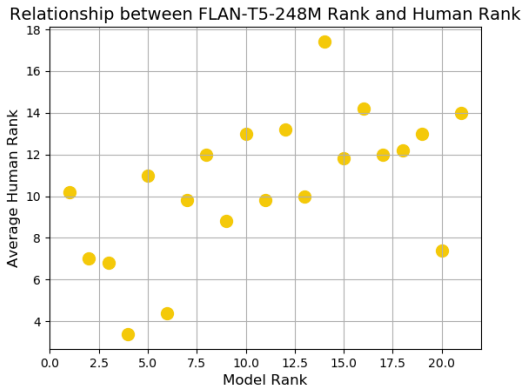

Figure 4.6: The comparison of FLAN-T5-248M rankings with average human rankings. The average human ranking represents the mean of human benchmark rankings for candidate answers that have the same model rank.

Question	Answer	Model rank	Human rank
When did Alice send the payment to Bill?	Alice sent the payment to Bill.	1	19
	Alice sent the payment to Bill on March 24th 2024.	2	1
What is the connection between Alice and Bill?	Alice is connected to Bill.	1	21
	On March 24th 2024, Alice sent a payment of 2000 euro to Bill.	3	3

Table 4.2: Examples of the first issue observed in the FLAN-T5-248M model, where bland answers lacking requested information are ranked higher than concrete answers preferred by humans.

As shown in Figure 4.5 and Figure 4.6, both the OPT and FLAN-T5 models tend to rank the answers most favored by humans around the third to fifth position. Additionally, after the fifth position, model rankings fluctuate but generally increase as human rankings go higher. This indicates that the models are not completely misaligned with human preferences.

However, a key issue arises when these models place bland, uninformative answers—those rated lowest by humans—at the top, while concrete, well-constructed answers that humans strongly prefer are placed lower. Table 4.2 provides examples where FLAN-T5 ranks answers that lack the requested information higher than more informative responses.

Another issue is the models' difficulty in prioritizing natural, fluent answers among candidate answers that contain the requested information. Examples of this issue are shown in Table 4.3.

Though the examples used are from OPT-350M, and FLAN-T5-248M, the pattern is not unique to these models. It is observed in GPT-2, OPT, and other FLAN-T5 variants across different parameter sizes.

The first issue, in particular, has broader implications for the design of other components in a data-to-text conversational system. If the text scorer cannot effectively filter out the uninformative answers, the query system must precisely extract the fine-grained information requested by the user, and the generative grammar must ensure this information is included dur-

Question	Answer	Model rank	Human rank
When did Alice send the payment to Bill?	Alice sent the payment on March 24th 2024 to Bill.	12	15
	Alice, on March 24th 2024, sent the payment to Bill.	17	12
Who sent Bill 2000 euro on March 24th 2024?	A Dutch woman named Alice sent 2000 euro on March 24th 2024.	20	12
	A Dutch woman named Alice sent 2000 euro on March 24th 2024 to Bill.	13	14

Table 4.3: Examples of the second issue in the FLAN-T5-248M model, where the model struggles to prioritize natural and fluent answers.

ing text generation.

Only a few uninformative answers were included in the original human preference dataset. To further investigate LLMs' behavior when scoring uninformative answers, a new dataset was created.

Informativeness dataset This dataset consists of 20 wh-questions related to the transaction information described in Section 3.2.1. Each question is paired with two candidate answers, both of which represent possible outputs from the generative grammar. The sole difference between the answers lies in the inclusion or omission of the requested information, which is the result of a single decision divergence in the grammar derivation process. Examples from this dataset are provided in Table 4.4.

Since humans are expected to prefer the informative answer, a model aligned with human preferences should assign a higher score to these answers compared to their uninformative counterparts. The model's performance is evaluated by the percentage of questions where it correctly scores the informative answer higher.

From this point forward, this dataset will be referred to as the "informativeness dataset", and the dataset collected from human participants as the "human preference dataset". Both datasets assess the model's ability to align with human preferences in selecting candidate answers. The informativeness dataset specifically focuses on answers that differ in whether they

Question	Uninformative Answer	Informative Answer
Who sent Bill 2000 euro on March 24th 2024?	2000 euro was sent to Bill on March 24th 2024.	2000 euro was sent to Bill on March 24th 2024 by Alice, a Dutch woman.
How much was the payment Alice sent to Bill?	Alice paid Bill.	Alice paid Bill 2000 euro.
When did Alice send the payment to Bill?	Alice sent Bill the payment.	On March 24th 2024, Alice sent Bill the payment.
Who was paid by Alice on March 24th 2024?	A person was paid by her on March 24th 2024.	A 37-year-old man named Bill was paid by her on March 24th 2024.

Table 4.4: Examples from the informativeness dataset. Each question is paired with two answers, one lacking requested information (uninformative) and one providing the requested information (informative).

Model	Percentage of correctly scored questions
GPT2	0.5
OPT-350M	0.4
OPT-13B	0.3
T5	0.1
FLAN-T5-248M	0.4
FLAN-T5-11.3B	0.45

Table 4.5: Model performance on the informativeness dataset. The performance is measured by the percentage of questions where each model correctly assigns a higher score to the informative answer.

contain the requested information. The human preference dataset provides a broader overview of how well the model aligns with human judgment across all types of generated answers.

As shown in Table 4.5, all models, regardless of architecture or scale, performed poorly in prioritizing informative answers. The worst-performing model, T5, assigned a higher score to the bland answer in 90% of the cases. Even the best-performing model, GPT2, selected the informative answer only half the time, which remains far from ideal.

This counter-intuitive behavior, where neural language models tend to favor bland, uninformative texts, has been documented by researchers and is referred to as the "likelihood trap"(see section 2.5). One explanation for this is that uninformative responses appear more frequently in training data, leading the model to assign them higher likelihoods. In contrast, informative responses, especially those containing specific details not seen

Model	Percentage of correctly scored questions
GPT2	0.5
MMI	0.7
GPT2+MMI	0.8

Table 4.6: MMI Model performance on the informativeness dataset. The performance is measured by the percentage of questions where each model correctly assigns a higher score to the informative answer.

during training, are penalized with lower likelihoods.

4.4 Re-ranking Model

One solution to mitigate the likelihood trap is the use of re-ranking models. The Maximum Mutual Information (MMI) scoring function is often employed in conversational systems to re-rank the top N candidate responses generated by a standard forward language model. MMI operates as essentially a pre-trained backward model, predicting the likelihood of the source question given the response, i.e., P(Question | Answer). The idea behind this is that uninformative responses are likely to correspond to many possible questions, leading to a lower probability for the specific question.

To evaluate the effectiveness of MMI re-ranking, a comparison was made between three approaches: using the forward model (GPT2) alone, using the backward model (MMI) alone, and combining both models. This comparison was conducted on both the human preference and informativeness datasets. GPT2 was selected for comparison because the MMI model shares the same architecture as GPT2 and is specifically designed to improve the quality of responses generated by GPT2-like models. In the combined approach, the score from the MMI model is added to the score from GPT2 for each answer.

As shown in Table 4.6, MMI performs better than GPT2 at filtering out uninformative answers. However, as shown in Table 4.7, MMI performs similarly to GPT2 on the general human preference dataset. The two main issues identified earlier—the inability to filter out uninformative answers and the failure to prioritize natural, fluent answers—are key factors in the

Model	Average correlation between model and human
GP2	0.21
MMI	0.19
GPT2+MMI	0.20

Table 4.7: MMI Model performance on the human preference dataset. The performance is measured by the average Kendall's τ between model and human rankings across all questions.

model's poor alignment with human preferences. MMI's success in filtering out uninformative answers, yet its poor overall alignment with human preferences, suggests it struggles to prioritize natural and fluent responses. Similarly, the combined approach does not offer any improvement in aligning with overall human preferences. Though it is the most effective at filtering out uninformative answers, its results are still far from ideal. The 20% rate of selecting bland answers over informative ones indicates that other components of the system must share the responsibility of information selection. Thus, MMI re-ranking is considered ineffective in improving model performance in the text-scoring task.

4.5 The Use of Prompts

The transaction data involved in the Q&A is external knowledge not available to LLMs during training. As a result, instead of assigning a high likelihood to answers containing unfamiliar details, LLMs tend to favor more generic responses. In this section, the impact of providing this external knowledge to LLMs through prompt is investigated.

Each prompt consists of a short task description, the question, and the relevant transaction data. For example:

```
Answer the following question naturally and informatively
based on the transaction data:
Who sent Bill 2000 euro on March 24th 2024?

Transaction data:
From: Alice, Dutch, Female
```

```
To: Bill, 37-year-old, Male
Amount: 2000 euro
Date: March 24th 2024
```

In this section, all results are based on $P(Answer|Prompt)$ rather than $P(Answer|Question)$.

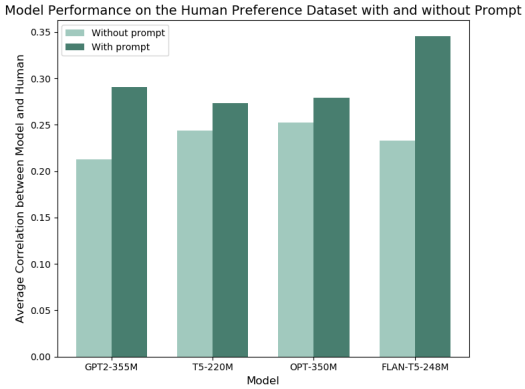

Figure 4.7: Model Performance on the Human Preference Dataset with and without Prompt. The performance is measured by the average Kendall's τ between model and human rankings across all questions.

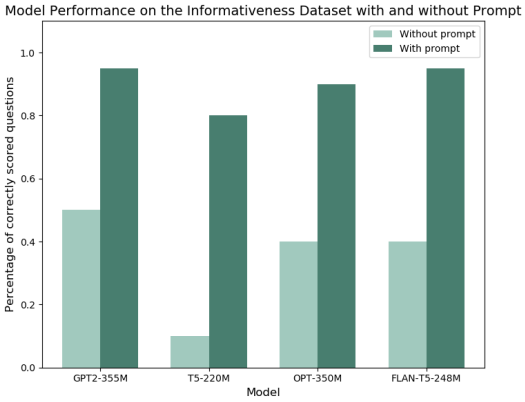

Figure 4.8: Model Performance on the Informativeness Dataset with and without Prompt. The performance is measured by the percentage of questions where each model correctly assigns a higher score to the informative answer.

As illustrated in Figures 4.7 and 4.8, using prompts improves the performance of all models on both the human preference and informativeness datasets.

In Figure 4.7, T5 and FLAN-T5, which have similar parameter sizes, perform similarly without prompts in terms of alignment with human preferences. However, after using prompts, FLAN-T5 shows a significant per-

formance growth, far surpassing T5. This performance growth can be attributed to FLAN-T5's instruction fine-tuning on over 1000 tasks, which is the key difference between the two models.

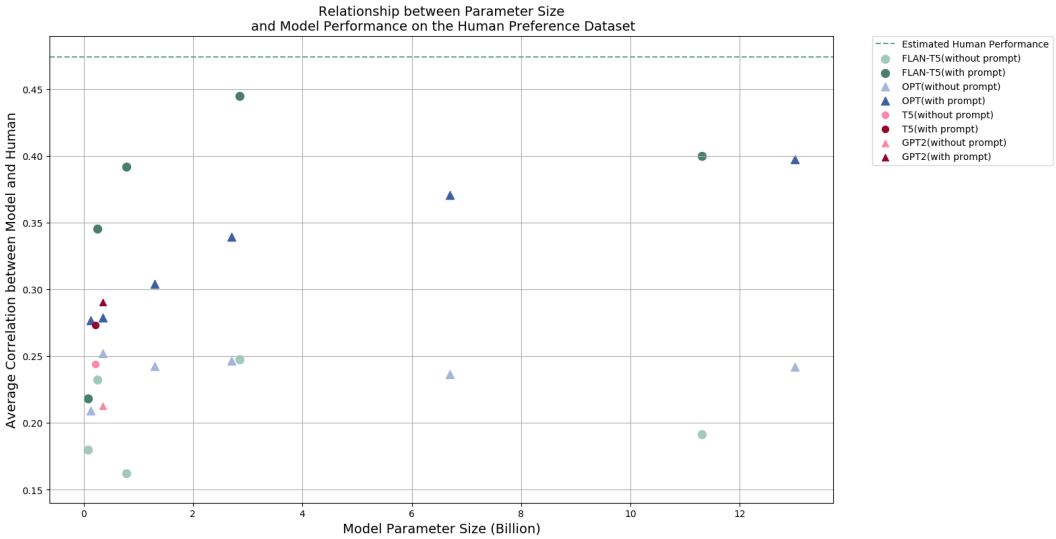

Figure 4.9: Relationship between Parameter Size and Model Performance on the Human Preference Dataset. The performance is measured by the average Kendall's τ correlation between model and human rankings across all questions.

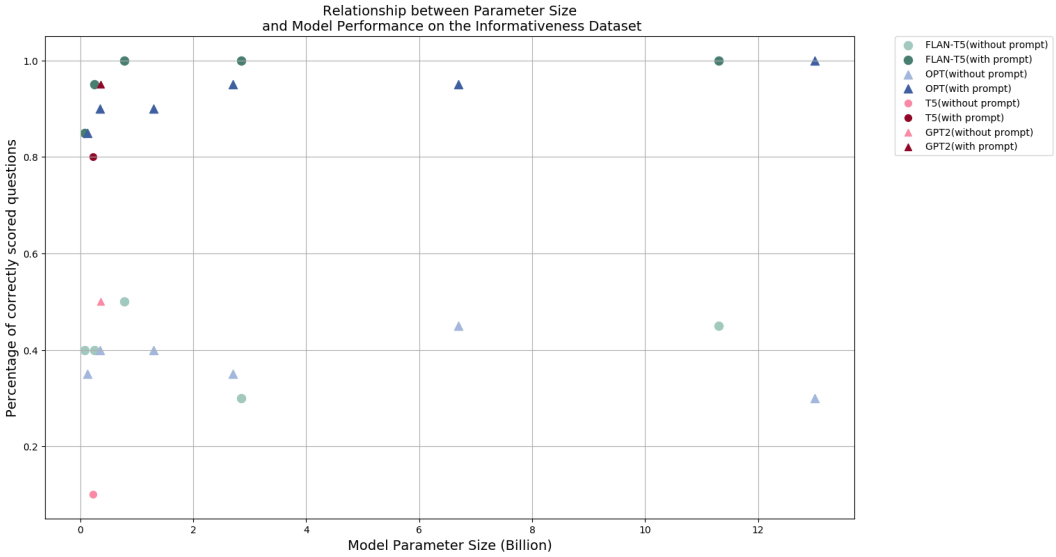

Figure 4.10: Relationship between Parameter Size and Model Performance on the Informativeness Dataset. The performance is measured by the percentage of questions where each model correctly assigns a higher score to the informative answer.

Scaling up model size, however, does not guarantee better performance. As illustrated in Figure 4.9, after using prompts, OPT with 350M parame-

Prompt content	Informativeness	Human preference
Question	0.5	0.16
Task description + Question	0.4	0.17
Data + Question	1.0	0.30
Shuffled data + Question	1.0	0.29
Task description + Data + Question	1.0	0.39

Table 4.8: The effect of varying prompt content on FLAN-T5-783M's performance across the informativeness and human preference datasets. The performance on the informativeness dataset is measured by the percentage of questions where each model correctly assigns a higher score to the informative answer. The performance on the human preference dataset is measured by the average Kendall's τ correlation between model and human rankings across all questions.

ters is outperformed by GPT2, which has a similar parameter size but was trained on a much smaller dataset. Similarly, OPT-1.3B, despite having more parameters and a larger training set than GPT2, does not show a significant performance advantage. Figure 4.10 further demonstrates that GPT2 performs as well as or better than larger OPT models when using prompts.

While it can be observed that OPT's performance generally improves with an increase in parameters, the comparison between GPT2 and OPT models suggests that simply scaling up model size is not a necessary condition for improving performance in the text-scoring task.

Overall, FLAN-T5 demonstrated superior performance across both datasets after the use of prompts. FLAN-T5-2.85B aligns most closely with human preferences, achieving a Kendall's correlation of 0.44, which approaches the estimated human performance of 0.47. FLAN-T5-783M stands out as the smallest model with a 100% success rate in filtering out uninformative answers, while also achieving a Kendall's correlation of 0.39—comparable to human performance. Among models with fewer than 1 billion parameters, it has the best performance.

To investigate the impact of different prompt component on model performance, a series of comparison experiments was conducted, with results shown in Table 4.8. On both datasets, using only the task description as the prompt led to performance similar to not using a prompt at all. In contrast, using only the data as the prompt resulted in a significant improvement

compared with no prompt. When data was removed from the full prompt, performance dropped sharply, while removing the task description caused only a slight change. This suggests that the performance boost from using prompts is largely due to the inclusion of the data.

To further explore the role of adding data in the prompt, the sequence of words describing the data was randomly shuffled several times, and the model's average performance was computed. An example of shuffled data is:

```
March 2000 24th Male Female data: Transaction Alice,
Dutch, 2024 From: Date: 37-year-old, euro Bill, Amount: To:
```

As shown in Table 4.8, the model's performance on both datasets remained almost unchanged whether the data was shuffled or not. Although shuffled data cannot convey accurate information, its presence in the context still led to the same performance boost as the unshuffled data. This suggests that the primary role of the data is not to communicate specific information, but rather to introduce rare data-related tokens into the context so that answers containing these details are not penalized due to their under-representation in the training data.

4.6 Grammar and Text Scorer Integration

To evaluate the effectiveness of the proposed neural-symbolic approach in enhancing data-to-text systems, this experiment constructs a basic grammar and tests the LLM text scorers using actual grammar-generated outputs, following the workflow illustrated in Figure 1.1. The information retrieval system is not included in this experiment. The same transaction information from Section 3.2.1 was used to simulate the information retrieval system's output.

The grammar used in the experiment is a basic feature-based context-free grammar, primarily designed to ensure the factual accuracy of the generated texts while allowing diversity in how the same transaction information is described. However, other aspects of text generation, such as well-

Question	Answer	Time(second)
When did Alice send the payment to Bill?	Alice sent the payment on March 24th 2024.	163.84
Who sent Bill 2000 euro on March 24th 2024?	Alice was sending Bill 2000 euro on march 24th 2024.	171.92
How much money did Alice send to Bill on March 24th 2024?	Alice sent Bill a 2000euro payment on March 24th 2024	182.24
Who received a payment from Alice on March 24th 2024?	Bill received a payment of 2000 euro on March 24th 2024 from her.	170.95
What is the connection between Alice and Bill?	Bill received a payment of 2000 euro on March 24th 2024 from her.	163.64
Tell me more about Alice.	The Dutch woman sent a 2000 euro payment on march 24th 2024.	165.43
Describe the transaction between Bill and the Dutch woman.	Bill received 2000 euro from her on March 24th 2024.	170.74

Table 4.9: Responses generated by the system for each question, along with the total time taken for sentence generation and text scoring. The information retrieval system is not included in this experiment.

formedness, were not carefully considered, leading to the potential for ill-formed sentences. For example, the grammar may produce sentences like "A man named Bill named Bill" or "The payment from Alice was sent by Alice," where certain phrases are redundantly repeated. Based on previous results (Section 4.5), FLAN-T5-783M was selected as the text scorer, as it performed best among models with fewer than 1 billion parameters, demonstrating near-human performance. The experiment was conducted on an NVIDIA L4 GPU. Sentences were processed in batches of 50 to reduce inference time.

For each input question, the grammar generated 10,000 distinct sentences describing the transaction information. The text scorer then evaluated these sentences and selected the one with the highest score. The total time was calculated as the sum of the sentence generation and text scoring times. The number 10,000 was chosen to ensure broad coverage of the generated sentences, providing a diverse set of potential outputs to comprehensively assess the text scorer's ability. The fine-tuning of this number is left for future studies, as it may depend on the complexity of the grammar and the specific task requirements.

As shown in Table 4.9, the generated responses remain faithful to the transaction information, demonstrating that the grammar component ensures factual accuracy. At the same time, the results show that the text scorer effectively filters out ill-formed sentences. This suggests that the linguistic knowledge acquired by the LLM can fill gaps left by the grammar. This supports the idea that the grammar design can be simplified within the neural-symbolic system, as discussed in Section 2.1, shifting the burden of encoding linguistic knowledge to the neural model while maintaining con-

trol over factual accuracy. Examples of generated texts and their scores are shown below.

```
What is the connection between Alice and Bill?


Bill received a payment of 2000 euro on March 24th 2024 from her.
-1.05
Bill was receiving 2000 euro on March 24th 2024.
-1.35
Bill has received the payment from Alice.
-1.84
Alice has given the 2000 euro payment to the man on March 24th 2024.
-2.20
Bill named Bill named Bill had been receiving a 2000 euro payment.
-3.43
She had been sending a man it.
-4.68
Alice was giving him it.
-5.01
A payment was given by Alice named Alice named Alice.
-5.62
```

Additionally, repetitive and bland answers were assigned lower scores compared to more informative ones, as illustrated in the examples above. This indicates that the "likelihood trap" issue—where models favor uninformative responses—is effectively mitigated by the text scorer.

Moreover, the combined system of grammar and text scorer, without further optimization, can generate well-formed answers in approximately 3 minutes. The system also demonstrates the ability to handle various query types, not limited to wh-questions.

These findings suggest that the neural-symbolic approach, which integrates a grammar component with a large language model as the text scorer, is effective for the data-to-text generation task.

5. Discussion and Future Study

5.1 Word Frequency

As discussed in Section 2.2, Lau et al. highlighted that while human perception of text is unaffected by factors like sentence length and word frequency, language model (LM) judgments are influenced by both. Their experiments demonstrated that normalizing sentence length and word frequency can significantly improve the correlation between model judgments and human preferences. In this study, however, only sentence length was normalized. This is primarily because, unlike sentence length, word frequency is difficult to measure accurately, due to the lack of transparency in LLM training data. Furthermore, preliminary experiments revealed that performance gains were largely attributed to sentence length normalization, with little added benefit from explicit word frequency normalization. This may be because the influence of word frequency cannot be easily mitigated through simple normalization, especially when the word frequency data may not accurately reflect real-world usage. Interestingly, the experiments showed that even shuffled prompts, which contain no coherent information, provided similar performance boosts as unshuffled prompts. Given that these prompts consist largely of rare, data-related tokens, we raise the possible interpretation that the prompt may be functioning as a factor that mitigate the influence of word frequency from the sentence score.

5.2 Scale

The experimental results indicate that increasing the model scale does not always result in better performance on the text-scoring task. For example, while the FLAN-T5 model improves as its parameter count increases from 77M to 2.83B, the 11.3B variant performs similarly to the much smaller

783M model (see Figure 4.9). Similarly, the OPT-1.3B model, despite having more parameters and data than GPT2, achieves performance comparable to GPT2.

This suggests that larger models are not necessarily more effective for the specific capabilities required by the text-scoring task. As discussed in Section 2.7, LLMs' linguistic competence can be divided into formal linguistic competence—the ability to produce and comprehend language structures—and functional linguistic competence, which involves cognitive functions used when applying language in real-world contexts.

Given these definitions, we propose that the text-scoring task likely relies more on formal linguistic competence than on functional competence. Previous studies have found that improving functional linguistic abilities requires significantly more training [45]. Therefore, a plausible explanation for the similar performance of larger and smaller models is that the superiority of larger models in functional competencies, such as formal reasoning and world knowledge, is not reflected in this task.

Additionally, earlier research suggests that language models need only around 10M to 100M words to acquire fundamental linguistic knowledge [46]. Thus, a potential explanation for the similar performance of FLAN-T5-783M and FLAN-T5-11.3B is that the model may have reached a plateau in the formal linguistic competence required for the text-scoring task at around 783M parameters, with further scaling contributing more to functional rather than formal capacities. The better performance of FLAN-T5-2.83B compared to both FLAN-T5-783M and FLAN-T5-11.3B could be attributed to performance variability, given the relatively small dataset. Similarly, the comparable performance between OPT-1.3B and GPT2 may reflect similar levels of formal linguistic competence, despite differences in functional capabilities.

5.3 Close-source LLM

The scope of this study is limited to open-source LLMs in the exploration of text scoring. This is due to the fact that this study focuses on the use of

Metric	Question 1	Question 2	Question 3	Question 4	Question 5	Average
Kendall's τ	0.49	0.65	0.30	0.50	0.44	0.48

Table 5.1: Kendall's τ between human rankings and GPT-4o rankings for each question. The average value is the correlation coefficients averaged across questions.

Text generation choices	ChatGPT 4o
the use of voice and tense	0.62
the use of core verb	0.09
the use of noun phrases	0.26
whether to use certain prepositional phrases	0.61
the position of the used prepositional phrases	0.40

Table 5.2: Kendall's τ between human rankings and GPT-4o rankings for each text generation choice, averaged across all questions.

sentence-scoring methods and sentence-scoring methods are not applicable to closed-source LLMs. Calculating sentence scores requires access to the logits of the input tokens, which are available in open-source models. In contrast, this information is not transparent in closed-source models. For instance, models like Gemini do not expose logit information, while others, such as GPT series models, only provide logits for generated tokens. Additionally, closed-source models often contain hundreds of billions of parameters. The use of these models also comes with a price. This runs counter to the objective of minimizing the scale and cost of the model, which is a key consideration in this study.

To provide a reference to the performance of current close-source LLMs in the task, GPT-4o, the latest model in the ChatGPT family, was prompted to rank the candidate answers for each question. The prompts were written in a manner similar to that of the surveys offered to humans. Each prompt contains the context, the transaction information, a question, and all candidate answers to that question.

As shown in Table 5.1, GPT-4o exhibits near-human performance. However, it aligns poorly with human performance in terms of core verb usage. Though GPT-4o outperforms all open-source LLMs, the comparison is not fair due to the fundamental differences in the ranking methods applied to these two types of models. For GPT-4o, the ranking task was reformulated

as a text generation task, where the model generates ranked candidate answers. In contrast, for the open-source models, the task was essentially approached as a language modeling problem, where rankings were based on the probabilities of different word sequences. It is notably that GPT-4o's response is not consistent. Feeding the same prompt to GPT-4o several times yields different rankings. In addition, it also exhibits the behaviour of skipping certain candidate answers during ranking as well as putting the same candidate answers in different places in a single shot ranking.

5.4 Future Study

Due to the capacity limit of this study, more fine-grained text-generation choices, such as the use of determiners, were not considered during dataset construction. Additionally, while collecting human preferences through pairwise comparison would likely improve the reliability of human benchmark rankings, absolute ratings were used due to the limited capacity of the participants. Future research could address these limitations by constructing a more comprehensive dataset.

This study focused on the text-scoring task using sentence scoring, which is only applicable to open-source LLMs. Other approaches, such as comparative assessment, could be explored in future work.

Another promising direction for future research is the application of knowledge distillation techniques to reduce model size. A recent study by Fu et al. demonstrated that smaller models can achieve strong performance on specific tasks by concentrating their modeling power on the target task [53].

While the use of LLMs as text scorers simplifies grammar rule design, the computational cost may increase if the text scorer is tasked with filtering out a large number of problematic texts generated by the grammar. Therefore, future research should explore the trade-off between reducing grammar development efforts and controlling the computational costs of the text scorer. Similarly, the balance between the information retrieval system and the text

scorer requires further investigation. If the information retrieval system can retrieve fine-grained data tailored to answer specific queries, the scale and computational cost of the text scorer may be reduced further. Studying these trade-offs will be key to improving overall system efficiency.

6. Conclusion

To summarize, this study focused on investigating the use of the LLM as the text scorer component in a neural-symbolic data-to-text conversational system. The key objective is to assess whether LLMs can align with human judgments regarding grammar-generated answers for a given question.

A benchmark dataset of human preferences was constructed, revealing that text generation choices with a significant impact on human preferences tend to influence how much information is conveyed, while less impactful choices affect the style in which the information is presented. It is also found that human participants have low preference agreement on certain candidate answers, reflecting the natural variation in individual preferences.

Sentence scoring was used to evaluate model judgments, and the experiments demonstrated that encoder-only models performed poorly on this task. All models struggled with the "likelihood trap", where they favored bland, uninformative responses over more informative ones. Reranking models were ineffective in addressing this issue, whereas the use of prompts, even when shuffled, proved effective. Since the shuffled prompt is also effective, we raise the possible interpretation that the prompt may be functioning as a factor that mitigates the influence of word frequency on the sentence score.

It was also observed that increasing the model scale does not guarantee better performance on the text-scoring task. We hypothesize that the scale growth primarily contributed to the enhancement of functional linguistic competencies, such as world knowledge, rather than formal linguistic competence, which is more critical for the text-scoring task. As a result, increasing the model scale does not necessarily improve performance in this context. The FLAN-T5 model, possibly benefiting from instruction fine-tuning, outperformed other models, with the 783M-parameter variant

achieving near-human performance.

Finally, the integration of a basic generative grammar with the LLM text scorer demonstrated the effectiveness of the neural-symbolic approach. The LLM's rich linguistic knowledge simplifies the grammar design, while the grammar guarantees accurate data representation in the generated text. This approach balances naturalness with factual accuracy, as well as the trade-off between the development cost of grammar and the computational cost of the LLM, highlighting the potential of the neural-symbolic method for data-to-text generation.

A. Preliminary Experiment

A.1 Background

As discussed in Section 2.4, modifying sentence probabilities can produce scores that align better with the intended scoring objectives. In both neural machine translation and human acceptability judgment prediction, length normalization is commonly applied to prevent shorter sentences from being favored by pure probability scores [20], [33]. Additionally, Lau et al. argue that word frequency should be adjusted in sentence probabilities, as human acceptability judgments are not affected by this factor [16], [20].

The goal of this preliminary experiment is to compare various normalization methods from previous studies to identify the most effective approach for use in subsequent experiments.

A.2 Methodology

A.2.1 Dataset

The preliminary experiment uses the BNC and ENWIKI datasets[1] from Lau et al.'s study, both containing a diverse set of sentences with human acceptability judgments [13], [16]. The BNC dataset consists of sentences from the British National Corpus, while the ENWIKI dataset includes sentences from English Wikipedia. Each dataset contains 2,500 sentences, with approximately 10 human ratings per sentence. The performance of each method is evaluated using the Pearson correlation coefficient, which measures the correlation between the normalized sentence probability and the average human rating.

[1] https://gu-clasp.github.io/projects/smog/experiments/

Normalization method	Equation
LogProb	$\log P_m(W)$
MeanLP [15]	$\frac{\log P_m(W)}{\|W\|}$
PenLP [33]	$\frac{\log P_m(W)}{((5+\|W\|)/(5+1))^{\alpha}}$
NormLP(token) [16]	$-\frac{\log P_m(W)}{\log P_{token}(W)}$
NormLP(word)	$-\frac{\log P_m(W)}{\log P_{word}(W)}$
SLOR [54]	$\frac{\log P_m(W) - \log P_{token}(W)}{\|W\|}$
Delta [34]	$\frac{\log P_m(W) - \log P_m(W')}{\|W\|}$

Table A.1: Overview of the normalization methods used in the preliminary experiment. The log probability of a sentence, $\log P_m(W)$, is computed using Equation A.1, while $\log P_{token}(W)$ and $\log P_{word}(W)$ represent the token-level and word-level unigram probabilities, respectively. Calculation of the unigram prabability follows Equation A.2 and A.3. W' refers to the word-level shuffled version of the original sentence W, and $|W|$ denotes the sentence length in tokens. The parameter α is set to 0.8 following the approach used by Lau et al [20].

A.2.2 Normalization Methods

In this experiment, the language model used was GPT2. Given a sentence $W = \{w_1, ..., w_{|W|}\}$, where w_i is the i-th token of W and $|W|$ is the sentence length, the log probability of W is calculated as:

$$\log P_m(W) = \sum_{t=1}^{|W|} \log P_m(w_t | W_{<t}) \tag{A.1}$$

Here, $W_{<t}$ represents the sequence of tokens prior to w_t. The frequency of a token w_i in the training data is computed as:

$$\frac{n_{train}(w_i)}{N_{train}} \tag{A.2}$$

Here, $n_{train}(w_i)$ is the number of occurrences of w_i, and N_{train} is the total number of tokens in the training set. Using this, the unigram probability of the sentence is defined as:

$$\log P_{token}(W) = \sum_{t=1}^{|W|} \log \frac{n_{train}(w_t)}{N_{train}} \tag{A.3}$$

Since GPT2's training data is not publicly available, the unigram probability was approximated using an open-source re-implementation of GPT-2's training corpus[2].

Lau et al. proposed NormLP(uni) to account for word frequency in sentence probabilities, but this method conflates token frequency and word frequency, as P_{token} reflects token frequency [16]. Due to GPT-2's Byte Pair Encoding (BPE) tokenization, token frequency does not directly correspond to word frequency. BPE often splits words into subword units, causing rare short words to have higher unigram probabilities than longer, more frequent words that are split into multiple tokens.

To address this, we propose NormLP(word), which normalizes using word frequency instead of token frequency. The word-level unigram probability P_{word} is calculated similarly to P_{token}, but the smallest unit is the word rather than the subwords.

A.3 Results

As shown in Table A.2, all normalization methods significantly outperform the baseline LogProb, indicating that normalizing sentence length, word frequency, or both can improve alignment between sentence probabilities and human judgments.

MeanLP and PenLP, which normalize sentence length, both show improved performance. Notably, MeanLP, which simply averages LogProb by

[2]https://skylion007.github.io/OpenWebTextCorpus/

Normalization method	BNC	ENWIKI
LogProb	0.33	0.33
MeanLP	0.62	0.60
PenLP	0.54	0.53
NormLP(token)	0.58	0.55
NormLP(word)	0.60	0.56
SLOR	0.53	0.56
Delta	0.62	0.57

Table A.2: Pearson correlation coefficients between the normalized sentence probabilities and average human ratings for different normalization methods on the BNC and ENWIKI datasets. A higher correlation score indicates better alignment with human judgments.

sentence length, outperforms the more complex PenLP. PenLP was originally developed for neural machine translation, where the parameter α is typically fine-tuned on a development set. Its relatively poor performance here may be due to the lack of fine-tuning or its unsuitability for predicting human acceptability judgments.

NormLP(token) and NormLP(word), which both intend to normalize word frequency, show that NormLP(word) outperforms NormLP(token). We speculate that this could be because token frequency has a limited impact on both sentence probability and human acceptability judgments, while word frequency might influence sentence probability more directly without affecting human acceptability. Thus, the normalization of word frequency appears more effective than token frequency normalization.

SLOR and Delta both account for both sentence length and word frequency. SLOR performs worse than Delta, despite showing strong performance with N-gram and RNN-based language models in the same task. This suggests that SLOR may not be well-suited for normalizing probabilities in LLMs. Delta performs similarly to the top method, MeanLP, but its higher computational cost (requiring sentence probabilities to be computed twice) makes it impractical for use in subsequent experiments.

In conclusion, MeanLP is selected for the following experiments due to its simplicity and superior performance in aligning with human judgments.

B. Ethics and Privacy Quick Scan Report

This research followed the ethics and privacy regulations of the Utrecht University Research Institute of Information and Computing Sciences. The Ethics and Privacy Quick Scan Report is provided in this annex, with non-applicable questions omitted and the "Your Information" section removed for anonymization.

Ethics and Privacy Quick Scan (version: 15 July 2024)

Section 1. Research projects involving human participants

		Yes	No
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 X, Reddit) without directly recruiting participants, please answer no.	**Yes**	

If no, continue with Section 2; if yes, fill in the following questions.

Recruitment

		Yes	No
P2	Does your project involve participants younger than 16 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?[1]		**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[2] group, other than those listed above?		**No**

If the answer to all of P2-P6 is no, continue with P8.

		Yes	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?	**Yes**	

[1] For informed consent people need to be able to (1) understand information provided relevant to making the consent decision, (2) retain this information long enough to be able to make a decision, (3) weigh the information, (4) communicate the decision.

[2] Vulnerable people include those who are legally incompetent, who may have difficulty giving or withholding consent, or who may suffer highly adverse consequences if their personal data were to become publicly available or from participating. Examples include irregular immigrants, refugees, sex workers, dissidents and traumatized people at risk of re-traumatization.

If the answer to P8 is yes, please answer P9.

		Yes	No
P9	Is it made clear to potential participants that not participating will in no way impact them (e.g. it will not directly impact their grade in a class)?	**Yes**	

If the answer to P9 is yes, then continue with PC1.

Consent Procedures		Yes	No	Not applicable
PC1	Do you have set procedures that you will use for obtaining *informed* consent prior to collecting data 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[3].)	**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**		

If the answer to PC1-PC7 is yes, then continue with PC8.

		Yes	No
PC8	Does your project involve concealment[4] or deliberate misleading of participants?		**No**

[3] uu.nl/en/research/institute-of-information-and-computing-sciences/ethics-and-privacy

[4] This may for example involve concealment of the study aim, of the identity of the researcher, or subliminal messaging during the study.

If the answer to PC8 is no, continue with Section 2.

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.

		Yes	No
D1	Are you gathering or using personal data (defined as any information relating to an identified or identifiable living person[5])?		**No**

If the answer to D1 is no, continue with Section 3.

Section 3: Research that may cause harm

Research may harm 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.

		Yes	No
H1	Does your project give rise to a realistic risk to the national security of any country?[6]		**No**
H2	Does your project give rise to a realistic risk of aiding human rights abuses in any country?[7]		**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?[8]		**No**

[5] This includes people's name, postal address, unique ID, IP address, voice, photo, video etc. When a person can be identified by combining multiple data points (e.g. gender + age + job role), this also constitutes personal data. When a person can be identified by a simple search online (e.g. with the content of a tweet) this also constitutes personal data. Note that Survey tool Qualtrics by default collects IP addresses and that the survey needs to be anonymized before distribution to prevent this.

[6] For example, research that can be used for autonomous armed vehicles/drones/robots, research on automated detection of objects, research on AI-enhanced forgery of video/audio data.

[7] For example, research on natural language/video/audio processing for automated identification of people's identity, sentiments, or opinions.

[8] For example, research that involves potentially violent participants such as criminals, research in likely unsafe locations such as war zones, research on an emotionally highly challenging topic,

H7	Is there a realistic risk of any participant experiencing physical or psychological harm or discomfort?[9]		**No**
H8	Is there a realistic risk of any participant experiencing a detriment to their interests as a result of participation?[10]		**No**
H9	Is there a realistic risk of other types of negative externalities?[11]		**No**

If the answer to H1-H9 is no, continue with Section 4.

Section 4: Conflicts of interest

		Yes	No
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 or imbalance of power between researchers and participants?		**No**

If the answer to C1-C2 is no, continue with Section 5.

Section 5: Your information

research in which the researcher is alone with a not previously known participant in the participant's home.

[9] For example, research that involves strenuous physical activity, research that stresses participants, research on an emotionally challenging topic.

[10] Detriment to participants' interests may include risks to participants' reputation if the data was disclosed, risks to their livelihoods, risks of prosecution or persecution, etc.

[11] A negative externality is a harm produced to a third party, social group, society in general, or the environment. For instance, intended or unintended negative ethical (e.g. bad governance or management practices), social (e.g. consumerism, inequality, stigmatization) or environmental effects (e.g. large CO_2 footprint or e-waste production) of your project.

Bibliography

[1] R. Tian, S. Narayan, T. Sellam, and A. P. Parikh, "Sticking to the facts: Confident decoding for faithful data-to-text generation," *arXiv preprint arXiv:1910.08684*, 2019.

[2] O. Dušek and Z. Kasner, "Evaluating semantic accuracy of data-to-text generation with natural language inference," *arXiv preprint arXiv:2011.10819*, 2020.

[3] C. van der Lee, E. Krahmer, and S. Wubben, "Automated learning of templates for data-to-text generation: Comparing rule-based, statistical and neural methods," in *Proceedings of the 11th International Conference on Natural Language Generation*, 2018, pp. 35–45.

[4] K. Knight and V. Hatzivassiloglou, "Two-level, many-paths generation," *arXiv preprint cmp-lg/9506010*, 1995.

[5] Z. Ji, N. Lee, R. Frieske, *et al.*, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[6] Z. Xu, S. Jain, and M. Kankanhalli, "Hallucination is inevitable: An innate limitation of large language models," *arXiv preprint arXiv:2401.11817*, 2024.

[7] H. Harkous, I. Groves, and A. Saffari, "Have your text and use it too! end-to-end neural data-to-text generation with semantic fidelity," *arXiv preprint arXiv:2004.06577*, 2020.

[8] E. Goldberg, N. Driedger, and R. I. Kittredge, "Using natural-language processing to produce weather forecasts," *IEEE Expert*, vol. 9, no. 2, pp. 45–53, 1994.

[9] T.-H. Wen, M. Gasic, N. Mrksic, P.-H. Su, D. Vandyke, and S. Young, "Semantically conditioned lstm-based natural language generation for spoken dialogue systems," *arXiv preprint arXiv:1508.01745*, 2015.

[10] G. Goodall, *The Cambridge handbook of experimental syntax*. Cambridge University Press, 2021.

[11] J. Myers, *Acceptability judgments*, Sep. 2017. DOI: 10.1093/acrefore/9780199384655.013.333. [Online]. Available: https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-333.

[12] N. Chomsky, *Aspects of the Theory of Syntax*. MIT press, 2014.

[13] J. H. Lau, A. Clark, and S. Lappin, "Measuring gradience in speakers' grammaticality judgements," in *Proceedings of the annual meeting of the cognitive science society*, vol. 36, 2014.

[14] J. Sprouse, C. T. Schütze, and D. Almeida, "A comparison of informal and formal acceptability judgments using a random sample

from linguistic inquiry 2001–2010," *Lingua*, vol. 134, pp. 219–248, 2013.

[15] A. Clark, G. Giorgolo, and S. Lappin, "Statistical representation of grammaticality judgements: The limits of n-gram models," in *Proceedings of the fourth annual workshop on cognitive modeling and computational linguistics (CMCL)*, 2013, pp. 28–36.

[16] J. H. Lau, A. Clark, and S. Lappin, "Unsupervised prediction of acceptability judgements," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 1618–1628.

[17] J. H. Lau, A. Clark, and S. Lappin, "Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge," *Cognitive science*, vol. 41, no. 5, pp. 1202–1241, 2017.

[18] J.-P. Bernardy, S. Lappin, and J. H. Lau, "The influence of context on sentence acceptability judgements," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 456–461.

[19] A. Ek, J.-P. Bernardy, and S. Lappin, "Language modeling with syntactic and semantic representation for sentence acceptability predictions," in *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 2019, pp. 76–85.

[20] J. H. Lau, C. Armendariz, S. Lappin, M. Purver, and C. Shu, "How furiously can colorless green ideas sleep? sentence acceptability in context," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 296–310, 2020.

[21] S. C. Wallbridge, C. Lai, and P. Bell, "Investigating perception of spoken dialogue acceptability through surprisal.," in *INTERSPEECH*, 2022, pp. 4506–4510.

[22] I. Groves, Y. Tian, and I. Douratsos, "Treat the system like a human student: Automatic naturalness evaluation of generated text without reference texts," in *Proceedings of the 11th International Conference on Natural Language Generation*, 2018, pp. 109–118.

[23] A. Warstadt, A. Singh, and S. R. Bowman, "Neural network acceptability judgments," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 625–641, 2019.

[24] A. Warstadt, A. Parrish, H. Liu, *et al.*, "Blimp: The benchmark of linguistic minimal pairs for english," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 377–392, 2020.

[25] B. Jacquet, A. Hullin, J. Baratgin, and F. Jamet, "The impact of the gricean maxims of quality, quantity and manner in chatbots," in *2019 international conference on information and digital technologies (idt)*, IEEE, 2019, pp. 180–189.

[26] Y. Nam, H. Chung, and U. Hong, "Language artificial intelligences' communicative performance quantified through the gricean con-

versation theory," *Cyberpsychology, Behavior, and Social Networking*, vol. 26, no. 12, pp. 919–923, 2023.

[27] S. Mehri and M. Eskenazi, "Unsupervised evaluation of interactive dialog with dialogpt," *arXiv preprint arXiv:2006.12719*, 2020.

[28] K. Song, Y. Leng, X. Tan, Y. Zou, T. Qin, and D. Li, "Transcormer: Transformer for sentence scoring with sliding language modeling," *Advances in Neural Information Processing Systems*, vol. 35, pp. 11 160–11 174, 2022.

[29] J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff, "Masked language model scoring," *arXiv preprint arXiv:1910.14659*, 2019.

[30] J. Shin, Y. Lee, and K. Jung, "Effective sentence scoring method using bidirectional language model for speech recognition," *arXiv preprint arXiv:1905.06655*, 2019.

[31] A. Wang and K. Cho, "Bert has a mouth, and it must speak: Bert as a markov random field language model," *arXiv preprint arXiv:1902.04094*, 2019.

[32] W. Yuan, G. Neubig, and P. Liu, "Bartscore: Evaluating generated text as text generation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27 263–27 277, 2021.

[33] Y. Wu, M. Schuster, Z. Chen, *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[34] Z. Xie, M. Li, T. Cohn, and J. Lau, "Deltascore: Fine-grained story evaluation with perturbations," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 5317–5331.

[35] J. Fu, S.-K. Ng, Z. Jiang, and P. Liu, "Gptscore: Evaluate as you desire," *arXiv preprint arXiv:2302.04166*, 2023.

[36] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," *arXiv preprint arXiv:1510.03055*, 2015.

[37] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," *arXiv preprint arXiv:1904.09751*, 2019.

[38] H. Zhang, D. Duckworth, D. Ippolito, and A. Neelakantan, "Trading off diversity and quality in natural language generation," *arXiv preprint arXiv:2004.10450*, 2020.

[39] Y. Zhang, S. Sun, M. Galley, *et al.*, "DIALOGPT : Large-scale generative pre-training for conversational response generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, A. Celikyilmaz and T.-H. Wen, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 270–278. DOI: 10.18653/v1/2020.acl-demos.30. [Online]. Available: https://aclanthology.org/2020.acl-demos.30.

[40] X. Gao, Y. Zhang, M. Galley, C. Brockett, and B. Dolan, "Dialogue response ranking training with large-scale human feedback data," *arXiv preprint arXiv:2009.06978*, 2020.

[41] A. Liusie, P. Manakul, and M. Gales, "Llm comparative assessment: Zero-shot nlg evaluation through pairwise comparisons using large language models," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 139–151.

[42] A. Liusie, V. Raina, Y. Fathullah, and M. Gales, "Efficient llm comparative assessment: A product of experts framework for pairwise comparisons," *arXiv preprint arXiv:2405.05894*, 2024.

[43] J. Kaplan, S. McCandlish, T. Henighan, *et al.*, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[44] J. W. Rae, S. Borgeaud, T. Cai, *et al.*, "Scaling language models: Methods, analysis & insights from training gopher," *arXiv preprint arXiv:2112.11446*, 2021.

[45] K. Mahowald, A. A. Ivanova, I. A. Blank, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko, "Dissociating language and thought in large language models," *Trends in Cognitive Sciences*, 2024.

[46] Y. Zhang, A. Warstadt, H.-S. Li, and S. R. Bowman, "When do you need billions of words of pretraining data?" *arXiv preprint arXiv:2011.04946*, 2020.

[47] P. A. Huebner, E. Sulem, F. Cynthia, and D. Roth, "Babyberta: Learning more grammar with small-scale child-directed language," in *Proceedings of the 25th conference on computational natural language learning*, 2021, pp. 624–646.

[48] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019. [Online]. Available: `https://api.semanticscholar.org/CorpusID:160025533`.

[49] Y. Liu, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[50] C. Raffel, N. Shazeer, A. Roberts, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[51] S. Zhang, S. Roller, N. Goyal, *et al.*, "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.

[52] H. W. Chung, L. Hou, S. Longpre, *et al.*, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.

[53] Y. Fu, H. Peng, L. Ou, A. Sabharwal, and T. Khot, "Specializing smaller language models towards multi-step reasoning," in *International Conference on Machine Learning*, PMLR, 2023, pp. 10 421–10 430.

[54] A. Pauls and D. Klein, "Large-scale syntactic language modeling with treelets," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, H. Li, C.-Y. Lin, M. Osborne, G. G. Lee, and J. C. Park, Eds., Jeju Island, Korea:

Association for Computational Linguistics, Jul. 2012, pp. 959–968. [Online]. Available: https://aclanthology.org/P12-1101.