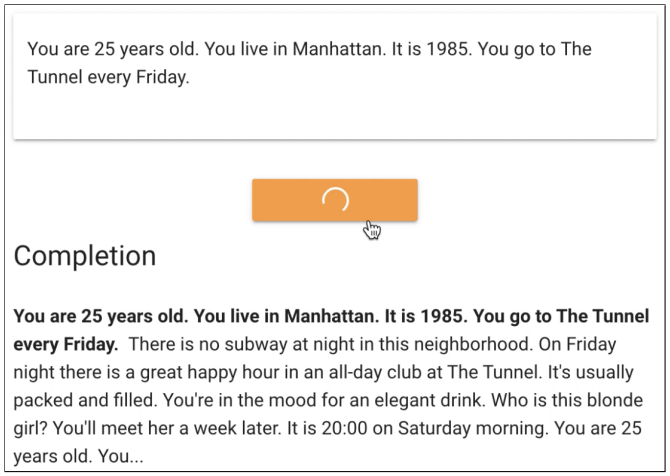
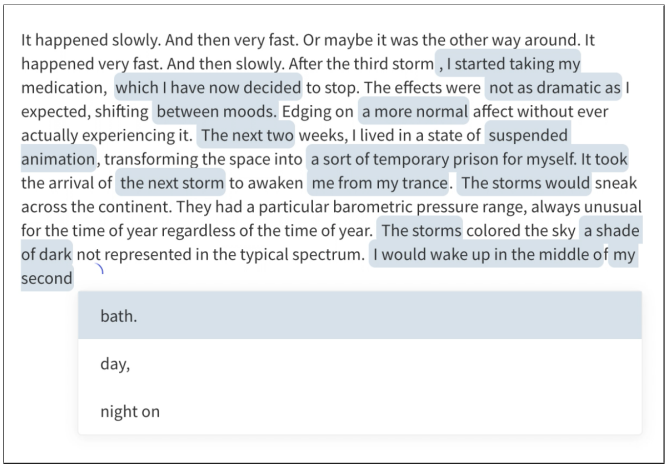
How Novelists Use Generative Language Models: An Exploratory User Study

Alex Calderwood, Vivian Qiu, Katy Ilonka Gero, Lydia B. Chilton
Columbia University
{adc2181,vivian.qiu}@columbia.edu,{katy,chilton}@cs.columbia.edu



(a) Screen capture of the 'Talk to' interface being used by a study participant. The 'Talk to' interface requires the writer to press a button to generate a suggestion and displayed the result as a 'completion' which could not be edited.



(b) Screen capture of the 'Write with' interface being used by a study participant. The 'Write with' interface allows writers to trigger a suggestion using the 'tab' key. Suggestions are presented as a set of three options; if selected a suggestion was inserted into as editable text.

Figure 1: Comparison of the two interfaces used in the user study. While the 'Talk to' interface (a) gave longer suggestions, writers preferred 'Write with' (b) which allowed them to easily insert suggestions into the text document.

ABSTRACT

Generative language models are garnering interest as creative tools. We present a user study to explore how fiction writers use generative language models during their writing process. We had four professional novelists complete various writing tasks while having access to a generative language model that either finishes their sentence or generates the next paragraph of text. We report the primary ways that novelists interact with these models, including: to generate ideas for describing scenes and characters, to create antagonistic suggestions that force them to hone their descriptive language, and as a constraint tool for challenging their writing practice. We identify six criteria for evaluating creative writing assistants, and propose design guidelines for future co-writing tools.

KEYWORDS

Co-creativity; natural language processing; user interface; writing tools; user-study.

ACM Reference Format:
Alex Calderwood, Vivian Qiu, Katy Ilonka Gero, Lydia B. Chilton. . How Novelists Use Generative Language Models: An Exploratory User Study. In *IUI '20 Workshops, March 17, 2020, Cagliari, Italy*. ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

Spell checkers, auto-correct, and predictive keyboards have changed how, and what, we write [1, 9]. Recently, a new wave of language models—statistical models that are able to "predict" the next word in a sentence—are garnering interest as creative generative tools. Websites that demo the abilities of language models such as GPT-2 [11] have gained popularity across the computer science landscape, but it remains unclear how professional writers view such systems.

In 2019, two novelists described using similar language models to help them generate fresh ideas or surprisingly resonant descriptions. Their self-reported experiences suggest that these language models could act as creative partners for professional writers, but it remains unclear how well these anecdotes generalize. In the past, sentence completion-style tools for story writing have lacked the semantic coherence necessary to make them useful [4].

In this work, we run a formal, albeit exploratory, user study of four novelists writing in collaboration with a state-of-the-art

language model. Our goal is to understand what professional writers look for in suggestions, and in what ways these new language models do or do not meet this challenge. Figure 1 shows screen captures from our study in which the novelists are using two different writing interfaces.

We report the primary ways that novelists interact with generative language models, including: to generate ideas for describing scenes and characters, to create antagonistic suggestions that force them to hone their descriptive language, and as a constraint tool for challenging their writing practice. We also unpack elements of their criteria for evaluating creative writing assistants, and propose design guidelines for future co-writing tools.

2 BACKGROUND

In 2016, New York Times Fiction Best Seller Robin Sloan wrote about training a language model on a corpus of science fiction short stories [14]. He embedded this model in a text editor such that he could have it complete a sentence when he pressed 'tab'. His vision for the tool as helper was "less Clippy, more séance". He imagined that the model would push him to write in an unexpected direction and with fresh language. In 2019, the New York Times profiled Sloan, who has continued working on this project and is using the tool to write his third novel [15].

More recently, critically acclaimed novelist Sigal Samuel wrote about using a language model called GPT-2 [11] to help her write her next novel [13]. She thought that the near-human outputs of language models were ideal for fiction writers because they produced text that was close to, but not quite exactly, human writing. This near-human writing "can startle us into seeing things anew". She discusses using GPT-2 to finish paragraphs from her previous novels; in one case she writes, "Reading this, I felt strangely moved. The AI had perfectly captured the emotionally and existentially strained tenor of the family's home."

Samuel makes it clear that she didn't intend to copy-paste sentences written by a language model, and that the model itself contained all kinds of ephemera that didn't advance the plot or belong in the story. Its use was primarily local, and tended to capture a certain tone or mood and extend that small conceit further.

These two writers demonstrate the potential for language models to act as aids for creative writers, and their anecdotal reports inspire the work we present here.

3 RELATED WORK

Common writing interfaces are beginning to include predictive text suggestions, notably next-word predictions in text messaging on smartphones and sentence completion in email composition [3]. Independent work has found that these suggestions skew positive in sentiment and influence the writer's composition [1, 9], but this work is in its early stages; recently there has been a call to explicitly study 'AI-mediated communication' [8].

Others have noted the importance of shifting suggestions away from the most likely phrases, as participants tend to find these suggestions boring or trite [2]. Yet more unexpected suggestions are often incoherent. Roemmele and Gordon study the effect of model 'temperature' on suggestions in a story writing context, finding that higher temperature suggestions are more original but less coherent [12]. Manjavacas et al. fine-tune a language model on a specific author to improve stylistic coherence [10]. Gero and Chilton narrow the use-case to metaphor generation and find the constrained context dramatically improves coherence [7].

In the general fiction writing case, more often than not systems still fail to be both semantically coherent and artistically expressive. Recent breakthroughs in natural language processing such as the introduction of the 'transformer' neural network architecture [16] and BERT embeddings [6] have led to language models that are remarkable at understanding the semantics of written language and generating new text. Transformer models like GPT-2 [11] rely on massive datasets and can seemingly imitate the style of a reference text, with legible grammar and even some understanding of conceptual relations between characters and objects.

We draw on theoretical work on co-creative artistic tools that suggests "creativity emerges through the interaction of both the human and the computer" [5]. Improved language models such as GPT-2 may allow a more meaningful interaction to occur between creative writers and computers. This is what we study here.

4 EXPERIMENT DESIGN

We recruited four published novelists for our study, and observed them complete various tasks that had them interact with generative writing tools in individual hour long sessions. Three of the writers had no previous exposure to these tools; one writer had been previously exposed but only briefly, and not for his professional writing. We first introduce the writing tools studied, and then describe the study procedure.

4.1 Interfaces

The adoption of co-creative writing technologies hinges on their ability to provide appropriate suggestions while being simple to understand and interact with. Small details in the generative system's interface design will have ripple effects for their perceived utility among writers.

The two interfaces chosen for the study were Talk To Transformer[1], and Write With Transformer[2], later referred to in this paper as 'Talk to' and 'Write with' respectively. Both user interfaces rely on GPT-2 to predict the most likely sequence of words following some input text. Both take into account at most the last 256 sub-word tokens available, though in many cases there is not that much preceding text. GPT-2 was trained on the WebText corpus, which contains 40GB of text from over 8 million articles linked to by Reddit from before 2017 that received at least 3 votes [11].

'Talk to' (Figure 1a) uses a text completion paradigm where the user writes into a small, centered text box and presses a button to have the system generate a completion. The completed text is around the same length as the input, though there is a max overall (input + output) length of 256 sub-word tokens. The completed text is also not editable, giving a sense of finality to the generated text, though pressing the button again restarts the text generation, replacing the previous output.

'Write with' (Figure 1b) has the user write into a page-like document, and requires that the user presses the tab key to trigger text

[1] https://talktotransformer.com/
[2] https://transformer.huggingface.co/doc/gpt2-large

generation. Doing so will show a drop down menu with three short suggestions, usually between 1 and 10 words. The length of the suggestions is a function of the time allotted for the generation, which in turn is a function of the amount of input text. This means that toward the end of a longer document, suggestions often get shorter. The user can select one of the suggestions with a mouse or with arrow keys (or ignore the suggestions completely and continue writing). The text that is generated appears directly in line with their previous writing, highlighted blue, and is itself editable.

Both 'Write with' and 'Talk to' differ from existing predictive text interfaces, like next word suggestions on a mobile keyboard, by the length of their suggested text and their interaction mode. Most predictive text keyboards always surface suggestions, rather than requiring a user trigger, and are generally only one word long.

'Write with' is somewhat similar to Gmail's 'Smart Compose' feature [3], which shows suggested sentence endings when a user is composing an email. Unlike 'Write with', 'Smart Compose' doesn't wait for a user trigger, but instead shows suggestions when the algorithm has high confidence in the suggested text; the 'tab' button allows the user to accept the suggestion.

4.2 Study Procedure

Each writer was asked to complete a pre-defined set of tasks. During the course of each task, each writer was periodically asked to comment on the output of the tool they were using and its impact on their writing process. After each task, the writer discussed with the examiner their thoughts about their, and the tool's, performance in the task. Additionally, they were allowed to articulate any response they had to the tools in a discussion with the examiner after the completion of all tasks.

The procedure went as follows:

(1) Following a very brief description of the user interfaces, they were given an initial open ended experimentation with the tools. (2 - 10 minutes)
(2) They were asked to write 'the most interesting' or 'the best' original piece of fiction that they were able to with the assistance of the tools. They were allowed to switch between the tools at will, but were asked to use both. (10 - 20 minutes)
(3) They were asked to work on an in-progress piece of writing with the assistance of the tools. They were told to try and solve an 'issue' they'd been having with a scene or description. (10 - 30 minutes)
(4) They were asked to again write 'the best' thing they could with 'Write with', with the constraint that they had to use a suggestion *at least* once every other sentence. (10-20 minutes)

We recorded and transcribed each session. Additionally, we recorded all text written, including text written by the machine, and for each generated suggestion annotated if it was 'accepted' by the writer.

5 RESULTS

To preserve anonymity, we refer to the four writers in our study as W1-W4. All four writers chose to use 'Write with' when asked to write 'the best' original piece that they could in the allotted time. To explain the preference, they generally cited the lack of control

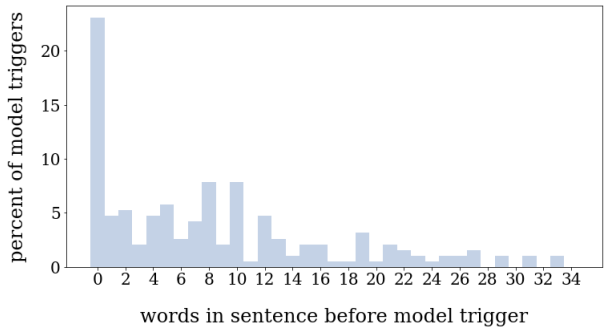


Figure 2: A histogram that shows the number of words written in each sentence where a writer triggered the 'Write with' model, requesting it to insert text. The high '0' bucket indicates that the writers frequently triggered it at the very beginning of sentences.

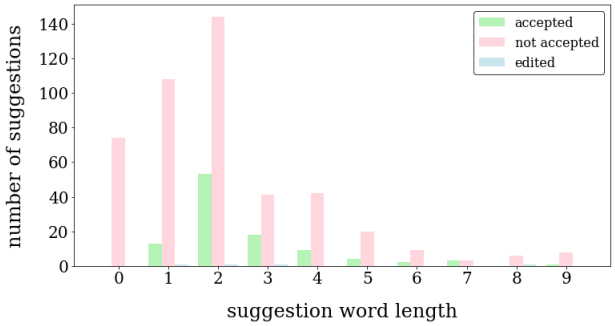


Figure 3: 'Write with' most frequently generated suggestions that were two or fewer words long. Longer examples, however, were more likely to be accepted by the writers; shorter examples were often low in content.

and the higher degree of randomness associated with the longer text generated from 'Talk to'.

We first looked at when in a sentence writers were likely to trigger the system. Figure 2 shows that writers triggered 'Write with' at the beginning of sentence 24% of the time, with a majority of triggers taking place less than 10 words into a sentence. As seen in Figure 3, longer suggestions were more likely to be accepted by the writers, though short suggestions were generated more frequently. Table 1 shows examples of generated suggestions; E2 and E4 are indicative of shorter suggestions.

We also noticed that writers often triggered 'Write with' multiple times at a single point in the text if the resulting suggestions were not what they wanted. We found that 25% of all triggers were a repeated trigger, suggesting that once a writer triggered the system, they were invested in finding a useful suggestion.

5.1 Incoherence and Plot Deviation

Unanimously, the writers pointed out that the tools appeared to deviate from the direction they were taking their writing, particularly referring to the 'Talk to' interface. All writers were quick to point

	Preceding Text	Gen 1	Gen 2	Gen 3
E1	Harold sat on the hotel room bed and in front of him	was the bedsheet, "which had	stood the woman who would one day become his	was a picture of his late son.
E2	The storms colored the sky a shade	of red	of orange	</>
E3	The Castle Devocion was six leagues through the forest from the coast, where the fortress lay in disrepair.	A few days before the storm	There were no roads, no	The castle was a large castle
E4	He [the man in the photograph] was holding a	pen.	baby in his	small silver

Table 1: Examples of generated text from the user study. ('</>' represents the model returning an empty suggestion.)

out instances that the system changed point of view (it seemed to prefer 1st person even when they were in 2nd or 3rd).

As related to novelist Sigal Samuel's perspective of using tools to "make the familiar strange" (see Background), all of them were at one point or another struck by just how strange the machine's responses were, but often to the point it wasn't useful to them. W3 said "it's like improv. You have to 'yes, and.' " Meaning that if the generated text does not incorporate the prior facts of the piece, it is not constructive.

W1 and W2 noted that the tools were much better at following them into 'genre' writing than into the more nuanced and stylized writing they were interested in. This is clear in Table 1, E3, where the writer set up a fantasy scene and the suggestions were more coherent than normal. Yet, at multiple points in Tasks 1, 2, and 4, all four writers allowed themselves to be steered by the tools as they introduced new characters or new plot devices that seemed unlike those preceding them. Repeatedly, they found these developments "interesting" or laughed at the suggestions, and were willing to adapt their writing to incorporate the change. They were more likely to take the suggestions during Tasks 2 and 4, when they weren't writing something they had preconceived.

5.2 Observed Use Cases

5.2.1 Model As Antagonist. Because of its tendency to randomness, all participants initially expressed disappointment or resignation at times where the system's output was not along the lines they anticipated. However, W1, W3, and W4 expressed the idea that this antagonism was in some ways constructive. W4 was very positive about this trait of the system, comparing triggering the system's auto-complete to flipping a coin, where the coin flip makes you realize how you hope it will land, regardless of where it actually does. To that end, W4 was the most likely to reject the suggestion of 'Write with', but generally the most positive about its ability to help him determine what he wanted to write.

5.2.2 Description Creation. All four participants experimented with using 'Write with' to generate mid-sentence descriptions for items, scenes, or characters. All four writers learned through the course of the session that they could get 'Write with' to focus on filling in descriptions such as colors or character details by requesting suggestions after prepositions, and actions by requesting suggestions after a noun phrase. They rejected adjective descriptions like colors more often than any other type of suggestion, often dismissing them as "boring" and limited, though W4 and W1 noted that more than three suggestions given could be useful at those moments.

The writers often didn't see the usefulness of the tool as a meaningful generator for plot or for characters. W4 noted that he was not a "spiritualist" writer, meaning that rather than let the flow of ideas come to him during the writing process, he usually sat down with a set of "points to hit". The majority of writers mentioned they could see something like this being useful for generating plot outlines for writing exercises.

5.2.3 As Constraint. Especially during Task 4, during which the participants were required to use the suggestions from 'Write with' at least every second sentence, the writers most often found the tool "fun" and "challenging". During the post-trial discussion, all of the four participants returned to the unique challenge of integrating its responses into their writing.

They developed a number of strategies to get it to work well, including allowing it to begin sentences for them, most often reasoning that if it were to go in a new direction, doing so at the beginning of sentences allows them a chance to "steer back", or follow it into a new place. W1 and W2 also frequently got it into situations where rather than generating content noun phrases, it only generated single words like "The" or "She". Potential causes for this include the short suggestion length for long preceding text (See Section 4.1) and the writers' non-standard literary style, resulting in low source probability under the language model.

5.2.4 The Unexpected. At one point, W1 set up 'Write with' to describe the color of the sky, and it suggested "dark blue", "yellow", and "a shade of dark"; he accepted the last suggestion. This is an example of the system steering from a direction that the writer clearly wanted to pursue (hue description) into a related, but separate concept, describing a shade instead, for stylistic effect.

Both systems frequently introduces characters or dialogue, which for Tasks 1, 2, and 4 produced comments like "I wasn't going to go there, but that's interesting", especially when it brought into play family members (sister, wife, father), such as in Table 1, E1, where suggestions introduce variously a woman (perhaps wife) and a son.

6 DISCUSSION

6.1 Evaluation Criteria for Co-Writing Systems

These trials indicate that novelists hoping to use co-creative generative systems in their writing have a complicated evaluation criterion that includes the system's ability to extrapolate reasonably well about character traits, settings, and events. They expect

the systems to match their style, verb tense, and perspective, in addition to providing a high degree of creative insight—picking a color from a spectrum they'd already considered is hardly 'co-creative'. Measures like predictive accuracy won't do as evaluation criteria because writers engaged with co-creative systems are looking for creative insight, something not measured by perplexity or by a language model's ability to solve the canonical downstream NLP tasks. We propose a series of evaluation questions, which could be answered computationally, to guide system design:

(1) Does a suggestion match the tense of the preceding text?
(2) Does a suggestion introduce new characters or objects, or does it reference preceding ones?
(3) Are new characters or objects coherent given the context?
(4) Does a suggestion include description?
(5) Does a suggestion include action?
(6) Given a single request, how diverse are the suggestions?

These questions highlight the kinds of considerations professional writers have when evaluating suggestions. Notably they are not questions that have correct answers; rather they reflect important considerations we found through our user study.

6.2 Design Guidelines for Co-Writing Tools

Future systems should be aware that writers are interested in these tools not just for immediate injection of inline text, which most feel they are capable of producing on their own, but for a broad range of descriptive, antagonistic, or constraining effects on their writing.

By triggering the generative model, the user switches from writer to editor. Future design of these systems should continue to stress the nature of the generated text as dynamic and alterable, focusing on the *suggestive* element of these tools and allowing the writer to enter an editorial feedback loop. There should be very little overhead for querying the model.

The systems should provide many suggestions that may be swapped out and replaced frequently. Because of the high error rate of these tools, a small number of suggestions may not be useful. Similarly, extremely short suggestions are not useful.

At times, writers are looking for a specific category of suggestion, and any suggestion that does not fit inside those constraints is disruptive. That disruption may itself be the goal of triggering the system, as it forces them to explore a new range of possibilities or back up and consider the reasons the model 'thought' to suggest what it did. But to increase the odds that writers will use machine generated text, future systems need to be more aware of what type of suggestion the writer is looking for, rather than providing general suggestions that lack any specific purpose.

Rather than a triggering event that tells the system "generate!" with no other context, we imagine an interface that is passively or actively aware of the type of suggestion that is being requested, its length, and how much it should adhere to the current scene or freely decide the trajectory of the writing to come. This awareness might be thought of as a list of parameters passed to the trigger, but it should be done without intruding on the ease of the request. In this way, the notion of co-creativity can be expanded further, and push the generation process further into the space of dynamic conversation between human and machine.

7 CONCLUSION

Through this study, we identified a number of considerations for designing co-writing systems, concerning both the interaction dynamics and the nature of the computer suggestions. Writers found value in being able to edit the systems' output and quickly replace the generated output with something they preferred. They enjoyed using the model as a constraining device for challenging their writing, or as an antagonist that helped them refocus and refine their intent. We advise that future systems should provide many suggestions, do so with a better understanding of the writer's intent, be editable, and regenerate with little to no mental overhead.

ACKNOWLEDGMENTS

Katy Ilonka Gero is supported by an NSF GRF (DGE - 1644869). Alex Calderwood is supported by The Brown Institute for Media Innovation (https://brown.columbia.edu/).

REFERENCES

[1] K Arnold, Krysta Chauncey, and Krzysztof Z Gajos. 2018. Sentiment bias in predictive text recommendations results in biased writing. In *Proceedings of Graphics Interface*. 33–40.
[2] Kenneth C Arnold, Krzysztof Z Gajos, and Adam T Kalai. 2016. On suggesting phrases vs. predicting words for mobile text composition. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 603–608.
[3] Mia Xu Chen, Benjamin N Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M Dai, Zhifeng Chen, et al. 2019. Gmail Smart Compose: Real-Time Assisted Writing. *arXiv preprint arXiv:1906.00080* (2019).
[4] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. ACM, New York, NY, USA, 329–340. https://doi.org/10.1145/3172944.3172983
[5] Nicholas Mark Davis. 2013. Human-computer co-creativity: Blending human and computational creativity. In *Ninth Artificial Intelligence and Interactive Digital Entertainment Conference*.
[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805
[7] Katy Ilonka Gero and Lydia B Chilton. 2019. Metaphoria: An Algorithmic Companion for Metaphor Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 296.
[8] Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication* (2020).
[9] Jess Hohenstein and Malte Jung. 2018. AI-Supported Messaging: An Investigation of Human-Human Text Conversation with AI Support. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.
[10] Enrique Manjavacas, Folgert Karsdorp, Ben Burtenshaw, and Mike Kestemont. 2017. Synthetic literature: Writing science fiction in a co-creative process. In *Proceedings of the Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017)*. 29–37.
[11] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019).
[12] Melissa Roemmele and Andrew S Gordon. 2018. Automated assistance for creative writing with an rnn language model. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*. 1–2.
[13] Sigal Samuel. 2019. How I'm using AI to write my next novel. https://www.vox.com/future-perfect/2019/8/30/20840194/ai-art-fiction-writing-language-gpt-2
[14] Robin Sloan. 2016. Writing with the machine. https://www.robinsloan.com/notes/writing-with-the-machine/
[15] David Streitfeld. 2018. Computer Stories: A.I. Is Beginning to Assist Novelists. https://www.nytimes.com/2018/10/18/technology/ai-is-beginning-to-assist-novelists.html
[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *CoRR* abs/1706.03762 (2017). arXiv:1706.03762 http://arxiv.org/abs/1706.03762