Semantic Accuracy in Natural Language Generation: A 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
With the fast-growing popularity of current large pre-trained language models (LLMs), it is necessary to dedicate efforts to making them more reliable. In this thesis proposal, we aim to improve the reliability of natural language generation systems (NLG) by researching the semantic accuracy of their outputs. We look at this problem from the outside (evaluation) and from the inside (interpretability). We propose a novel method for evaluating semantic accuracy and discuss the importance of working towards a unified and objective benchmark for NLG metrics. We also review interpretability approaches which could help us pinpoint the sources of inaccuracies within the models and explore potential mitigation strategies.

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 year, large pre-trained language models (LLMs) have managed to permeate into the hands and minds of millions of users worldwide (Ouyang et al., 2022; Touvron 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 their 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, and failing to fulfill this expectation can lead to distrust, or in a worse scenario, even to the spread of misinformation.

We believe it is worth trying to propose evaluation schemes that could incentivize institutions and companies to optimize their models for reliability rather than just fluency and impressiveness. The proposed thesis aims to take a step in this direction by investigating semantic accuracy in a data-to-text generation setting. We consider a text semantically accurate if it faithfully represents the underlying input data.

Despite the fact that inaccurate does not always mean wrong (Maynez et al., 2020), i.e. conflicting with our current understanding of the world, we argue that an NLG system should produce semantically accurate texts to be considered reliable. We still consider it important to research NLG through the lens of semantic accuracy, without the intent of explicitly fact-checking (Thorne et al., 2018), for the following reasons:

• It is important to alert the user about the output text deviating from the data so they do not overlook it and can evaluate the factuality themselves.

• The NLG system stores a representation of its training data in its parameters. However, some of that information might be outdated and therefore is no longer accurate. If we supply an NLG system with input data containing updated information, such as the name of a new prime minister, we want this to take precedence over the information learned during training.

• In some use cases, such as in task-oriented dialogue systems, we want full control of the output to maintain a high level of reliability. This is especially important if explicit dialogue state tracking is used so that the system has an accurate representation of what was already communicated to the user.

Thesis Objectives The main objective of this thesis is to answer the question: “How can we make data-to-text Natural Language Generation more reliable?” We hope to achieve this objective by carefully studying NLG systems, namely LLMs, with respect to semantic accuracy, from the outside
(evaluating their outputs) as well as from the inside (inspecting their hidden layers).

It is valuable to quantify how reliable an NLG system is before attempting to increase its reliability to measure the magnitude of such an increase. Furthermore, we hope to provide insights into the operation of NLG systems and the limitations they have. This will allow for a more informed design of NLG systems to tackle the detected problems.

**Thesis Structure** The first part of the thesis, described in Section 2, is dedicated to NLG evaluation. We propose a novel approach for evaluating the semantic accuracy of a generated text given the source data. We also intend to contribute a benchmarking dataset for evaluating NLG metrics focused on semantic accuracy. Thomson and Reiter (2021) have presented such a dataset with high-quality human annotations, however, due to the high costs of human annotation it is very modest in size. Therefore, we share our idea of constructing a larger dataset automatically.

In the second part of the thesis, described in Section 3, we will use interpretability techniques to explore where inaccuracies appear. We aim to then use these insights to learn how to guide the NLG system to produce outputs that are more faithful to the input data.

**Applications** This thesis’ most visible contribution will be in the task of data-to-text natural language generation as it is our primary goal. We anticipate our insights will also be helpful in dialogue systems and retrieval-augmented generation (Lewis et al., 2020). Furthermore, it is our intention to extend the described approaches to abstractive summarization as the task is similar to ours. Finally, we consider omission – omitting some information from the source data in the target text.

**2 Evaluating Semantic Accuracy**

Many aspects of NLG system outputs can be evaluated: fluency, grammatical correctness, acceptability with respect to a context, or similarity to a given reference text, etc (Howcroft et al., 2020). In this thesis, we focus solely on the aspect of semantic accuracy which is far from being solved.

We aspire to evaluate how accurately a target text represents given source data either in a set of semantic triples (subject-predicate-object), a table, or a different structured form. Our proposed output is not only the numeric result of the metric which can be used in a development or research setting, but primarily a set of alignments between the text and the data (Dou and Neubig, 2021) This will allow for an intuitive visualization for a user in a fact-checking setting.

We consider three major types of semantic inaccuracy, following Maynez et al. (2020) The first is **extrinsic hallucination** – a phenomenon where the text includes additional information that is not directly inferable from the input data, such as introducing new entities. The second and more subtle way of introducing semantic inaccuracy is **intrinsic hallucination** – creating new relations between entities that are not described in the input data. Finally, we consider omission – omitting some information from the source data in the target text.

### 2.1 SoTA in Semantic Accuracy Evaluation

We review state-of-the-art semantic accuracy metrics and discuss the limitations we aim to address in our work. We refer to Celikyilmaz et al. (2020) and Sai et al. (2022) for a broader overview.

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. The major difference between these metrics and the method we propose later on in this section is that instead of comparing the target text with the source data, they compare it with a reference text. This means the methods can only be applied to examples where a reference is available. 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). 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 (with the exception of 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. A work by...
Goyal and Durrett (2020) mitigates this by applying entailment to dependency trees. This method is not equipped to deal with negation and omission which we aim to address in our work.

Finally, we review a text-level error detection metric for table-to-text generation presented by Kasner et al. (2021). This metric uses rules to construct a set of sentences that can be derived from the input data and measure the semantic similarity between them and the evaluated sentence. We aspire to reach a better result by crafting a synthetic pre-training set containing more intricate hallucinations as described later on in this section.

2.2 Metric Evaluation

To our knowledge, there is not yet an objective way of evaluating how well semantic accuracy metrics perform in finding inaccurate information. We might not fully achieve objective evaluation of metrics but we argue it is important to move towards this goal as it will lead to better evaluation methods. The most prevalent method of measuring metric performance is comparing the scores given to selected evaluated examples to human judgment. However, such evaluation is not easily reproducible and does not give us enough information to compare the metrics among themselves (Belz et al., 2021).

Data-to-text datasets such as WebNLG (Gardent et al., 2017), Enriched WebNLG (Castro Ferreira et al., 2018), DART (Nan et al., 2021) are not sufficient for benchmarking evaluation metrics. As datasets intended as NLG system data, they generally do not contain phenomena like hallucination, but in the rare cases when they do, they are not marked as such. The closest to our goals is the dataset presented by Thomson and Reiter (2021) intended for error detection in table-to-text generation. It contains high-quality human annotation at the drawback of being small in size – ~90 examples across train and validation sets combined. Maynez et al. (2020) created such a dataset for the task of abstractive summarization by extending the XSum dataset (Narayan et al., 2018). They conducted a human annotation experiment to tag hallucinations in the generated summaries. While we hope we can extend our evaluation method to abstractive summarization, this dataset is not directly suitable for evaluating data-to-text generation. A similar benchmarking dataset is available for dialogue systems (Dziri et al., 2022). This dataset contains annotations with manually evaluated judgments about whether a system response is fully attributable to a relevant large unstructured source of information. Such task is out of scope for this thesis.

To create a unified way of evaluating and comparing NLG metric performance, we propose a construction of a dataset designed for data-to-text metric evaluation which will contain examples of semantically accurate texts, both extrinsic and intrinsic hallucination, and omission. This will allow for a fine-grained diagnostic of the metric performance in a fully automated setting.

A portion of the data-to-text datasets mentioned above will serve as positive examples containing no hallucinations or omissions. Hallucinations could be automatically generated by dropping semantic triples. We selected this format as our starting point for several reasons:

- It is widely used in the datasets we considered.
- Other formats (tables, graphs, name-value slot pairs) can be losslessly transferred to semantic triples.\(^1\)

In case we drop a triple where both the subject and object are included in other triples, we are creating an intrinsic hallucination, since the only thing being removed is the relation between the two. Otherwise, we are creating an extrinsic hallucination.

Generating examples of omission could be done by dropping a sentence from the reference text whenever there are more sentences. More intricate examples could be generated by dropping a subtree from the dependency tree of the reference.

A portion of the dataset should also include categorized outputs produced by various NLG systems. This will ensure that the metric itself is properly evaluated on the data it was designed for. There is no scarcity of erroneous NLG outputs, however, the bottleneck will be the need for human annotation and categorization. For this reason, we intend to start with a small set of such data and slowly expand it.

Creating such a benchmarking dataset would help us compare the performance of existing metrics on the three categories of inaccuracies and to understand their limits.

\(^1\)We consider graphs as tuples \(G = (V, E)\) where \(V\) is a set of vertices and \(E\) is a set of edges. We propose that the edges can be converted to predicates and vertices can be converted to subjects and objects in the semantic triples.
2.3 Evaluation Method

We propose a novel method to evaluate semantic accuracy based on alignments between source data and target text. Using the alignment method introduced by Dou and Neubig (2021), we intend to align portions of the data, e.g. semantic triples, to phrases in the target text. To reach phrase-level granularity, we aim to use dependency trees – inspired by the work of Vamvas and Sennrich (2022) and Goyal and Durrett (2020).

If a portion of the data cannot be aligned with any combination of the phrases, it means the information was omitted. On the other hand, if a phrase cannot be aligned with any portion of the data, it is likely indicative of a hallucination. We are aware this could also happen with filler words or phrases. We can handle such cases during dependency parsing or filter them through their perplexity – filler phrases generally have a lower perplexity than information-bearing phrases.

The main output of this method is the set of alignments that can be used to flag any suspicious parts. However, in a development setting, it is desirable to have a numerical output quantifying the quality of an evaluated system. This can be obtained either as a total distance between the aligned embeddings in the embedding space or the percentage of embeddings not aligned. Both scores can be normalized for sequence length.

The advantage of this method is that it allows us to track the source of all information in the target text, not only the inaccurate parts. This can be useful in a setting where the alignments are presented directly to the user because if visualized properly, it could make fact-checking faster and easier.

Expected Qualities  We aspire for the evaluation method to have the following qualities:

- **Explainable** Instead of just outputting a numerical value to characterize the accuracy of a target text given the source data, it also identifies the hallucination spans. Therefore, it should be able to point out precisely which parts of the text are not supported by the data or which parts of the data were omitted from the text.

- **Reference-less** The metric is designed to evaluate novel texts where no reference text is available. This corresponds to the task of quality estimation (Dušek et al., 2019; Specia et al., 2013). While this might seem like a limitation, recent work by Kocmi and Federmann (2023) shows that neural metrics are capable of reaching better results when not presented with a reference.

- **Robust** The metric is robust with respect to lexical diversity. The choice of words should not matter as long as they are semantically similar. We expect to approach this quality by working with embeddings rather than n-grams.

- **Automatic** While the metric can be used to help a user, it should not require any input from the user.

### Alternative Approach as Tagging

Finding hallucinations and omissions in the text can also be approached as a BIO tagging problem (Ramshaw and Marcus, 1995). In our case, we aim to classify every token as the beginning of a hallucination or omission. This approach has been previously explored on a more narrow task of error detection (Kasner et al., 2021) trained on data from Thomson and Reiter (2021).

We believe that training a BIO tagger could benefit from our proposed benchmarking dataset from Section 2 could be used for training such a tagger. The hallucination and omission spans can then be automatically annotated using the alignments from our main evaluation method. Even in case the alignments prove to be worse quality than anticipated, we will investigate whether adding this data as a pre-training step and then refining on high-quality data from Thomson and Reiter (2021) will lead to better performance.

3 Mitigating Inaccuracies with Interpretability

In the second part of the thesis, we will use various techniques to uncover the sources of semantic inaccuracies within networks. We will then use the gained knowledge to improve the semantic accuracy of the generated text.

In the first subsection, we discuss the methods we intend to explore. In the second subsection, we name the research questions we seek to answer.

### 3.1 Methods

We will investigate LLMs with openly accessible weights (Touvron et al., 2023; Taori et al., 2023; Chung et al., 2022; Wang et al., 2022). In our
experiments, we will aim to always have a mixture of encoder-decoder models vs decoder-only models, to explore whether the model architecture makes a difference. We will also compare models fine-tuned on instructions to those that were not to investigate whether this training schema is beneficial in increasing semantic accuracy.

Attention Visualization The first step in our search for semantic inaccuracies is using Attention Visualization (Vig, 2019). The goal is to look for an intuitive insight into what happens inside the networks while inaccuracies are generated. We will search for any reoccurring patterns that can be addressed by pruning. We bear in mind that the results might be hard to interpret or even misleading (Mareček et al., 2020; Wiegrefe and Pinter, 2019). Nevertheless, we consider this method a good place to start in our interpretability research.

Probing We anticipate that the major part of our analysis will be done using probing (Ettinger et al., 2016; Adi et al., 2017; Conneau et al., 2018). Probing aims to extract information from the network’s hidden layers by applying a classifier of an investigated linguistic phenomenon on top of them.

In this thesis, we will mostly be interested in extracting graph structures as we are equally interested in entities (nodes) and relations among them (edges). This will be inspired by extracting syntactic properties (Hewitt and Manning, 2019), and discourse structures (Huber and Carenni, 2022) from hidden layers. The core idea of both works is applying linear transformations to the activations, considering the result as a distance metric which was then applied to construct trees directly or using dynamic programming.

Our idea of utilizing this approach is to extract the structures in a similar manner and to try to match them to the input data. This can be done on multiple levels to look for the precise point when a hallucination forms by the introduction of new information into the structure or when a part of the input data is forgotten.

We also plan to build upon the work of Schuster and Linzen (2022), who show that Transformer-based models do not yet have entity tracking capabilities and can introduce new entities, which is an instance of extrinsic hallucination (Schmidtova, 2022). Klaftka and Ettinger (2020) use probing to obtain information about the surrounding words from a given word. This approach could help us reveal intrinsic hallucination in case we retrieve information about a predicate not supported by the data. We will also look into probing via prompting an LLM (Li et al., 2022) as this approach does not require a trained probe.

Pruning After identifying a potential source of inaccuracy, one of the most natural mitigation strategies is attention head pruning – removing some of the attention heads after training. Voita et al. (2019) and Behnke and Heatfield (2020) observed a comparable model performance in machine translation before and after strategically pruning attention heads.

Our aim is to identify attention heads that consistently contribute to hallucination via copying from the training data instead of attending to the input data via attention visualization and probing. In case we succeed, there is a possibility of improving a model’s semantic accuracy by pruning those heads.

Fine-tuning Fine-tuning a large pre-trained language model can be computationally very demanding. Most LLMs which achieve state-of-the-art results are simply too large to fine-tune using traditional methods on hardware accessible to a Ph.D. student. Therefore, we aim to explore methods such as LoRA (Hu et al., 2021) and QLoRA (Dettmers et al., 2023) to fine-tune LLMs using the available data-to-text generation datasets to reach higher semantic accuracy.

Furthermore, in case we find recurring hallucination patterns through attention visualization and probing, we can use the matrix injection method described by Hu et al. (2021) to remove hallucinations before they can even appear in the generated text.

Modelling Uncertainty In case a model is not confident enough in its answer, it should rather say ‘I don’t know’ instead of hallucinating a plausiblesounding response. Goldberg (2023) argues that such behavior cannot be learned in a supervised manner, as we ourselves do not know what knowledge is stored in the model.

We aim to explore Bayesian methods to estimate the model uncertainty. Wu et al. (2022) model aleatory (data) and epistemic (model) uncertainty (Kiureghian and Ditlevsen, 2009) to detect out-of-domain queries fed to dialogue systems. Our intentions are the opposite – instead of using this method on the system inputs, we aim to focus on the outputs. We intend to leverage this method to
model epistemic uncertainty and use the modeled values to update the system weights.

We believe this will be a promising research area as this is the kind of interaction humans intuitively expect.

**Prompt Engineering** The performance of LLMs largely depends on the prompts they receive. We will investigate to what extent prompt choice can influence the semantic accuracy of the produced texts. There are already many strategies and courses for prompt engineering (Bach et al., 2022; Sanh et al., 2022; Liu et al., 2021; Ng and Fulford, 2023), however, the suggested strategies for hallucination mitigation are often not very effective. We will seek the boundaries of semantic accuracy that can be achieved through prompt engineering.

We aim to experiment with zero-shot prompting (Chang et al., 2008; Palatucci et al., 2009), few-shot prompting (Brown et al., 2020), and chain-of-thought prompting (Wei et al., 2023). We are aware that a prompt that will mitigate hallucinations for one model might not be so successful for another one and we are willing to modify the prompts for specific models. We plan to experiment with many aspects of the prompt such as sentence length, unambiguity, word choice, using placeholders, special symbols as delimiters etc.

The advantage of prompt engineering is that the results will be applicable immediately. We expect to observe a wide range in LLM performance based on prompt choice.

### 3.2 Research Questions

Through our interpretability research, we aim to answer the following questions:

- Are there reoccurring patterns in attention that appear when the model is hallucinating?

- Can we use probing to identify the layers where hallucinated information infiltrates the input data?

- Is it possible to teach the network to estimate its confidence in a fact before replying? Would such confidence be reliable or arbitrary?

- Is it possible to minimize the influence of the prompt on semantic accuracy by manipulating the model by fine-tuning, pruning attention heads, or using reinforcement learning to estimate model confidence?

- How significantly can we increase semantic accuracy through modifying the model’s inner properties (weight updates, skip connections, or attention head pruning) compared to the increase we can achieve through less resource-intensive prompt engineering?

### 4 Conclusion

This thesis proposal has outlined the importance of investigating semantic accuracy in natural language generation. By focusing on this important aspect, we aim to address the challenge of ensuring that NLG systems generate text that represents the underlying data more faithfully.

We proposed a unified benchmark for NLG metrics focusing on semantic accuracy, which will enable researchers to compare them in an objective and standardized manner. Additionally, we introduced a novel semantic accuracy evaluation method, which measures how accurately the generated text represents the underlying data while also providing data-text alignments.

Furthermore, we discussed ways to investigate where inaccuracies appear inside NLG models, with the aim of identifying potential areas for improvement. Our proposed approach includes attention visualization and probing, which provide insights into the decision-making process of the models and enhance their interpretability. The mitigation strategies we aim to use with this knowledge are attention head pruning, fine-tuning, and updating the weights using estimated uncertainty. We also aim to explore how prompt engineering can contribute to more semantically accurate texts.

We hope our research will lead to improved communication between humans and machines, enhanced user experiences, and more trust from the public.

**Challenges** There is a possibility that certain LLMs may have already encountered the development and testing portions of the datasets that we plan to use for evaluation during their training process. We will be very mindful of this while conducting all evaluations and aim to use training data extraction techniques (Carlini et al., 2021) to verify whether this is the case for a particular set of data and a given LLM. However, searching for new unseen data will be challenging and is definitely something that should be addressed by a wider scientific community.
Acknowledgements

This research was supported by SVV 260575 and by the European Research Council (Grant agreement No. 101039303 NG-NLG). I would like to thank Ondřej Dušek, Mateusz Lango, Tom Kočmí, and the anonymous reviewers for their helpful feedback.

References


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


United States. Association for Computational Linguistics.


Andrew Ng and Isa Fulford. 2023. Guidelines for prompting.


Teven Le Scao and Angela Fan et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.


Jesse Vig. 2019. Visualizing attention in transformer-based language representation models.


