Sparks: Inspiration for Science Writing using Language Models

ANONYMOUS AUTHOR(S)

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.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; *Natural language interfaces*; • **Information systems** → *Language models.*

Additional Key Words and Phrases: creativity support tools, writing support, co-creativity, science writing, natural language processing

ACM Reference Format:

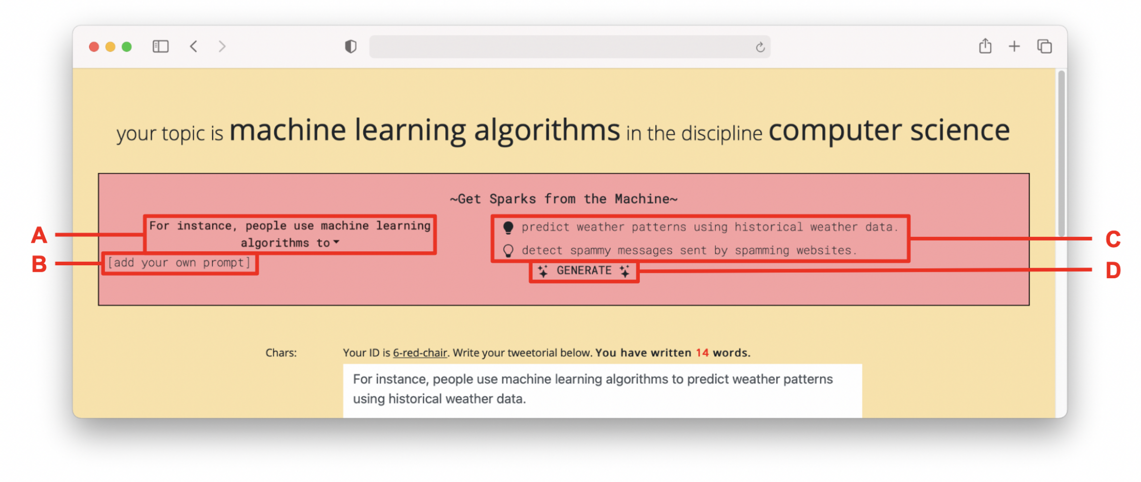
Anonymous Author(s). 2018. Sparks: Inspiration for Science Writing using Language Models. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

In this work we study how large-scale language models can be applied to a real-world, high-impact writing task: science writing. Working on science writing requires a system to demonstrate proficiency within an area of expertise. This introduces challenges different from those in traditional creative writing tasks, which tend to deal with common objects and relations. We structure our work around 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" [1]. 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 [9]. 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.

Our system generates sparks using a mid-sized language model (GPT-2 [8]) and a custom decoding method to encourage specific and diverse outputs. 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 uses a language model to generate "sparks" related to a scientific concept, including a custom decoding method for generating from a pre-trained language model;
- a user study with 13 graduate students showing three main use cases, as well as how spark quality relates to participant satisfaction.



your topic is machine learning algorithms in the discipline computer science

~Get Sparks from the Machine~

For instance, people use machine learning algorithms to ▾

[add your own prompt]

● predict weather patterns using historical weather data.

○ detect spammy messages sent by spamming websites.

Chars: Your ID is 6-red-chair. Write your tweetorial below. **You have written 14 words.**

For instance, people use machine learning algorithms to predict weather patterns using historical weather data.

Fig. 1. Example screenshot of our system generates sparks. A: writers can select from 10 templates of prompts in a drop-down menu. B: writers can add their own prompt to the drop-down menu. C: sparks are generated with a lightbulb icon to the left, if writers click the lightbulb it will highlight and the spark is copied into the text area. D: writers can hit the generate button in order to generate a new spark.

2 SYSTEM DESIGN

Generating sparks. We use GPT-2, an open source language model trained on 40GB of text from the web [8].[1] We designed a decoding method to increase the specificity and diversity of outputs: first, we modify the probability distribution using a normalized inverse word frequency; we then perform beam search using only the top 50 highest ranking tokens.[2] We also increase the diversity of outputs by forcing the first token of each output to be unique.[3] Finally, in order to keep the sparks succinct and generating quickly, we only generate 10 tokens after the prompt and cut off the generation as soon as each sentence has been completed. We implement this using the huggingface library [11].[4]

Interface. Figure 1 shows a screenshot of the system (a web interface) with its important features marked. The website consists of a single textbox for writing, and a 'prompt box' above it that allows writers to interact with the sparks. Writers can select a templated prompt or type in their own prompt. When a prompt is selected, if they press 'GENERATE' the language model will generate a single spark. Writers can 'star' a spark by clicking on the lightbulb icon—this fills in the lightbulb and also pastes the spark into the textbox. If a writer selects a different prompt, the sparks already generated are preserved such that if they return to a previous prompt their generated sparks will be shown again. The writing area is split into two sections with a line of dashes. The area above the line is reserved for brainstorming and notes, and the area below is reserved for tweetorial writing. A word count for the writer's tweetorial draft is displayed at the top of the textbox, and a character count for each tweet is displayed to the left. The website is implemented using Python 3.7 and the Flask web framework.[5]

[1] While larger open source models are available (though only to some—for example, at the time this work was done, GPT-3 [2] was only accessible to those that had been granted access by OpenAI) we wanted to limit the size of the model to reduce costs and speed up generation.
[2] The effect of this is to ensure that the modified probability distribution doesn't introduce incoherencies, for example by dramatically increasing the rank of a token very far down in the original probability distributions.
[3] While several more sophisticated methods have been proposed to increase diversity while retaining the coherence of beam search (e.g. [10]), in testing we found none were as effective as simply enforcing the first token to be unique.
[4] Link to code to be added after anonymous review.
[5] Link to demo to be released after anonymous review.

3 STUDY METHODOLOGY

In this study we sought to understand how writers make use of sparks when writing, and how spark quality relates to this usage. In particular, we pose the following research questions: *RQ1: In what ways do writers make use of language model outputs? RQ2: What attributes of language model outputs correlate with writer usage and satisfaction?* Participants were asked to write approximately the first 100 words (or about five tweets) of a tweetorial on a topic of their own expertise. We use graduate students as they are eager to participate in science writing [6], demonstrating that this is a writing task our participants may conceivably want to engage in on their own. This study was approved by the relevant IRB and run remotely via video chat and screen sharing. We recruited 13 STEM graduate students. Participants were given an introduction to tweetorials and the Sparks system, which typically took 10-15 minutes. They were then asked to pick a topic to write about, as well as provide a 'context area' to help the system interpret their topic. Participants were then given 20 minutes to interact with the system and complete the writing task. Mouse clicks and key presses while each participant interacted with the system were collected, as well as all sparks generated. After this, the participant filled out a short survey, which included the Creativity Support Index [3] and partook in a semi-structured interview with the facilitator. The entire study took about an hour and participants were compensated $40 USD.

4 STUDY RESULTS

In what ways do writers make use of language model outputs? Of our 13 participants, nine spoke in great detail about the ways in which sparks helped them. The remaining four reported that they did not find the sparks helpful. To answer our first research question we focus on the nine participants who found the system useful. In a later section of the analysis, we will analyze factors that may explain why four participants did not find the sparks helpful. Participants made use of sparks in three distinct ways: as *inspiration* for new ideas, to help with *translation* of their ideas into concrete sentences, and to understand the *perspective* of their reader. Table 1 shows examples of the three main use cases participants reported. In addition, we wanted to investigate how these three use cases correlated with participants' actual interaction with the system. To do this, we labeled each participant with a single use case, where participants who mentioned more than one use case were labeled based on the use case they said was the most prominent or that they discussed the most.[6] We then looked at: 1) the number of sparks generated and number of prompts used, 2) the number of times a user swaps between generating sparks and writing, and 3) the average longest common substring between a selected spark and what participant wrote. The 'translation' participants requested more sparks and used a higher variety of prompts to do so than others; they also swapped between writing and generating sparks more often, and copied longer portions of the sparks into their tweetorial. This demonstrates that different kinds of support result in distinct interaction patterns. [7]

What attributes of LM outputs correlate with writer action and satisfaction? We look at the quality of individual sparks, as well as the aggregate quality seen per participant, and hold the following hypotheses: *H1: Writers are more likely to star higher quality (↑ coherence, ↓ error rate) sparks. H2: Writers who see on average higher quality (↑ coherence, ↑ diversity, ↓ error rates) sparks are more likely to find the system useful.* Of the 224 sparks seen by participants, 67 were starred, which amounts to 30% of sparks seen. Using the Fisher exact test, we find that sparks without errors are significantly more likely to be starred by participants ($p < .01$). Similarly, using the Welch's t-test, we find that starred

[6] This resulted in four participants for 'inspiration', three for 'translation', and two for 'perspective'. The remaining four said sparks were not helpful.
[7] Interestingly, the number of starred sparks, as well as the number of starred sparks divided by total sparks seen, is not noticeably different between the groups, suggesting that different use cases does not mean different levels of usefulness. Participants who said the sparks were not helpful had quite varied interaction patterns, suggesting that interaction pattern alone is not enough to determine utility of a writing support tool.

Use Case	Example Usage and Quote
inspiration	spark: People care about glacier retreat over the holocene because *glaciers affect sea level rise.* what participant wrote: ...Second, *the glaciers in South America have had an outsized impact on sea level rise.* xxx% of the current sea level rise has actually be attributed to the retreat of glaciers in South America! ... quote: “My specialty is very specific and technical. And it’s often hard to figure out how to spin things in ways that feel relevant to people who don’t study this. Sea level rise is something that people would find relevant.”
translation	spark: In sociology, a deprivation index measures *societal conditions affecting individuals’ abilities to obtain goods.* what participant wrote: ...relative deprivation experienced by individuals relative to others. It can be defined as *societal conditions affecting individuals’ ability to obtain goods*, poverty levels relative to medium household income, among other definitions. ... quote:“Most of the time it [the system] was articulating the ideas that were already in my head in a way that’s short and concise.”
perspective	spark: One attribute of measurement of sexism is *that measuring sexism involves measuring attitudes towards men versus.* what participant wrote: The researchers in my study wanted to answer the question: "Does the level of sexism somewhere *impact that area’s rate of gender-based violence?*" quote: “That was helpful because the research that I do around sexism is not concerned with people’s attitudes, and instead concerned about things like incomes or legal rights or education levels. And so I wouldn’t have even thought to talk about like sexism as it relates to people’s attitudes.”

Table 1. Results of thematic analysis on reasons sparks were helpful. We report the three main use cases. Italics added by researchers to highlight where sparks influenced participant writing.

sparks have significantly higher coherence than those not starred ($p < .01$). **Thus we confirm H1: Writers are more likely to star higher quality sparks.** To test our second set of hypotheses, we look for correlations between measures of spark quality (coherence, diversity, and error rate) and the results of the Creativity Support Index (CSI) survey. The CSI nicely matches our interview data, where the four participants who reported that the system was not useful had the four lowest scores. We calculate the Pearson correlation coefficient and p-value to look for a linear relationship and find no significant correlations. **We cannot confirm H2: Writers who see higher quality sparks are more likely to find the system useful.** The interview data allows us to explore why spark quality may not correlate with perceived usefulness. We can consider the different experiences of participants P10, P12, and P13; all were graduate students in the school of public health, and all commented that sometimes the sparks misinterpreted their topic. But P10 and P12 had some of the highest CSI scores, and P13 had the lowest. P10 saw value in a spark annotated as low quality, because although it misinterpreted her topic, it gave her unique perspectives. P12 generally found the sparks useful, but when sparks were not helpful, she blamed herself for the error. In contrast, P13 described the same situation as an error on the part of the system—finding it useless. **This suggests that confounds like participant attitudes made spark quality insufficient as an explanation of perceived usefulness.** We also saw that while participants had no ownership concerns, they did report concerns about plagiarism. Some participants said “they weren’t sure where the sparks came from” and almost all said they wouldn’t want to copy too much verbatim due to these concerns. Generative models pose interesting questions about authorship, but few have taken on the implication this has for plagiarizing [4].

Research on how users develop mental models of AI systems has shown that some people are more likely to blame the system when something goes wrong, and others are more likely to blame themselves [5]. An important area of inquiry is how openness to intervention (computer or otherwise) and preconceptions of technology correlates with how useful participants find generative systems.

REFERENCES

[1] 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

[2] 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.

[3] Erin Cherry and Celine Latulipe. 2014. Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 4 (2014), 1–25.

[4] Nassim Dehouche. 2021. Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3). *Ethics in Science and Environmental Politics* 21 (2021), 17–23.

[5] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel, and Wei Zhang. 2020. Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–12. https://doi.org/10.1145/3313831.3376316

[6] Emily L. Howell, Julia Nepper, Dominique Brossard, Michael A. Xenos, and Dietram A. Scheufele. 2019. Engagement present and future: Graduate student and faculty perceptions of social media and the role of the public in science engagement. *PLOS ONE* 14, 5 (May 2019), e0216274. https://doi.org/10.1371/journal.pone.0216274

[7] Bonnie J. F. Meyer and Melissa N. Ray. 2017. Structure strategy interventions: Increasing reading comprehension of expository text. *International Electronic Journal of Elementary Education* 4, 1 (Aug. 2017), 127–152. https://www.iejee.com/index.php/IEJEE/article/view/217

[8] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog* (2019), 24.

[9] 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

[10] Ashwin K. Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models. *arXiv:1610.02424 [cs]* (Oct. 2018). http://arxiv.org/abs/1610.02424 arXiv: 1610.02424.

[11] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

A PROMPT DESIGN

We craft a 'prefix' prompt to pre-pend to any prompt used by a writer, and then hand-crafted suggested templates for prompts. We pre-pend all prompts with the following: "{topic} is an important topic in {context area}" where {topic} and {context area} are provided by the writer. In hand-crafting our prompts, we wanted to make sure our prompts captured a range of relevant angles, so our system could flexibly work with any technical discipline. We manually developed these prompts according to established frameworks within narrative and expository theory. Our prompts within the categories of *instantiation, goal, antecedent,* and *role* draw upon the constructionist framework of inferences, specifically the following categories: case structure role assignment, causal antecedent, the presence of superordinate goals, and the instantiation of a noun category (respectively). Less formally, *instantiation* prompt templates suggest completions that instantiate where and in what ways topic X may occur in the real world. *Goals* prompt templates suggest completions that represent how topic X is used in the real world. *Causes* prompt templates suggest completions for how topic X might interact in cause and effect chains. *Roles* prompt templates cover entities involved with topic X. As tweetorials exhibit both elements of narrative and expository writing, we also borrowed signal phrases from Meyer's framework for expository text [7]—e.g. "specifically", "such as", "attribute"—and folded them within our prompt templates.

Table 2. Prompt templates designed for science writing task.

category	prompt
expository	One attribute of {topic} is Specifically, {topic} has qualities such as
instantiation	One application of {topic} in the real world is {topic} occurs in the real world when
goal	For instance, people use {topic} to {topic} is used for
causal	{topic} happen because For example, {topic} causes
role	{topic} is used by {topic} is studied by

In testing we found that participants often wanted to 'follow up' on an output by entering in their own prompt. For this reason, we added the ability for writers to add their own prompts, though this prompt would also be pre-pended with our prefix.