Sparks: Inspiration for Science Writing using Language Models

Katy Ilonka Gero and **Vivian Liu** and **Lydia B. Chilton**
Columbia University
katy@cs.columbia.edu, vl2463@columbia.edu,
chilton@cs.columbia.edu

Abstract

Large-scale language models are rapidly improving, performing well on a variety of tasks with little to no customization. In this work we investigate how language models can support science writing, a challenging writing task that is both open-ended and highly constrained. We present a system for generating "sparks", sentences related to a scientific concept intended to inspire writers. We run a user study with 13 STEM graduate students and find three main use cases of sparks—*inspiration*, *translation*, and *perspective*—each of which correlates with a unique interaction pattern. We also find that while participants were more likely to select higher quality sparks, the overall quality of sparks seen by a given participant did not correlate with their satisfaction with the tool.[1]

1 Introduction

New developments in large-scale language models have produced models that are capable of generating coherent, convincing text in a wide variety of domains (Vaswani et al., 2017; Brown et al., 2020; Adiwardana et al., 2020). Their success has spurred improvements on many tasks, from classification and summarization (Brown et al., 2020) to creative writing support (Coenen et al., 2021). These improvements demonstrate that language models have the potential to support writers in real-world, high-impact domains.

Despite their successes, language models continue to exhibit known problems, such as generic outputs (Holtzman et al., 2020), lack of diversity in their outputs (Ippolito et al., 2019), and factually false or contradictory information (Lin et al., 2021). Additionally, there remain many unknowns about how this technology will interface with people in real-world writing tasks, such as how language models can best contribute to different writing forms (Calderwood et al., 2018) and how to mitigate the bias that language models encode (Bender et al., 2021).

In this work we study how language models can be applied to a real-world, high-impact writing task: science writing. This introduces challenges different to those in traditional creative writing tasks which tend to deal with common objects and relations. Science writing requires a system to demonstrate proficiency within an area of expertise. We pose the following research question: *How can language model outputs support writers in a creative but constrained writing task?*

As a test-bed, we use a science writing form called "tweetorials" (Breu, 2020). Tweetorials are short, technical explanations of around 500 words written on Twitter for a general audience; they have a low-barrier to entry and are gaining popularity as a science writing form (Soragni and Maitra, 2019). We present a system that aims to inspire writers when writing tweetorials on a topic of their expertise. This system provides what we call "sparks": sentences generated with a language model intended to spark ideas in the writer.

We report on a study in which we have 13 graduate students from five STEM disciplines write tweetorials with our system and report on how they thought about and made use of the sparks. We make the following contributions:

- a system that generates "sparks" related to a scientific concept, including a custom decoding method for generating sparks from a pre-trained language model;
- an evaluation demonstrating that sparks are more coherent and diverse than a baseline, and approach a human gold standard;
- a user study with 13 graduate students showing three main use cases of sparks and corresponding interaction patterns, as well as an analysis on how spark quality relates to participant satisfaction.

[1]This extended abstract summarizes work published in Designing Interactive Systems (Gero et al., 2022).

Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022), pages 83 - 84

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a Human-like Open-Domain Chatbot. *arXiv:2001.09977 [cs, stat]* (Feb. 2020). http://arxiv.org/abs/2001.09977 arXiv: 2001.09977.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 610–623. https://doi.org/10.1145/3442188.3445922

Anthony C. Breu. 2020. From Tweetstorm to Tweetorials: Threaded Tweets as a Tool for Medical Education and Knowledge Dissemination. *Seminars in Nephrology* 40, 3 (May 2020), 273–278. https://doi.org/10.1016/j.semnephrol.2020.04.005

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. *arXiv:2005.14165 [cs]* (July 2020). http://arxiv.org/abs/2005.14165 arXiv: 2005.14165.

Alex Calderwood, Vivian Qiu, Katy Ilonka Gero, and Lydia B Chilton. 2018. How Novelists Use Generative Language Models: An Exploratory User Study. In *23rd International Conference on Intelligent User Interfaces*. ACM.

Andy Coenen, Luke Davis, Daphne Ippolito, Emily Reif, and Ann Yuan. 2021. Wordcraft: a Human-AI Collaborative Editor for Story Writing. *arXiv:2107.07430 [cs]* (July 2021). http://arxiv.org/abs/2107.07430 arXiv: 2107.07430.

Katy Ilonka Gero, Vivian Liu, and Lydia B. Chilton. 2022. Sparks: Inspiration for Science Writing using Language Models. In *Designing Interactive Systems Conference 2022*. ACM, Virtual Event USA.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. *arXiv:1904.09751 [cs]* (Feb. 2020). http://arxiv.org/abs/1904.09751 arXiv: 1904.09751.

Daphne Ippolito, Reno Kriz, Maria Kustikova, João Sedoc, and Chris Callison-Burch. 2019. Comparison of Diverse Decoding Methods from Conditional Language Models. *arXiv:1906.06362 [cs]* (June 2019). http://arxiv.org/abs/1906.06362 arXiv: 1906.06362.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv preprint arXiv:2109.07958* (2021), 13. https://arxiv.org/abs/2109.07958

Alice Soragni and Anirban Maitra. 2019. Of scientists and tweets. *Nature Reviews Cancer* 19, 9 (Sept. 2019), 479–480. https://doi.org/10.1038/s41568-019-0170-4

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 *Advances in Neural Information Processing Systems*, Vol. 30.