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

Proceedings of the 61st Annual Meeting of the
Association for Computational Linguistics - Student Research Workshop, pages 352–361
July 10-12, 2023 ©2023 Association for Computational Linguistics

(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 believe that the evaluation method presented in Section 2 could even be used for evaluating human-written texts. While it is not intended as a fact-checking method by itself, it could be used as an aid for users who perform fact-checking to warn them about text parts not consistent with the data.

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 inferrable 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; Wiegreffe 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 Carenini, 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). Klafka 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 Heafield (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 plausible-sounding 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 is 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 Kocmi, and the anonymous reviewers for their helpful feedback.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations*.

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Alshaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 2022. PromptSource: An integrated development environment and repository for natural language prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, Dublin, Ireland. Association for Computational Linguistics.

Maximiliana Behnke and Kenneth Heafield. 2020. Losing heads in the lottery: Pruning transformer attention in neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2664–2674, Online. Association for Computational Linguistics.

Anya Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021. The ReproGen shared task on reproducibility of human evaluations in NLG: Overview and results. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 249–258, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Thiago Castro Ferreira, Diego Moussallem, Sander Wubben, and Emiel Krahmer. 2018. Enriching the webnlg corpus. In *Proceedings of the 11th International Conference on Natural Language Generation*, INLG'18, Tilburg, The Netherlands. Association for Computational Linguistics.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *CoRR*, abs/2006.14799.

Ming-Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, page 830–835. AAAI Press.

Yanran Chen and Steffen Eger. 2023. Menli: Robust evaluation metrics from natural language inference.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European*

Chapter of the Association for Computational Linguistics: Main Volume, pages 2112–2128, Online. Association for Computational Linguistics.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

Ondřej Dušek, Karin Sevegnani, Ioannis Konstas, and Verena Rieser. 2019. Automatic quality estimation for natural language generation: Ranting (jointly rating and ranking). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 369–376, Tokyo, Japan. Association for Computational Linguistics.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022. Evaluating attribution in dialogue systems: The BEGIN benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, Berlin, Germany. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 179–188. Association for Computational Linguistics.

Yoav Goldberg. 2023. Reinforcement learning for language models. Accessed on May 3rd, 2023.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Patrick Huber and Giuseppe Carenini. 2022. Towards understanding large-scale discourse structures in pre-trained and fine-tuned language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2376–2394, Seattle, United States. Association for Computational Linguistics.

Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. 2020. NUBIA: NeUral based interchangeability assessor for text generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 28–37, Online (Dublin, Ireland). Association for Computational Linguistics.

Zdeněk Kasner, Simon Mille, and Ondřej Dušek. 2021. Text-in-context: Token-level error detection for table-to-text generation. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 259–265, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112. Risk Acceptance and Risk Communication.

Josef Klafka and Allyson Ettinger. 2020. Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4801–4811, Online. Association for Computational Linguistics.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Jiaoda Li, Ryan Cotterell, and Mrinmaya Sachan. 2022. Probing via prompting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1144–1157, Seattle, United States. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

David Mareček, Jindřich Libovický, Tomáš Musil, Rudolf Rosa, and Tomasz Limisiewicz. 2020. *Hidden in the Layers: Interpretation of Neural Networks for Natural Language Processing*, volume 20 of *Studies in Computational and Theoretical Linguistics*. Institute of Formal and Applied Linguistics, Prague, Czechia.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. DART: Open-domain structured data record to text generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447, Online. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745.

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

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Mark Palatucci, Dean Pomerleau, Geoffrey Hinton, and Tom M. Mitchell. 2009. Zero-shot learning with semantic output codes. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, NIPS'09, page 1410–1418, Red Hook, NY, USA. Curran Associates Inc.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv.*, 55(2).

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization.

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

Patricia Schmidtova. 2022. Theatre play generation. Master's thesis, Charles University.

Sebastian Schuster and Tal Linzen. 2022. When a sentence does not introduce a discourse entity, transformer-based models still sometimes refer to it. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 969–982, Seattle, United States. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. QuEst - a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Craig Thomson and Ehud Reiter. 2021. Generation challenges: Results of the accuracy evaluation shared task. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 240–248, Aberdeen, Scotland, UK. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Jannis Vamvas and Rico Sennrich. 2022. As little as possible, as much as necessary: Detecting over- and undertranslations with contrastive conditioning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 490–500, Dublin, Ireland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

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

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, A. Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, M. Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddharth Deepak Mishra, Sujan C. Reddy, Sumanta Patro, Tanay Dixit, Xu dong Shen, Chitta Baral, Yejin Choi, Hannaneh Hajishirzi, Noah A. Smith, and Daniel Khashabi. 2022. Benchmarking generalization via in-context instructions on 1,600+ language tasks.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Yanan Wu, Zhiyuan Zeng, Keqing He, Yutao Mou, Pei Wang, and Weiran Xu. 2022. Distribution calibration for out-of-domain detection with Bayesian approximation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 608–615, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.