Who Are All The Stochastic Parrots Imitating? They Should Tell Us!

Sagi Shaier[▽] **Lawrence E. Hunter**[†] **Katharina von der Wense**[▽◇♠]
[▽]University of Colorado Boulder
[†]University of Colorado Denver
[◇]Johannes Gutenberg University Mainz
[▽]{sagi.shaier, katharina.kann}@colorado.edu
[†]larry.hunter@cuanschutz.edu

Abstract

Both standalone language models (LMs) as well as LMs within downstream-task systems have been shown to generate statements which are factually untrue. This problem is especially severe for low-resource languages, where training data is scarce and of worse quality than for high-resource languages. In this opinion piece, we argue that LMs in their current state will never be fully trustworthy in critical settings and suggest a possible novel strategy to handle this issue: by building LMs such that can cite their sources – i.e., point a user to the parts of their training data that back up their outputs. We first discuss which current NLP tasks would or would not benefit from such models. We then highlight the expected benefits such models would bring, e.g., quick verifiability of statements. We end by outlining the individual tasks that would need to be solved on the way to developing LMs with the ability to cite. We hope to start a discussion about the field's current approach to building LMs, especially for low-resource languages, and the role of the training data in explaining model generations.

Figure 1: An actual conversation with ChatGPT in Hebrew on the effects of not drinking enough water. ChatGPT is unable to point the user to its sources and instead falls back to a general answer ("I am ChatGPT, an OpenAI model based on the GPT-3.5 deep learning model. I am powered by OpenAI's learning set, which has been raised with the help of machine learning techniques on Internet culture, including websites, books, articles, quotes, and more"). We argue that ChatGPT and similar models should be able to direct the user to the sources of their information, which will have multiple benefits, such as quick verifiability of model statements.

1 Introduction

Transformers (Vaswani et al., 2017) and related models have been improving rapidly, with applications in a surprisingly large number of domains, such as natural language generation (Zhang et al., 2019), machine translation (Wang et al., 2019), question answering (Akermi et al., 2020), and code generation (Svyatkovskiy et al., 2020), based on the ability to generate sensible outputs to prompts over a nearly limitless input domain.

Despite impressive performance on a wide array of benchmark tasks, these models are known to produce "AI-splaining," confident sounding but incorrect statements: "To the extent that a use case places importance on the truth of the outputs provided, it is not a good fit for GPT-3" (Dale, 2021); see also Church et al. (2022) and Marcus (2019).

This problem has proven to be especially true for models trained on low-resource languages (Guerreiro et al., 2023), where data may not only be scarce (Mager et al., 2018), but also not well curated with respect to correctness or quality, in comparison to higher-resource languages (Hedderich et al., 2021). Furthermore, model hallucination in such settings can result in toxic patterns that can be found in the training data (Guerreiro et al., 2023).

In accordance with the large LMs and low-resource languages theme track, we argue that while the performance and factuality of LMs has been improving, both in high-resource and low-resource settings, in their existing state, LMs will realistically never be fully trustworthy. Thus, in settings in which factuality is required, such as medicine, they are dangerous and unemployable.

♠Formerly: Katharina Kann

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), pages 113–120
November 1–4, 2023. ©2023 Association for Computational Linguistics

This is further noted in Menick et al. (2022), who state that users cannot trust any claim a model makes without fact-checking.

Our proposal to address these concerns suggests both technical development and a simple regulatory framework: as we often ask students, journalists and scholars, **we should ask our models to name their sources and provide evidence for their assertions.** Currently, even popular LMs often fail at this, as seen on the ChatGPT (OpenAI, 2023) example in Figure 1. In the case of generative models, either the model itself or a post-hoc procedure could – and, under certain circumstances, *should be required to* – be designed to produce evidentiary justification for its output.

In this position paper, we overview which NLP tasks would benefit from such citation models, discuss the benefits they would bring, and present a roadmap to develop such models. Our goal is to motivate the field to start thinking about what is necessary to make current models truly useful in all sorts of – potentially critical – scenarios.

2 Background

Factuality and the Lack Thereof LMs store factual knowledge (Dong et al., 2022; De Cao et al., 2021; Elazar et al., 2021; Jiang et al., 2019) and previous work have shown that LMs can act as knowledge bases (Petroni et al., 2019; Sung et al., 2021). However, there is no guarantee that the retrieved knowledge is indeed factual, and unfortunately, often it is not. This can be seen in many areas, such as question answering (Xu et al., 2021), dialogue systems (Dziri et al., 2021; Shuster et al., 2021; Testoni and Bernardi, 2021), image captioning (Rohrbach et al., 2018), text summarization (Zhao et al., 2020b; Cao et al., 2022; Maynez et al., 2020) and translation (Raunak et al., 2021; Jeblick et al., 2022). This is especially true in low-resource settings (Guerreiro et al., 2023). In order for LMs to be fully utilized as such knowledge bases and in settings where factuality is crucial, the retrieved knowledge must first be factual. But, without knowing the source of such the model's knowledge, verifying its factuality is a challenge.

Citation Generation Although LMs, particularly those intended to produce scientific text, such as Meta's Galactica (Taylor et al., 2022), already produce text that looks as if it is a citation, frequently there is no document corresponding to the apparent citation or the cited document does not support the statement associated with it. Many existing approaches to citation recommendation offer productive avenues to explore for factuality testing, post-hoc generation of support, hybrid architectures, or creation of training data (Ali et al., 2022; Krasnova et al., 2023). There has also been work on citation generation, where the task is either: 1) given two documents, generate an explanation for the relation between them (Luu et al., 2020), or 2) generate a citation for an already existing text (Gu and Hahnloser, 2022; Xing et al., 2020; Wu et al., 2021; Fetahu et al., 2016). This is different from our suggestion to generate statements and citations simultaneously, and also not optimal: as LMs are being trained on massive datasets, evaluating whether each statement came from each of the potentially millions of article becomes impractical. Lastly, many existing systems that can in fact provide citations are based on search engines or retrieval models (Menick et al., 2022; Glaese et al., 2022), see also Perplexity AI♠, YouChat♠, or the ALCE benchmark (Gao et al., 2023). This is problematic because 1) it is far more time consuming than directly generating citations together with text; 2) access to the information sources needs to be provided at all times; 3) in contrast to our proposed approach, it does not increase model interpretability; and 4) for low-resource languages the quantity and quality of the data is often limited, and hence result in difficulties retrieving the relevant, factual source.

3 Citations and Their Pros and Cons

In this section, we will first discuss which NLP tasks – according to us – require LMs with an ability to cite their sources. We will then discuss the benefits and, subsequently, risks of such models.

3.1 Which Tasks Require Citations?

We propose to classify tasks via two questions: (1) Is the source of the generated text obvious? (2) Is the generated text an objective truth or a subjective statement? See Table 1 for examples.

If the answer to the first question is *yes*, no further citation is required. This is the case, e.g., for machine translation (Brants et al., 2007): the content of the generated text comes from the input sentence. The same holds true for summarization (See et al., 2017) and paraphrase generation (Zhou

♠https://www.perplexity.ai/
♠https://you.com/

and Bhat, 2021). However, this is only partially the case for text simplification (Sheang and Saggion, 2021): while most of the content comes from the original text, simpler versions of text sometimes contain additional explanations, which do require citations. In contrast, for many other tasks the input does not act as the source for text generation – instead, the output comes from information stored in the model parameters and, thus, originally from the training data. An ideal system would be able to cite the part of its training data responsible for any given output. This is the case for the popular NLP tasks of closed-book free-text question answering (Roberts et al., 2020), dialogue generation (Zhao et al., 2020a), or creative writing (Xu et al., 2020).

For tasks for which the answer to Question 1 is *no*, we then turn to the second aforementioned question and ask if the generated text without clear sources of information in the input contains what should be objective truths. This is typically true for closed-book free-text question answering, which, as a consequence, according to our rules does require citations. However, this is *sometimes* the case for other tasks too, such as the generation of additional explanations during text simplification or image captioning. Similarly, for dialogue generation, objective truths and subjective statements could be mixed within the same conversation. As a result, some generated statements for those tasks do require citations, while others are good without.

3.2 Benefits of Citations

Citations allow us to verify the factuality of generated text easily. In contrast, without knowing where the text came from we are often unable to verify that it is correct. Moreover, knowing what portion of the text is copied verbatim allows us to give credit to the author and prevent copyright violations. Citations also increase the explainability of the answer and allow users to learn more about interesting topics. Additionally, recent work in prompt engineering have shown that models providing justifications for their assertions (even when only partially correct) can improve the correctness of the outputs (Jung et al., 2022). Trustworthiness judgments among people often include a social aspect, so by doing a good job of identifying sources and influences has the potential to increase both the trust in AI systems and their trustworthiness. For example, human trustworthiness judgments about scientific claims are influenced by the interests of the authors (Gierth and Bromme, 2020).

3.3 Risks of Citations

Unfortunately, citations also come with risks. Just by having a citation next to a generated text, users are more likely to trust it (Thornley et al., 2015). However, it is likely that users will not examine each and every citation manually to verify that the text is indeed factual, or that the source is trustworthy (Simkin and Roychowdhury, 2002; Thornley et al., 2015). This will be exacerbated by the fact that it is incredibly unlikely that any automated system will ever produce 100% correct citations at all times, and may result in either users' diminishing trust and usage of such systems or a potential harm.

There is also the risk of decreased readability: backing up every statement with many citations, as the text may appear in multiple places, will reduce the readability of the text and may hinder users from reading or understanding it. Lastly, privacy concerns also arise from the training process of LMs. For example, state of the art LMs are often trained on a massive automatically extracted text (Radford et al., 2018). But, as manual examination of each text is not feasible for its size, there is a possibility that it may contain private user information, such as patient records. This may result in LMs cite information that should stay private.

3.4 Citations vs. Explainability

The goal to understand why a model generates any given output is shared with research on model explainability (Danilevsky et al., 2020). However, in contrast to the latter, we are not interested in the effect of certain input on the output. In addition, we do not necessarily require that the model describes its reasoning by providing citations – what we care about instead is that the citations back up the model's answer. This enables humans to verify the output – even if the cited source should not actually in the technical sense have been the reason for the model's output.

4 Road Map

4.1 The Big Picture

Meta-information Currently, the standard in the field is to train models on text, disjoint from its origin. Even though some models are trained on data that contain text with citations (e.g., Taylor et al. (2022)), the citations are only "attached" to statements taken from other sources, while any

Task	Q1	Q2	Citation?	Example
Creative writing	No	Sometimes	Sometimes	Penguins are known for their ability to survive in harsh Antarctic conditions [CITATION], but few people know that they also possess the power of telekinesis which they use to build intricate nests out of ice blocks.
Dialogue generation	No	Sometimes	Sometimes	Did you know that penguins can jump up to 6 feet out of water when leaping onto land or ice floes? [CITATION]. I think elephants can do the same.
Free-text QA	No	Yes	Yes	The current president is not a penguin [CITATION].
Image captioning	No	Sometimes	Sometimes	A group of penguins diving into the ocean to catch fresh fish for dinner, highlighting their impressive swimming abilities [CITATION], while one penguin emerges victorious with a giant fish twice its size.
Paraphrase generation	Yes	N/A	No	**Source text**: Penguins are social animals who live in large colonies. **Paraphrased sentence**: Penguins thrive in community living
Summarization	Yes	N/A	No	**Source text**: Emperor penguins are the largest species of penguin, standing up to 4 feet tall. They are skilled hunters, capable of catching fish and krill by diving hundreds of feet below the surface. **Summary**: Emperor penguins are notable for their size and hunting prowess, making them formidable predators in their environment.
Text simplification	Sometimes	Sometimes	Sometimes	**Source text**: Penguins have evolved unique adaptations that allow them to survive in environments as harsh as Antarctica, such as their countershaded dark and white plumage, which camouflages them from predators above and below the ice. **Simplified text**: Penguins live in Antarctica, which is year-round one of the coldest places on Earth [CITATION], and they look different than other birds so they don't get eaten.
Translation	Yes	N/A	No	**Source text**: Penguins are cool. **Translated text**: Pinguine sind cool.

Table 1: An overview of natural language generation tasks together with our opinion regarding if they require citations. Q1: *Obvious source?* Q2: *Objective truth?*

other text, even taken from the same article, does not have a citation attached to it. This results in LMs that can only sometimes, on a limited text, produce citations. In order to develop LMs that can cite their sources effectively, we need to give them the metadata which contain citation information.

Retrieval Say we trained a LM with the right data such that it has knowledge of which statement came from which article. How would we extract text with citations? One avenue for such knowledge extraction is to modify the pretraining, such that citation information is being generated together with every piece of generated text.

When To Cite? The above strategy would result in LMs that would always produce a citation. However, as mentioned in Section 3.1, not every task or statement requires a citation. For tasks that do require citations, we can just let the model always cite. For tasks that do not require citations, we can simply remove the citations. For tasks in between, where citation is sometimes required, we propose to utilize the existing subjectivity classification task (Wiebe et al., 1999).

4.2 Concrete Tasks to Master

Our goal is to lay out a roadmap for the community, which describes necessary steps for the development of models that can cite their sources. This is not trivial, as it requires improvements of models for existing tasks as well as the development of systems for novel challenges.

Simultaneous Citation and Text Generation As mentioned in Section 2, existing work mainly retrieve citations for already generated text, which becomes intractable as models are trained on ever more text and the number of possible source documents increases drastically. In contrast, we propose STANCE: the task of **S**imultaneous **T**ext **AN**d **C**itation g**E**neration. As an additional challenge, future work should also focus on MultiSTANCE: multihop citation generation, where the sources for a given text are spread across multiple texts. As the number of citations can be significant (though much smaller in the low-resource setting), we suggest to use topic modeling, as a potential avenue to reduce such large search space.

Subjectivity Classification As mentioned in Section 3.1, whether a task requires a citation partially depends on if the text is objective or subjective. This is not a novel task as the community has been working on subjectivity classification for quite some time (Wiebe et al., 1999; Wiebe and Riloff, 2005). However, to the best of our knowledge, models for this task have not been employed in the context of citations.

Citation-Text Correctness To ensure that the retrieval step (Section 4.1) is successful, we need to identify whether the statement appears in the source. For that, two existing tasks can be used: 1) identifying which part of the generated text refers to the citation (Wang et al., 2020). 2) Validate that the citation is appropriate for the selected text span (Karadzhov et al., 2017; Martín et al., 2021; Honovich et al., 2022; Mihaylova et al., 2018; Lee et al., 2020). Using such automatic methods instead of manually verifying citations will result in faster model development.

Source Trustworthiness We all know that Wikipedia is not a reliable source for citation. We propose CUE (**C**itation q**U**ality **E**valuation), the task of evaluating the quality of the source corresponding to a generated citation. We believe there are six main sub-tasks for CUE, which consist of classifying 1) the time of publication, 2) whether the source is credible, 3) how many times the source has been cited, 4) if the author is known, 5) if the source is unbiased, and 6) if the statement and citation are still relevant. For example, answering that the current US president is Barack Obama was *previously* factual, and may still show up in many source documents, but is not factual in 2023.

5 Conclusion

Language models (LMs) performance has been improving rapidly in a wide variety of areas. However, there is the crucial issue of their generated text often being nonfactual, especially for low-resource languages. We argue that, in order for LMs to be fully trustworthy, they must cite their sources – i.e., point users to the parts of their training data that back up their outputs. In this opinion piece, we discuss NLP tasks which would benefit from such citation models, highlight the benefits and risks such models would bring, and outline the individual tasks that would need to be solved on the way to develop such LMs.

Limitations

While developing the proposed language models that can cite their sources increase their utility, there is a risk that people would trust them more without actually verifying that the generated citations are actually correct. Such increase in trust would be especially problematic in time-critical scenarios where people cannot examine each citation manually. Ideally, our proposal will result in an increase of data cleaning, such that each citation is by default trustworthy. That being said, our approach does not solve copyright issues.

Ethics Statement

The main reason for this paper is to point out shortcomings of state-of-the-art language models, which can have significant social, health-related, and economic consequences. Future work should develop systems that can cite their sources in order to facilitate a verification of the factuality of generated statements.

Acknowledgments

We would like to thank the reviewers for taking the time to provide feedback on our work. Your insights have been valuable in helping us refine and enhance our manuscript. The authors acknowledge financial support from NIH grants OT2TR003422 and R01LM013400.

References

Imen Akermi, Johannes Heinecke, and Frédéric Herledan. 2020. Transformer based natural language generation for question-answering. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 349–359, Dublin, Ireland. Association for Computational Linguistics.

Zafar Ali, Guilin Qi, Khan Muhammad, Siddhartha Bhattacharyya, Irfan Ullah, and Waheed Abro. 2022. Citation recommendation employing heterogeneous bibliographic network embedding. *Neural Comput. Appl.*, 34(13):10229–10242.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages

858–867, Prague, Czech Republic. Association for Computational Linguistics.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

Kenneth Church, Valia Kordoni, Gary Marcus, Ernest Davis, Yanjun Ma, and Zeyu Chen. 2022. A gentle introduction to deep nets and opportunities for the future. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–6, Dublin, Ireland. Association for Computational Linguistics.

Robert Dale. 2021. Gpt-3: What's it good for? *Natural Language Engineering*, 27(1):113–118.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models.

Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. 2022. Calibrating factual knowledge in pretrained language models.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models.

Besnik Fetahu, Katja Markert, Wolfgang Nejdl, and Avishek Anand. 2016. Finding news citations for wikipedia. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations.

L. Gierth and R. Bromme. 2020. Beware of vested interests: Epistemic vigilance improves reasoning about scientific evidence (for some people). *PLoS One*, 15(4):e0231387.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements.

Nianlong Gu and Richard H. R. Hahnloser. 2022. Controllable citation text generation.

Nuno M. Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Ricke, and Michael Ingrisch. 2022. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2019. How can we know what language models know?

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations.

Georgi Karadzhov, Preslav Nakov, Lluis Marquez, Alberto Barron-Cedeno, and Ivan Koychev. 2017. Fully automated fact checking using external sources.

F. V. Krasnova, I. S. Smaznevicha, and E. N. Baskakova. 2023. Text sampling strategies for predicting missing bibliographic links.

Nayeon Lee, Belinda Z. Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact checkers?

Kelvin Luu, Rik Koncel-Kedziorski, Kyle Lo, Isabel Cachola, and Noah A Smith. 2020. Citation text generation. *arXiv preprint arXiv:2002.00317*.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Gary Ernest Davis Marcus. 2019. *Rebooting AI: Building Artificial Intelligence we can Trust*.

Alejandro Martín, Javier Huertas-Tato, Álvaro Huertas-García, Guillermo Villar-Rodríguez, and David Camacho. 2021. Facter-check: Semi-automated fact-checking through semantic similarity and natural language inference.

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.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes.

Tsvetomila Mihaylova, Preslav Nakov, Lluis Marquez, Alberto Barron-Cedeno, Mitra Mohtarami, Georgi Karadzhov, and James Glass. 2018. Fact checking in community forums.

OpenAI. 2023. Chatgpt: Optimizing language models for dialogue.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases?

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Kim Cheng Sheang and Horacio Saggion. 2021. Controllable sentence simplification with a unified text-to-text transfer transformer. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation.

M. V. Simkin and V. P. Roychowdhury. 2002. Read before you cite!

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases?

Alexey Svyatkovskiy, Shao Kun Deng, Shengyu Fu, and Neel Sundaresan. 2020. Intellicode compose: Code generation using transformer.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science.

Alberto Testoni and Raffaella Bernardi. 2021. "I've seen things you people wouldn't believe": Hallucinating entities in GuessWhat?! In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 101–111, Online. Association for Computational Linguistics.

Clare Thornley, Anthony Watkinson, David Nicholas, Rachel Volentine, Hamid R. Jamali, Eti Herman, Suzie Allard, Kenneth J. Levine, and Carol Tenopir. 2015. The role of trust and authority in the citation behaviour of researchers. *Inf. Res.*, 20.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Pancheng Wang, Shasha Li, Haifang Zhou, Jintao Tang, and Ting Wang. 2020. Toc-rwg: Explore the combination of topic model and citation information for automatic related work generation. *IEEE Access*, 8:13043–13055.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822, Florence, Italy. Association for Computational Linguistics.

Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *Conference on Intelligent Text Processing and Computational Linguistics*.

Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 246–253, College Park, Maryland, USA. Association for Computational Linguistics.

Jia-Yan Wu, Alexander Te-Wei Shieh, Shih-Ju Hsu, and Yun-Nung Chen. 2021. Towards generating citation sentences for multiple references with intent control.

Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan. 2020. Automatic generation of citation texts in scholarly papers: A pilot study. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6181–6190, Online. Association for Computational Linguistics.

Peng Xu, Davis Liang, Zhiheng Huang, and Bing Xiang. 2021. Attention-guided generative models for extractive question answering.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. Megatron-cntrl: Controllable story generation with external knowledge using large-scale language models.

Haoyu Zhang, Jingjing Cai, Jianjun Xu, and Ji Wang. 2019. Pretraining-based natural language generation for text summarization. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 789–797, Hong Kong, China. Association for Computational Linguistics.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020a. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online. Association for Computational Linguistics.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020b. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online. Association for Computational Linguistics.

Jianing Zhou and Suma Bhat. 2021. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.