Literature Review: Bounding the Capabilities of Large Language Models in Open Text Generation with Prompt Constraints

1 Summary

This paper delves into an in-depth examination of large language models (LLMs) with a primary focus on GPT-3, assessing its proficiency in generating open-ended text while adhering to particular stylistic and structural guidelines. The backbone of the research lies in an innovative methodology devised by the authors, which is structured around a taxonomy of constraints divided into stylistic and structural elements.

While the stylistic constraints encompass tone, mood, genre, and viewpoint, the structural constraints revolve around numerical stipulations, descriptive conditions, and specific format requirements, such as those needed for programming code, emails, and scholarly papers. In addition to this, the authors carried out sensitivity studies to evaluate how the size of the model influenced its grasp of these constraints.

The findings revealed several areas where GPT-3 had difficulties in meeting the constraints accurately. For instance, it often failed to produce the exact number of words, sentences, or paragraphs requested, and displayed a high level of inconsistency in output when given descriptive guidelines like 'long' or 'short.' Notably, the model also struggled with the formatting standards for academic papers. However, it was noted that supplementing the prompts with more in-context information typically enhanced the model's performance.

The study also compared GPT-3's performance with other LLMs like OPT-176B (Zheng et al., 2022), BLOOM-176B, and GLM-130B (Du et al., 2022). The results showed that apart from GPT-3, the other models had a substantial rate of degenerate responses, which might be due to noisier pre-training datasets and a lack of training that is aligned with the instructions.

While acknowledging the limitations of their work, such as the taxonomy not being exhaustive in terms of stylistic and structural constraints and its lack of empirical user-centric nature, the authors bring attention to the potential ethical issues concerning the misuse of styled text and the importance of safeguarding annotators from possibly harmful content.

Despite the challenges, this research provides substantial insights into the current capabilities of LLMs, highlighting the necessity for more refined mitigation approaches, better prompt optimization, and user-centric taxonomies in future investigations.

2 Strengths

This study offers several significant contributions to the field of Natural Language Processing, particularly in our understanding and application of Large Language Models (LLMs). A notable contribution is the development of a comprehensive methodology to evaluate the capabilities of LLMs. This innovative approach allows researchers to assess how well these models generate open-ended text under stylistic and structural constraints, providing a robust framework for future investigations.

The authors further enhance our understanding by introducing a detailed taxonomy of stylistic and structural constraints, including numerical, descriptive, and formatting constraints. This carefully curated taxonomy forms the bedrock for generating prompts used to evaluate the performance of LLMs. This innovative tool offers an organized and systematic way to challenge and analyze the capabilities of these models.

The researchers apply their methodology to conduct a comprehensive analysis of GPT-3's performance under the proposed taxonomy. The exploration extends to include a sensitivity study focused on the model size. The results reveal that the understanding and adherence to stylistic and structural constraints improve with increasing model size.

The study also includes an insightful compari-

son between GPT-3 and other publicly available LLMs, such as OPT-176B, BLOOM-176B, and GLM-130B. The findings from this comparative analysis are eye-opening, highlighting the strengths and weaknesses of each model, and providing valuable insights for future development.

Finally, the paper does not shy away from addressing the ethical considerations surrounding the use of LLMs. The authors emphasize the potential for misuse of styled text, and the need to safeguard annotators from potential harm. This honest discussion highlights the importance of ethical considerations in the ongoing development and use of these powerful tools.

The paper's suggestions correspond well with the paper's findings. Their proposed methodology and taxonomy are direct responses to identified issues with Large Language Models (LLMs) meeting stylistic and structural constraints. The findings about model performance variability inform the sensitivity study on model size, suggesting that larger models might handle constraints better. Ethical concerns raised by the study are also addressed, emphasizing careful and responsible use of these tools. Overall, the authors provide practical, insightful suggestions to navigate the challenges they discovered.

3 Weaknesses

While the paper offers insightful observations about Large Language Models (LLMs), there are a few areas that could benefit from further exploration. One of these is the limited coverage of the taxonomy of stylistic and structural constraints. Although the taxonomy is well-thoughtout and useful, it doesn't encapsulate all possible constraints an LLM might face. This may inadvertently paint an incomplete picture of the model's capabilities, leaving room for future research to delve into unexplored areas.

Another aspect to consider is the heavy dependence on prompt design. The performance of these models is largely influenced by how the prompt is framed. While the study assumes optimal prompt design, this may not always be the case in practical scenarios. A poorly designed prompt could negatively affect performance, and it might be beneficial for future research to explore this aspect more thoroughly.

Lastly, the study's focus on GPT-3 might limit the generalizability of its findings. GPT-3, like all models, has its unique strengths and weaknesses. Relying heavily on its performance could skew the results and not give an accurate depiction of the capabilities of other LLMs.

Therefore, expanding the study to include a wider range of models could present a more balanced view of LLMs' performance under stylistic and structural constraints. These considerations, while not detracting from the study's overall value, provide areas for future work to further enhance our understanding of LLMs.

There are certain improvements that could make this paper more impactful:

Firstly, the authors have done a commendable job acknowledging the study's limitations. Yet, a more detailed discussion would further strengthen the paper. For instance, elucidating why Large Language Models, particularly GPT-3, fail to meet certain numerical constraints, or why other LLMs produce a high rate of degenerate outputs could deepen the understanding of these models' behavior and limitations.

Secondly, while the paper briefly touches upon the ethical considerations surrounding LLMs, this critical aspect could be expanded. A more thorough exploration of potential safeguards or policies that could prevent the misuse of styled text would provide readers with a comprehensive perspective on this pressing issue.

Lastly, the authors propose using additional in-context information as a mitigation strategy, which is indeed promising. However, a more extensive exploration of this strategy, including its practical application and potential effectiveness, could significantly enhance the paper's relevance and applicability.

In summary, this paper is a substantial contribution to the field, but with some refinements and expansions, it could provide an even more comprehensive and impactful exploration of Large Language Models' capabilities and challenges.

References

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Lianmin Zheng, Zhuohan Li, Hao Zhang, Yonghao Zhuang, Zhifeng Chen, Yanping Huang, Yida Wang, Yuanzhong Xu, Danyang Zhuo, Eric P Xing, et al. 2022. Alpa: Automating inter-and {Intra-Operator}

parallelism for distributed deep learning. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, pages 559–578.