Semantic Accuracy in Natural Language Generation
Thesis Proposal

Patrícia Schmidtová
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic
schmidtova@ufal.mff.cuni.cz

Abstract
State-of-the-art generative large language models (LLMs) are capable of achieving impressive fluency in natural language generation (NLG) tasks, however, they still struggle with producing outputs that are semantically accurate with respect to their inputs. This poses a pressing issue since LLMs are no longer only used by researchers who understand how they work, but also by millions of users who tend to anthropomorphize this technology.

In the proposed thesis, we present our planned methods of examining which factors further contribute to the prevalence of semantic errors. By trying to understand how factors such as prompt design affect the models’ performance, we aim to identify ways in which the models can be modified to achieve better results. At the same time, evaluation in natural language generation remains far from standardized. Therefore, before examining the models’ performance at scale, we need to propose an evaluation schema. For this reason, we will dedicate significant attention to the challenge of how to properly evaluate whether an output is semantically accurate given the input. We intend to propose new metrics inspired by those used in machine translation quality estimation as well as to participate in community efforts to standardize evaluation in NLG.

1 Introduction
The introduction of the Transformer architecture (Vaswani et al., 2017) irreversibly changed the research landscape in natural language processing. Moreover, in the past couple of years, large pre-trained language models (LLMs) have managed to permeate into the hands and minds of millions of users worldwide (Ouyang et al., 2022; Jiang et al., 2023; Scao and et al., 2023). With a growing public interest in natural language generation (NLG) and dialogue systems, it is essential to thoroughly research LLMs’ reliability. If a human does not know the answer to a question, the socially acceptable behavior is to say ‘I do not know’ instead of making up a plausibly sounding lie. This is how many users expect intelligent systems to behave, which is further reinforced by practices such as chain-of-thought (Wei et al., 2023), where the LLM is prompted to ‘think step by step’, which may suggest to a layperson that a statistical model is capable of human-like thought. Cheng et al. (2024) show that there is a rising trend of anthropomorphism (attributing human-like characteristics to non-human entities) of LLMs in the media and even NLP research papers. If an LLM provides unreliable information, it can lead to distrust, or in a worse scenario, even to the spread of misinformation. In a notable case this year, an LLM-generated news article featured a quote from Prof. Emily Bender, however, she reported that she had never said those words (Mahadik, 2024).

In this thesis proposal, we will highlight the importance of semantic accuracy in natural language processing (NLP) research. We will look into how to evaluate it as well as how to understand and improve the semantic accuracy of LLM outputs.¹

In the remainder of this section, we will formulate the goals of this thesis (Section 1.1), define semantic accuracy (Section 1.2), specify the tasks in our scope (Section 1.3) and provide an outline of this proposal (Section 1.4).

1.1 Goals
This thesis has two objectives formulated as research questions:

1. How can we determine if a generated text is semantically accurate given its source data?

¹Since large language models have become the gold standard for natural language generation, we will sometimes use the terms NLG system and LLM interchangeably. Furthermore, we also alternate between the terms input vs. source and output vs. target.
2. Which internal or external mechanisms affect the semantic accuracy of an LLM’s output and how can we manipulate them to achieve better accuracy?

1.2 Semantic Accuracy

In the literature, semantic accuracy is frequently denoted by two terms: faithfulness and factuality (Maynez et al., 2020). Faithfulness means adherence to user-provided inputs — i.e. checking if the model correctly performs the task at hand without adding or omitting information. Factuality is the veracity of the text based on all current knowledge. Suppose we have input data defined as a semantic triple (subject-predicate-object): ACL 2024 – location – Thailand. The sentence "ACL 2024 will be held in Bangkok, Thailand" is factual, but not faithful since Bangkok was not mentioned in the input data.

We define semantic accuracy as faithfulness and recognize two types of semantic errors: hallucination (adding unsupported information) and omission (omitting information from the input in the generated text). While hallucination is an issue in every use case and task known to us, including creative text generation (Schmidtová et al., 2022), omission is not always an issue — especially in tasks such as summarization or simplification. For this reason, we put more emphasis on hallucination.

The majority of current approaches focus on factuality (Azaria and Mitchell, 2023; Li et al., 2023b; Lin et al., 2022). Our decision to focus on faithfulness is partially pragmatic — it is a better-defined problem that is easier to grasp within the limited scope of a dissertation thesis. But more importantly, even though inaccurate/unfaithful does not always mean wrong (Maynez et al., 2020), we argue that an NLG system should produce semantically accurate texts to be considered reliable and that research on faithfulness should not be overlooked. Our belief that faithfulness outside of fact-checking (Thorne et al., 2018) is an important research direction is motivated by the following reasons:

- It is important to alert a user of an LLM about the output text deviating from the source data so they do not overlook it and can evaluate the factuality themselves.
- Faithfulness can be more useful in some domains, such as journalism. The factuality of some statements can change (for example "Guenther Steiner is the team principal of the Haas F1 team." is no longer true). Suppose a journalist wants to write an AI-assisted article about the new team principal. They will do so while the information is fresh - i.e. before the information is overwritten in any databases or vector stores. The journalist will prefer a faithful system that uses the supplied new information over a factual one that will attempt to use outdated information stored in a database.

- In some downstream tasks, such as task-oriented dialogue systems or retrieval-augmented generation (Lewis et al., 2020a), we want full control of the output to maintain a high level of reliability. This is especially important if a system needs to keep track of what was already been communicated to the user.

1.3 Tasks in Scope

Large language models are capable of processing a wide range of NLG tasks, for example, data-to-text generation, summarization, or paraphrasing. We intend to cover as many of these tasks as possible. We acknowledge this might not be possible for all of the proposed research directions we plan to undertake. In such cases, we will focus on data-to-text generation and summarization due to their popularity and usefulness, as well as our familiarity with the tasks.

We note that multi-modal tasks such as image caption generation are also traditionally considered NLG tasks. However, in this thesis, we consider any multi-modal inputs and outputs out of scope and focus on purely textual data.

Furthermore, we will focus on tasks where all of the necessary data is given to the model as input and the model does not need to access any knowledge bases on its own. For simplification, we assume that the step of retrieving information from a vector store in retrieval-augmented generation (Lewis et al., 2020a) or accessing a database were already performed for us and we simply focus on composing a faithful output.

1.4 Outline

In Section 2, we will cover the theoretical foundations necessary for the proper understanding of this thesis. In Section 3, we intend to address the first research question. We will discuss the current
state of NLG evaluation in general and also delve into specifics of evaluating semantic accuracy. We will point out the challenges in conducting a proper human evaluation as well as the divided state of automatic evaluation in NLG. We will also present our published, ongoing, and future work in the realm of NLG evaluation.

In Section 4, we address the second research question. We will introduce the current strategies in the literature on how to improve semantic accuracy. We will also discuss how interpretability techniques are used to get a better understanding of the inner workings of large language models. Then we will combine these two approaches and present how we intend to gain a better understanding of factors that influence semantic errors and sketch strategies on how to leverage this knowledge in the making of more reliable models or best practice recommendations for the community. The thesis proposal is concluded in Section 5.

2 Theoretical Foundations

2.1 Transformer Architecture

Since the state-of-the-art language models are built using the Transformer architecture (Vaswani et al., 2017), we will dedicate this subsequence to a brief introduction to them.

The Transformer is generally composed of a stack of encoders and decoders, however, it is common for language models to have decoder-only architectures. Before the processing, a positional embedding is added to the token embeddings to add information about word order in the sequence. In the encoders, self-attention is applied to the inputs and followed by a feedforward neural network. The decoders have the self-attention layer as well as the feedforward neural network but have an additional layer of encoder-decoder attention between them. Each of the layers involves an additional residual connection (He et al., 2015) and layer normalization (Ba et al., 2016). Finally, to produce the next word, the decoder output is connected to a fully connected linear layer followed by a softmax layer. The softmax layer contains the next-token likelihood. There are various decoding strategies used for selecting the next token such as nucleus sampling (Ravfogel et al., 2023) or beam search (Meister et al., 2020)

In our interpretability experiments, we will focus on self-attention when exploring any attention-related techniques (Voita et al., 2019; Behnke and Heafield, 2020) and the feedforward layers when applying probing (described in Section 2.2).

Training Transformer models are usually pre-trained on vast unlabeled corpora using objectives such as masked language modeling (Devlin et al., 2019) or next-word prediction (Radford et al., 2019).

Fine-tuning can be used to adapt a pre-trained model to a new domain or task. Notably, low-rank adaptation (Hu et al., 2021; Dettmers et al., 2023) is a form of parameter-efficient fine-tuning that lowers the requirements on computational power and allows researchers to fine-tune large models. We intend to use this technique during the work on the proposed thesis.

Large language models can also be fine-tuned to follow instructions (Ouyang et al., 2022) which allows them to be trained for multiple tasks at once. Reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) is used to teach the LLMs how to produce text that humans will like.

Improvements Since the Transformer architecture was published in 2017, there have been many improvements to increase its performance, decrease the computational power needed to run them and prolong the sequences they are able to process. Mixture of Experts (Du et al., 2022; Jiang et al., 2024) originally introduced by (Jacobs et al., 1991) shows promising increases in performance. Their effect on semantic accuracy still needs to be explored.

Quantization (Jacob et al., 2017) helps especially academic researchers by lowering the amount of computational power needed to run a large model. Flash attention (Dao, 2023) allows for the processing of longer sequences by decreasing the space complexity of the attention mechanism. The last two improvements are important, however, they do not directly affect semantic accuracy.

2.2 Probing

Probing (Ettinger et al., 2016; Adi et al., 2017; Conneau et al., 2018) is a technique for examining the information stored in a model’s hidden layers. It is typically performed in several steps:

1. Training a model or using a pre-trained one.
2. Formulating a hypothesis, e.g.: “A model trained for masked language modeling learns
3.1 Human Evaluation

Human evaluation is currently the most reliable tool for evaluating natural language systems – when performed properly (Howcroft et al., 2020). In addition to scoring the overall impression of a system output, it also allows us to collect fine-grained feedback on various qualities of the generated text such as informativeness, fluency, or semantic accuracy. For tasks such as data-to-text generation or summarization, human annotators do not need a gold reference and can judge the quality of a generated text given the source data. Moreover, human evaluation is flexible – if needed, we can get more information than just numbers, we can ask annotators to tag error spans, rank the outputs of various systems, or share their thoughts in an unstructured way which can lead to interesting insights on the systems’ performance.

Challenges However, there are many pitfalls that can degrade the quality of human evaluation. When designing the evaluation process, it is crucial to properly name and define the qualities to be assessed by the annotators by referring to the best practices recommended by Howcroft et al. (2020). However, even when taking proper precautions in designing the evaluation, the results can be biased: Hosking et al. (2024) show that humans are often influenced by fluency even when judging unrelated qualities such as factuality.

Reproducibility Another downside of human evaluation is its limited reproducibility (Belz et al., 2023a,b). This poses an issue if we want to objectively compare the outputs of our system to a baseline from previous work. In such cases, the baseline must be re-evaluated within the same setting as the inspected system. The issue of reproducibility is now getting attention thanks to the ReproHum project (Belz et al., 2022) leading to recommendations and best practices for ensuring the best possible reproducibility of experiments that involve human evaluation. We (Lango et al., 2024)\(^2\) have participated in the collective effort as well.

Reliability During our reproducibility experiment (Lango et al., 2024), we also experienced issues with the reliability of crowdworkers. Some crowdworkers completed the tasks in an impossibly short amount of time (reading about 2000 words in less than 10 minutes). Another issue we noticed is that even some of the crowdworkers performed properly (Howcroft et al., 2020).

\(^2\)The author of this proposal participated in the setup of the experiment, creating the forms for gathering the annotations as well as analyzing the results.
ers who completed the task in a reasonable amount of time occasionally assigned the worst scores to the gold reference. This behavior was inconsistent with the majority of crowdworkers, suggesting that the problematic crowdworkers selected answers quickly at random and then waited before submitting responses to avoid suspicion. We experienced no such issues with in-house annotators. To make matters worse, human evaluation has also become less reliable with the introduction of large language models to the public – Veselovsky et al. (2023) estimate that 33-46% of crowdworkers use LLMs to complete the tasks for them. Filtering out low-quality annotators is also possible using techniques such as attention checks or hidden instructions for large language models (Hudecek et al., 2023).

Experts Another way to combat the above-mentioned issues is to use experts for the annotations. There are two obstacles to doing so: the experts are scarce to find and tend to be considerably more expensive than crowdworkers or even in-house annotators. Therefore, we will reserve them for tasks where domain knowledge is necessary or whenever we have reasons to assume layperson annotators would be biased by some other aspects of the evaluated texts rather than the qualities we are trying to measure.

Best Practices When conducting a human evaluation for our experiments, we will thoroughly define and explain the evaluated qualities to avoid confusion, implement attention checks, and employ filtering techniques. In general, we will adhere to the best practices shared by the NLG community (Thomson et al., 2024) to obtain reliable human evaluation results.

3.2 Automatic Evaluation

Automatic evaluation tends to be used as a proxy for human evaluation whenever human evaluation would be too expensive. In our work, we intend to focus on and contribute to automatically evaluating NLG for the following reasons:

1. It is quick and almost costless, therefore it can be used repeatedly as a part of a development cycle.

2. Metrics with good correlation with human judgment allow us to extrapolate the result of a human evaluation on a smaller sample, thus reducing the costs needed.

3. Automatic evaluation can be used to observe patterns that would go unnoticed by the human eye, such as average sentence length.

4. When using open-source data and trustworthy implementations of the metrics (Post, 2018; Grusky, 2023), automatic evaluation is inherently reproducible.

Word-Overlap Metrics Natural language generation is dominated by reference-based automatic metrics originally intended for machine translation, such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2020). In addition, ROUGE (Lin, 2004) is also used for various NLG tasks even though it was originally intended for summarization. We can see this trend in Figure 1 which observes the use of metrics at INLG 2023 and we further discuss it as a part of our ongoing research efforts in Section 3.5.

These metrics are the most commonly used, despite some researchers warning against relying too much on these metrics (Reiter and Belz, 2009) and showing limited correlation with human judgment (Novikova et al., 2017; Reiter, 2018; Akter et al., 2022). It is certainly partially caused by tradition and the momentum these metrics have gained. However, we hypothesize that this is difficult to change due to the diversity of input types and formats in NLG tasks: for data-to-text, we can see semantic triples, tables, knowledge graphs, jsons, etc., for text-to-text we have varying lengths of input text that do not always correspond to the output lengths (such as summarization), and some tasks even have multi-modal inputs, such as image or video captioning. For all these tasks, it is extremely difficult to design a metric that could take all of the different types of input into account and it is simply easier to compare the output text to a human-written reference.

Trainable Metrics Trainable metrics such as BERTScore (Zhang et al., 2020), Bleurt (Sellam et al., 2020), or PARENT (Dhingra et al., 2019) can be used to evaluate the semantic accuracy of a given text. All of these are reference-based metrics – they compare the evaluated texts to a given human-written reference solution instead of comparing the texts with the source data. This means that these
methods cannot be applied to freshly mined examples where a reference is not available (Kasner and Dušek, 2024). Furthermore, such metrics cannot explain why a text received a high or a low score – they can only measure the proximity to a reference.

The majority of metrics for evaluating the semantic accuracy of generated text utilize models pre-trained for the task of Natural Language Inference (NLI). NLI is a task where we classify the relation between a hypothesis and a premise into one of three classes: entailment, contradiction, and neutral. When evaluating semantic accuracy, NLI is usually applied to measure the entailment between the generated text and the input data or a reference. Such metrics include NUBIA (Kane et al., 2020), MENLI (Chen and Eger, 2023), and approaches presented by Maynez et al. (2020) and Dušek and Kasner (2020).

The advantage of NLI-based metrics is that they generally do not need a reference (except for NUBIA) and can handle lexical diversity. However, they are not easily interpretable by the user, because they natively do not show where the inaccuracies occur within the text. Furthermore, this method is not equipped to deal with structured inputs, because the NLI model is trained on plain text.

**Benchmarks** One of the most popular ways to evaluate and compare large language models is using benchmarks (Zellers et al., 2019; Nie et al., 2020; Lin et al., 2022) and leaderboards (Beeching et al., 2023). There is also a benchmark for hallucination recognition (Li et al., 2023a) that has been generated using ChatGPT.

We consider this trend out of the scope of this thesis for two reasons. First, research in this area is overly saturated, so we rather focus on a more qualitative approach when evaluating NLG systems. Second, there is no guarantee that even the open-source models such as Llama2 (Touvron et al., 2023) or Mistral (Jiang et al., 2023) have not been trained on benchmark data as their training datasets are proprietary and not accessible to researchers. For more discussion on the topic of data contamination and models being trained on benchmark data, please see Section 3.4.

**LLMs as Judges** Another interesting emerging practice is the use of large language models as judges when performing an evaluation. GEMBA (Kocmi and Federmann, 2023b), one of the earliest efforts in this direction, uses various OpenAI models for the evaluation of machine translation. Kocmi and Federmann (2023b) discovered that GPT-3.5-turbo and GPT-4 are in fact state-of-the-art evaluators of machine translation quality.

Within natural language generation, the results reported by previous work are mixed. Notably, Fu et al. (2023) show that large language models are currently not capable of making numerical judgments on a Likert scale that would correlate well with human ratings. However, Kasner and Dušek (2024) show a promising way to evaluate semantic accuracy errors as a sequence tagging problem. The inter-annotator agreement between human annotators and a GPT-4-based metric is comparable to the agreement solely between humans.

Chan et al. (2023) introduce an interesting approach where three large language models discuss the evaluation amongst themselves to reach a common judgment. Each of the three large language models receives a different set of instructions, modeling various human personalities and preferences.

An important issue to consider when using LLMs as judges is their bias towards their own outputs or the outputs of models within the same model family (Koo et al., 2023). This challenge can be addressed by reserving a model for evaluation rather than using it for the experiments. Practically, we intend to focus our research on open-source models such as Mistral (Jiang et al., 2023), Mistral (Jiang et al., 2024), or Llama2 (Touvron et al., 2023) for conducting the experiments. Then we can use even a closed-source model such as GPT3.5-turbo or GPT-4 for a quick evaluation. We intend to use this technique as a form of preliminary evaluation since it is non-reproducible by design of closed-source models.

We also have previous work trying this approach (Plátek et al., 2023) discussed in Section 3.4. Moreover, we have plans for future work in this area described in Section 3.6.

### 3.3 Previous Work

We also draw from the knowledge the author gained during the work on her Master’s thesis (Schmidtová, 2022) while working on the TheAItre project (Schmidtová et al., 2022). At that time, the biggest challenges were fluency (especially when generating texts longer than the context window of the model used), repetitiveness, contradiction, and hallucination. This called for the introduction of a
novel unconventional metric to evaluate whether a newly generated line contributes to the story. Contrary to current evaluation methods based on NLI, described earlier in this section, we used neutrality instead of entailment to measure repetitiveness and contradiction. It is interesting to observe the shift to the current models where fluency or repetitiveness are now no longer an issue, however, hallucination still remains an issue (Kasner and Dušek, 2024).

3.4 Published Work

Data Contamination Our most prominent published work in the area of NLG evaluation warns against the problem of data contamination – the presence of test data in models’ pre-training or fine-tuning data (Balloccu et al., 2024). While previous work focuses on data contamination in the pre-training data (Sainz et al., 2023; Golchin and Surdeanu, 2024), we have identified a new issue of researchers unknowingly leaking benchmark test data to OpenAI models by evaluating them using their web interface. In total, researchers have leaked ~4.7M samples coming from 263 benchmarks that cover every task imaginable in NLP.

Indirect data leaking is even more worrying because all of the data is received with detailed instructions on how to process the data, resulting in a novel gold-standard dataset for instruction tuning on benchmark data.

LLMs as Judges Our other published work in NLG evaluation examines various qualities of dialogue responses paired with a given dialogue context on a Likert scale (Plátek et al., 2023). The examined qualities were Appropriateness, Content Richness, Grammatical Correctness, and Relevance.

In this work, we found that GPT-NeoX-20B (Black et al., 2022), OPT-30B (Zhang et al., 2022), and TK-Instruct-11B (Wang et al., 2022) have a poor Spearman rank correlation with human judgment. This work also explored the performance of GPT-3.5-turbo and Llama2-7b-chat (Touvron et al., 2023), which had a moderate Spearman rank correlation with humans, nearing state-of-the-art in the task.

3.5 Ongoing Work

Reproducing Human Evaluation The first ongoing work we will introduce in this proposal is our paper on reproducing human evaluation in dialogue summarization (Lango et al., 2024), submitted to the HumEval workshop collocated with LREC-Coling. We reproduced the human evaluation done by Feng et al. (2021). We found that all of the system outputs tend to be judged more harshly by our annotators than originally reported in the study. This could be due to two reasons. First, the original annotation was done by the PhD students working in the same lab as the authors of the reproduced paper leading to a possible positive bias. Second, the paper came out before ChatGPT was released, which could have influenced the human perception of what a good system output looks like.

We also found that the inter-annotator agreement was significantly weaker during our reproduction than in the original study. This can be attributed to the fact that our annotators come from more diverse backgrounds (different native languages, locations, and educational backgrounds).

As mentioned in Section 3.1, we have run into issues with crowdworkers trying to cheat the system. The annotation itself was performed via Google Forms and there were no attention checks in place to conform with the ReproHum project’s guidelines. We have observed that some annotators completed the task quickly and produced results of visibly low quality. Moreover, some crowdworkers had annotations with a relatively high agreement with the remaining annotators despite fulfilling the task in a very short amount of time. We hypothesize that these annotators cheated the system by skipping the dialogue transcript and simply rating the summaries based on overall impression or fluency (these were not included in the qualities we were interested in).

This work was a very beneficial experience to realize that not all human evaluations are equally reliable. In our future work, we will be much more careful in designing the studies and attention checks to make our evaluation results more trustworthy.

Systematic Review of NLG Evaluation Our next ongoing work is an international team effort planned to be a submission for INLG 2024. In
this work, we systematically review papers from INLG 2023 and the Generation track of ACL 2023 to obtain information about what automatic metrics are used by the NLG community, the motivations to use those metrics, and the reported synergy between human and automatic evaluation. There are already lengthy surveys on the topic detailing which metrics exist for NLG and how they are supposed to be used (Sai et al., 2022; Celikyilmaz et al., 2020), however, they overlook the important aspect of what the community uses and why.

In Figure 1 we include preliminary results on which metrics were used in the 37 papers published at INLG 2023 excluding newly introduced metrics. We can see that there is no consensus in the community about which metrics should be reported outside of the word-overlap-based metrics. Moreover, 4 new metrics were introduced within the same venue.

Our objective with this work is to uncover evaluation patterns within NLG tasks, encourage wider adoption of the uncovered good practices, and motivate the community to try metrics with a better human correlation than BLEU or ROUGE. We note that a similar review exists for human evaluation (Howcroft et al., 2020) and it has been very influential in the field.

### 3.6 Future Work

#### MT-Inspired Approaches

As a follow-up work after the systematic review of NLG evaluation (Section 3.5), we intend to explore the relations between automatic metrics and human judgment to uncover which magnitudes of score differences in automatic evaluation results are significant. This plan is inspired by the work of Kocmi et al. (2024) in the field of machine translation, where they examine how increases in automatic metric scores correlate with increases in human-given scores.

We also intend to explore adapting machine translation quality estimation metrics into NLG tasks. Our initial experiments show that using XComet-XL (Guerreiro et al., 2023) out-of-the-box shows a poor correlation with semantic accuracy on tasks such as paraphrasing or definition modeling. On the other hand, CometKiwi23 (Rei et al., 2023) is showing a promising Spearman rank correlation with human judgment without any modifications. We will experiment with adapting these metrics to enable them to evaluate a wide variety of NLG tasks.

#### Using LLMs for Evaluation

As discussed in Section 3.2, there are already approaches assessing the reliability of using large language models as judges. However, the majority of existing work focuses on a single LLM, task, strategy on how to obtain the LLM’s judgment (numerical score, sequence tagging, ranking, etc.), and a single level of expertise of the human baseline. Furthermore, works attempting to score the examples on a Likert scale (Fu et al., 2023; Plátek et al., 2023) tend to show a weaker correlation with human judgment than works that approach the problem as sequence tagging of error spans (Kocmi and Federmann, 2023a; Kasner and Dušek, 2024). Since large language models have a limited understanding of numbers (Zhu et al., 2024), we will also explore how their performance as judges will change if we swap the numeric Likert scale for a variant that uses adjectives instead.

An additional factor to consider is that crowd-workers are the most frequently employed human baseline which is not the level of performance we should strive for. Instead, we should also explore the correlation to experts whose judgments are closer to the gold standard. Therefore, we intend to report the correlation with both classes of annotators to accurately assess the LLM’s capabilities.

By exploring more combinations of models, tasks, strategies, and human expertise, we aim to provide a set of fine-grained instructions on how to achieve the most reliable judgments. We reiterate that in our future research, we intend to use LLMs as judges to obtain frequent feedback dur-
ing our experiments and will prefer to report the results of a human evaluation when publishing our results. This research direction will be explored with colleagues from Edinburgh Napier University.

Alignment-Based Methods for Semantic Accuracy Evaluation As discussed in our previous work (Schmidtova, 2023), we still intend to contribute our own methods for evaluating semantic accuracy. Our main research direction in this area is also inspired by machine translation and will focus on leveraging source-target alignments (Dou and Neubig, 2021; Molfese et al., 2024) and using contrastive decoding (Sennrich et al., 2024). The advantages of these approaches are a high level of interpretability, robustness to lexical diversity, and being usable without a reference.

When we say “methods”, we refer to obtaining the source-target alignments rather than a single number. We choose to focus on the alignments because they are immensely helpful for explaining the relation between the source and the target on the sample level. Any numerical measurements will be directly related to the alignments. To create a metric with a numerical output out of this method, we can for example use a length-normalized proportion of the target that can be aligned to the source. We will also experiment with following more numerical parameters, such as the number of alignments made, or even the distances amongst representations.

The main technical challenge of adapting these methods to a wider range of NLG tasks is to ensure we can obtain reliable embeddings for many various formats of input data, such as semantic triples, knowledge graphs, tables, or JSON files. (Kasner and Dušek, 2024) showed that large language models are capable of handling a variety of input formats, including the ones mentioned above. We intend to explore using representations from the hidden layers or various open-source LLMs with accessible weights as embeddings in the alignment algorithms.

The second challenge is the adoption by other researchers than its authors and we intend to undertake several steps to increase its chances. First, we acknowledge that the NLG community scarcely uses referenceless metrics or methods. In Figure 1, we can see that some works report statistics that only rely on the output, such as average length or perplexity. However, there are only three referenceless metrics that take the input data into account: NLI, GPT-3.5, and GPT-4. Therefore, to establish the credibility of our method, we have to ensure a thorough human annotation and demonstrate a high correlation with human judgment.

Second, we can notice that all four most frequently used metrics are available via libraries such as HuggingFace or as PyPI packages. This suggests that the ease of access is important. Moreover, it is not always straightforward to get the code from GitHub running (Platek et al., 2023). It is therefore our intention to distribute our finished work as a PyPI package to limit any technical barriers that would prevent the adoption of our methods.

4 Improving and Understanding Semantic Accuracy in NLG

In this Section, we will discuss approaches for improving semantic accuracy. In Section 4.1, we will discuss the related work that attempts to improve semantic accuracy via engineering. We specifically dedicate Section 4.2 to approaches based on interpretability, which is a direction we would like to explore in our future work. In Section 4.3, we reveal our ongoing efforts to investigate the effect of grammatical errors on an LLM’s performance. Finally, in Section 4.4, we sketch our plans for interpretability-motivated improvements of semantic accuracy.

4.1 Engineering

Current work focuses on increasing reliability in limited domains (González Corbelle et al., 2022; Nie et al., 2019) or using external knowledge bases (Dziri et al., 2022) or forming the responses after performing information retrieval, i.e. retrieval-augmented generation (Lewis et al., 2020b). While it is not in the scope of this thesis to directly access external databases or knowledge bases and perform information retrieval, we believe they are essential to reliable large language models, because, unlike LLMs, external sources are easy to maintain and keep up-to-date. Therefore, we also keep this use case in mind and aim to make our methods compatible with it by considering that additional information can be added to the LLM input.

A recently observed, but as of now unexplained, phenomenon is the fact that the exact wording of instructions given to an LLM can have a dramatic impact on whether the LLM will hallucinate or not (Zhang et al., 2023). Specifically, one of the ways to decrease the prevalence of hallucinations a little
is to simply instruct the LLM explicitly to stick to facts provided on the input (Axelsson and Skantze, 2023). We are not aware of a work that would examine why this is the case. Therefore, we aim to investigate this phenomenon.

4.2 Interpretability

In natural language processing, probing was previously used to extract part of speech (Conneau et al., 2018), syntactic properties (Hewitt and Manning, 2019), and discourse structures (Huber and Carenini, 2022) from hidden layers of neural networks, suggesting that neural networks are capable of observing various linguistic properties without being taught they even exist. More recently, probing was used to seek out and modify facts stored in LLMs’ trained weights (Meng et al., 2022). Azaria and Mitchell (2023) use probing to classify if an LLM believes that a statement supplied by the user on the input is true. Li et al. (2023c) use a similar approach to identify which Transformer attention heads (Vaswani et al., 2017) are likely to produce untrue output and shift activations away from them.

The research community is divided in the question of whether attention can serve as a reliable explanation (Wiegreffe and Pinter, 2019) or not (Jain and Wallace, 2019). We note that we will consider the exploration of attention as a potential tool in our interpretability research, with no specific expectations on the outcomes.

4.3 Ongoing Work

We are currently exploring the effect of grammatical errors in the prompts on the performance of large language models in various tasks. We have observed that some papers draw strong conclusions about an LLM’s performance based on prompts that are objectively grammatically flawed, such as omitting the subject, wrong usage of verb tenses, or mixing up prepositions. There is some research on how the wording of the prompt affects the LLM’s performance (Leidinger et al., 2023). However, we argue that grammatical correctness is a more objective measure and an important aspect of the prompt design that should not be overlooked.

Based on the literature that examines the typical errors of L2 English speakers with various native languages (Hawkins and Buttery, 2008; Díaz-Negrillo and Valera, 2010; Mizumoto et al., 2012; Can, 2018) we construct 5 classes of errors: spelling, word grammar, phrase grammar, clause grammar, and word choice. For each of the analyzed tasks, we adopt prompts used in previous work and apply grammatical errors from each class. We then examine the effect of each of the grammatical error classes on the LLM’s performance. We are in the process of evaluating the results of the machine translation task and planning to explore summarization, sentiment analysis, and grammatical error correction.

We anticipate that our findings will be interesting to the community of researchers and active users of large language models regardless of the outcome. In case we find that LLMs are robust to grammatical errors in the prompts, it is good news for all L2 English speakers interacting with LLMs. If we discover that grammatical errors negatively impact the LLM’s performance, we will point out a previously unseen issue with a relatively simple solution: first using the LLM to correct the grammar of the prompt.

4.4 Future Work

LLM Meaning Representations

As a follow-up to our work on evaluation through alignments described in Section 3.6, we plan to explore the embedding space of NLG inputs and outputs. Inspired by work in multilingual embedding space (Sherborne et al., 2023), we will explore if we can decrease the distances between source and target representations through a fine-tuning objective. Furthermore, we will measure the effect of this change on the model’s overall semantic accuracy.

Effects of Prompt Wording on Semantic Accuracy

We also intend to explore the finding of Axelsson and Skantze (2023) that LLMs hallucinate less when they are instructed to only use the user-provided data by using probing. The findings of Azaria and Mitchell (2023) and Meng et al. (2022) show that large language models store information in their weights in ways that can be accessed and updated. This leads us to believe we could be able to observe differences in activation patterns when using different prompts. Once observed, we can either update the weights using LoRA parameter efficient fine-tuning (Hu et al., 2021) or draw inspiration from EasyEdit (Wang et al., 2024), a framework for seeking and replacing knowledge stored in the weights of large language models, and proceed with a targeted weight update.

An ideal outcome of such work would be a model that is more robust to various prompt wordings and always assumes it was asked to only use
the provided information. In case we would not be able to achieve such an outcome, we also believe there is value in reporting the found patterns and mechanisms along with a set of best practices on which of the examined strategies work better.

Challenges and Limitations We do not expect we will be able to solve the issue of hallucination within this thesis. Furthermore, we remain skeptical about whether the issue can be fully solved (Kalai and Vempala, 2024). Therefore, instead of trying to chase state-of-the-art results and reporting an X% decrease in hallucinations, we intend to dedicate our research efforts to understanding some of the faithfulness-related processes taking place inside the black box. We also acknowledge that the interpretability of large language models is a quickly evolving topic. This is the reason why our interpretability plans are significantly less elaborate than our next steps in NLG evaluation – there is a high risk of the plans going obsolete within a year.

5 Conclusion

Large language models produce outputs that are very fluent but yet frequently contain semantic errors (Kasner and Dušek, 2024). As described in Section 4, our understanding of the mechanisms causing these errors is limited, especially when we consider external factors such as the prompt design or the format of input data. In this thesis, we intend to employ interpretability techniques to help advance the current understanding of these mechanisms. Furthermore, we outline ways how this knowledge could help us modify models to make them more consistent and reliable.

When making any changes to a model, we naturally want to measure their impact. As we argue in Section 3, this poses an additional challenge because the majority of researchers are divided between evaluation on potentially compromised benchmarks and using machine translation metrics with a poor correlation with human judgment. We do not consider either of these directions reliable enough to use in our own research, therefore, we dedicate the first half of the thesis to defining a better evaluation approach. We intend to contribute to NLG evaluation by mapping the current state of research and encouraging the adoption of good practices. Moreover, we plan to adapt cutting-edge quality-estimation machine translation metrics and approaches to a wider range of NLG tasks. We will also thoroughly investigate the usability of LLMs as evaluators as a quick way to receive feedback on our experiments. Finally, we intend to measure our contributions by a combination of automatic metrics and carefully designed human evaluation.

References


Shahriar Golchin and Mihai Surdeanu. 2024. Time travel in llms: Tracing data contamination in large language models.


Tom Hosking, Phil Blunsom, and Max Bartolo. 2024. Human feedback is not gold standard.


Adam Tauman Kalai and Santosh S. Vempala. 2024. Calibrated language models must hallucinate.


Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality.


Mateusz Lango, Patrícia Schmidtová, Simone Balloccu, and Ondřej Dušek. 2024. Reproducing human evaluation of dialogue summaries: Investigating the impact of educational background and first language.

Submitted to the HumEval Workshop collocated with LREC-Coling.


Karan Mahadik. 2024. AI Invents Quote From Real Person in Article by Bihar News Site: A


Fangwei Zhu, Damai Dai, and Zhifang Sui. 2024. Language models understand numbers, at least partially.