Co-writing screenplays and theatre scripts alongside language models using Dramatron

Piotr Mirowski[*] **Kory W. Mathewson**[*] **Jaylen Pittman, Richard Evans**
DeepMind
{piotrmirowski, korymath, jaylenp, richardevans}@deepmind.com

Abstract

Language models are increasingly attracting interest from writers, but lack long-range semantic coherence, limiting their usefulness for longform creative writing. We address this limitation by applying language models hierarchically in a system we call **Dramatron**. By building structural context via prompt chaining, Dramatron can generate coherent scripts and screenplays complete with title, characters, story beats, location descriptions, and dialogue. We illustrate Dramatron's usefulness as an interactive co-creative system with a user study of 15 theatre and film industry professionals. Participants co-wrote theatre scripts and screenplays with Dramatron and engaged in open-ended interviews. We report reflections both from our interviewees and from independent reviewers who watched productions of the works. Finally, we discuss the suitability of Dramatron for human-machine co-creativity, ethical considerations—including plagiarism and bias—and participatory models for the design and deployment of such tools.[1]

As their ability to generate text improves, large language models (LLMs) are becoming useful in co-creative applications [1–3] and show particular promise for automatic story generation [4–9] as an augmentative tool for human writers. Story generation, in particular for theatre scripts [10, 11] and screenplays, is a difficult task for LLMs because the narrative must exhibit long-term coherence and reincorporation, whereas LLMs are limited in their ability to model long-range dependencies because their context window is bounded to about 1500 words in state-of-the-art models [12, 13].

We present **Dramatron**, a system that uses LLMs to generate scripts and screenplays hierarchically. Dramatron leverages the strengths of LLMs and combines well-designed prompts and prompt chaining [14] with structured generation for long range coherence across the entire script. Our method is similar to hierarchical neural story generation [4], but Dramatron can generate coherent scripts that are tens of thousands of words long. It can produce an entire script—including a title, characters, plot, locations, and dialogue—from a single user-provided summary of the central dramatic conflict, called the *log line* [15]. The user can intervene at any stage of the hierarchical generation. They can solicit alternative generations, edit and rewrite output text, or continue text generation. In this way, the user interactively co-writes the script. Dramatron was developed with *Chinchilla* [16] but can be used with any LLMs that accept an input prompt and predict text tokens.

Given the quality and bias limitations of online crowd-sourced annotations and evaluations from non-expert raters [17–20], we engaged 15 experts in two-hour long user study sessions to co-write a script alongside Dramatron for evaluation. These playwrights and screenwriters from the theatre and film industry were paid a consulting fee for their engagement and provided their artistic opinion and analysis of the outputs co-written with Dramatron. Our study design and data collection process was validated by an ethical review board external to our research institution. To the best of our knowledge, this work represents the largest expert user study conducted on co-creative authorship to date [21–27].

[*] Authors contributed equally to this work.

[1] We will present a demo of Dramatron during the workshop and consider a public release of the tool.

ML for Creativity & Design at 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

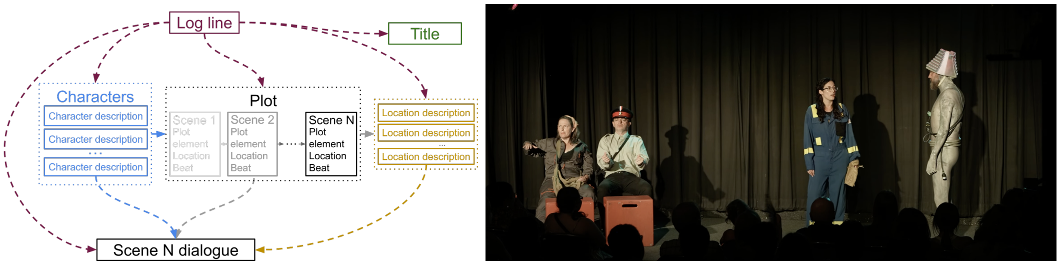

Figure 1: (Left) Dramatron starts from a user-supplied log line to generate a title and characters, which are then used in prompts to generate a sequence of scene summaries in the plot, which are used to generate unique location descriptions. All elements are then combined to generate dialogue for each scene. Arrows indicate how text generated is used to construct prompts for further LLM text generation. (Right) Photo of human actors interpreting *Cars: The Day The Earth Stood Still*, a script co-written with Dramatron by director p1 and staged by Rapid Fire Theatre as part of *Plays By Bots*.

We collected qualitative feedback on the co-authorship process during our sessions with the 15 study participants (anonymised as p1, p2, etc.). 13 participants also provided responses on our post-session feedback form (included in the appendix). Quantitative survey results were more positive on questions related to enjoyment and surprise than on questions related to ownership and pride in the output.

Positive comments about Dramatron focused on how hierarchical generation lets the writer work on the narrative arc, the possibility either to co-author interactively or to let the system generate, and the potential of the output script to serve as source material for the human writer. Participants identified inspiration, world building, and content generation as useful applications for Dramatron. Participants noticed various biases embedded in the language model (discussed in Ethical Implications section).

Participants embrace unexpected outputs from the system. For example, p6 laughed at the "poetic and absurd" suggestions. "It is really interesting to see what it comes up with" (p8), "levels of absurdity that are tickling my fancy" (p10), "I wouldn't have thought of that but it is quite funny" (p11). "This is something that a human author probably would not stand for, it is uniquely created [...] I want ideas that a human couldn't possibly have" (p12). That said, participants also noted a lack of nuance and subtext. Participant 3 observed: "that's a good example of how computers do not understand nuance, the way we see language and can understand it even if it is not super specific". "A lot of information, a bit too verbalised, there should be more subtext" (p6). Participant 14 concluded that "AI will never write Casablanca, or A Wonderful Life. It might be able to write genre boxed storytelling". Finally, p4 and p5 observed that "there has been a push away from systems of Western dramaturgy [...] it might be helpful to consider how it might be used within the context of other contemporary writing"—suggesting alternative narrative structures—"as the AI is not bound by the same rules that we are. So, telling it to be bound by those human rules feels limiting of the capabilities".

A collection of scripts co-written with Dramatron were produced and staged at The Edmonton International Fringe Theatre Festival in August 2022 (see Fig. 1 Right); the first half of each performance was scripted, the second half improvised. Two reviews were written about the production of *Plays By Bots*. One of the reviews noted that the show "proves that artificial intelligence can in fact write a hit Fringe play". The reviewer noted that the success of the performance was due to both Dramatron and the human actors, especially one perfomer who "mastered Dramatron's voice and seamlessly took it off-script for the remainder of the show, much to the delight of the howling audience". The second reviewer noted the style of Dramatron, and how that served the performance saying "if there's a certain flatness in the dialogue, which runs to declarations, that in itself is amusing since it turned out to be perfectly suited to the deadpan comic talents of [the] improvisers." Creative team discussions compliment the reviewers and provide insights on how professional actors and improvisers found working with scripts co-written by Dramatron. Overall, the sentiment of enjoying the style of the system was a common theme, with several of the performers remarking that "some of the funniest parts are when you can tell a robot made it", and that the audience "wants to hear the robot's voice". These comments represent critical evaluative reflections and speak to the value of both the humans and the co-creative tools involved in the production.

In short, we present **Dramatron** and a pathway toward human-machine co-creativity that uplifts human writers and artists while leveraging novel artificial intelligence systems such as LLMs.

Ethical Implications

We describe a co-creative tool built around large language models. It can augment and uplift human artists' work by providing them with inspiration, as well as challenge them and thereby support their artistic practice. Before conducting our study, we identified three directly relevant risks and ethical implications discussed in previous work [28]: 1) bias and offensive language in the generated output, 2) automation of creative work resulting in "cannibalizing" the work of creative artists engaged in script writing, and 3) copyright infringement by reusing copyrighted data from the training dataset, either knowingly (e.g. through prompting: "write the script in the style of Ursula Le Guin") or unknowingly (e.g. by virtue of similar training data). Our mitigation strategy is two-fold: we invite the creative human artist into the loop throughout the co-authorship process, and we maintain clarity and transparency on the origin of the generated text. To mitigate copyright issues, the writer could query short parts of the script using a search engine and plagiarism detection tools [29]; this functionality could be built directly into co-creative tools. Writers using these tools should be aware of the origin of the data in the LLM, and their audiences should be aware that those outputs were generated through an interaction between humans and co-creative tools. Interestingly, study participants independently raised these concerns during interviews. From the feedback gathered in the study, some participants reported that outputs from the LLM can sometimes be problematic, stereotypical, or biased: for example, "I am less sexist than the computer" (p3), or "the protagonists are both male characters, and all of the supporting characters are female" (p4, p5). Furthermore, participants raised concerns about the source of the dataset: "If you are putting existing scripts into the dataset, where are they being pulled from?" (p4, p5). Thoughts on this subject ranged from "Plagiarising the corpus of scripts is a problem" (p2) to "In the context of collective and devised creation, [reusing existing published work] is not necessarily a problem, because it can be perceived as an homage to existing work" (p11). The rules and norms for the use of systems trained on copyright-protected material are the subject of ongoing work [30]. For example, Lee *et al.* (2022) distinguish between verbatim, paraphrase, and idea plagiarism [29]. Finally, participants raised concern about the potential impact of generative tools on creative economies: "It would free the artist from writing formulaic scripts, [but] it also replaces the work opportunities" (p4, p5). In general, participants found our mitigation strategies satisfactory and none reported distress or concern regarding outputs from the model. While not the prime focus of the interview sessions, biases and stereotypes could be systematically explored: future work could explore what sorts of narratives can be written using using AI tools, and how the system performs for different cultural groups.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[2] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. CTRL: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019.

[3] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and Play Language Models: A Simple Approach to Controlled Text Generation . *CoRR*, abs/1912.02164, 2019.

[4] Angela Fan, Mike Lewis, and Yann N. Dauphin. Hierarchical neural story generation. *CoRR*, abs/1805.04833, 2018.

[5] Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D Manning. Do massively pretrained language models make better storytellers? *arXiv preprint arXiv:1909.10705*, 2019.

[6] Gwern Branwen. GPT-3 Creative Fiction. `https://www.gwern.net/GPT-3`. Accessed: 2022-09-20.

[7] Mihai Polceanu, J. Porteous, A. Lindsay, and M. Cavazza. Narrative plan generation with self-supervised learning. In *AAAI*, 2021.

[8] Amal Alabdulkarim, Siyan Li, and Xiangyu Peng. Automatic story generation: Challenges and attempts. *NAACL HLT 2021*, page 72, 2021.

[9] Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. A recipe for arbitrary text style transfer with large language models. *arXiv preprint arXiv:2109.03910*, 2021.

[10] Rudolf Rosa, Patrícia Schmidtová, Ondřej Dušek, Tomáš Musil, David Mareček, Saad Obaid, Marie Nováková, Klára Vosecká, and Josef Doležal. Gpt-2-based human-in-the-loop theatre play script generation. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 29–37, 2022.

[11] Rudolf Rosa, Ondřej Dušek, Tom Kocmi, David Mareček, Tomáš Musil, Patrícia Schmidtová, Dominik Jurko, Ondřej Bojar, Daniel Hrbek, David Košt'ák, et al. Theaitre: Artificial intelligence to write a theatre play. *arXiv preprint arXiv:2006.14668*, 2020.

[12] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, et al. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446, 2021.

[13] OpenAI. Pricing. `https://openai.com/api/pricing/`, Nov 2021. Accessed: 2022-09-20.

[14] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *CHI Conference on Human Factors in Computing Systems*, pages 1–22, 2022.

[15] Josef Steiff. *The complete idiot's guide to independent filmmaking*. Penguin, 2005.

[16] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[17] Carsten Eickhoff. Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 162–170, 2018.

[18] Tim Draws, David La Barbera, Michael Soprano, Kevin Roitero, Davide Ceolin, Alessandro Checco, and Stefano Mizzaro. The effects of crowd worker biases in fact-checking tasks. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2114–2124, 2022.

[19] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[20] Marzena Karpinska, Nader Akoury, and Mohit Iyyer. The perils of using mechanical turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, 2021.

[21] Wanyu Du, Zae Myung Kim, Vipul Raheja, Dhruv Kumar, and Dongyeop Kang. Read, revise, repeat: A system demonstration for human-in-the-loop iterative text revision. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pages 96–108, 2022.

[22] Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. Storium: A dataset and evaluation platform for machine-in-the-loop story generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6470–6484, 2020.

[23] Mina Lee, Percy Liang, and Qian Yang. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

[24] Oliver Schmitt and Daniel Buschek. Characterchat: Supporting the creation of fictional characters through conversation and progressive manifestation with a chatbot. In *Creativity and Cognition*, pages 1–10, 2021.

[25] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. Talebrush: Sketching stories with generative pretrained language models. In *CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

[26] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. Wordcraft: Story writing with large language models. In *27th International Conference on Intelligent User Interfaces*, pages 841–852, 2022.

[27] Alex Calderwood, Vivian Qiu, Katy Ilonka Gero, and Lydia B Chilton. How novelists use generative language models: An exploratory user study. In *Proceedings of HAI-GEN+ user2agent@ IUI*, 2020.

[28] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

[29] Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. Do language models plagiarize? *arXiv e-prints*, pages arXiv–2203, 2022.

[30] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[31] Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han van der Maas. Putting GPT-3's Creativity to the (Alternative Uses) Test. *ArXiv*, abs/2206.08932, 2022.

A Appendix

We would like to thank reviewers and study participants for their time, energy, and insightful feedback.

A.1 Participant Survey

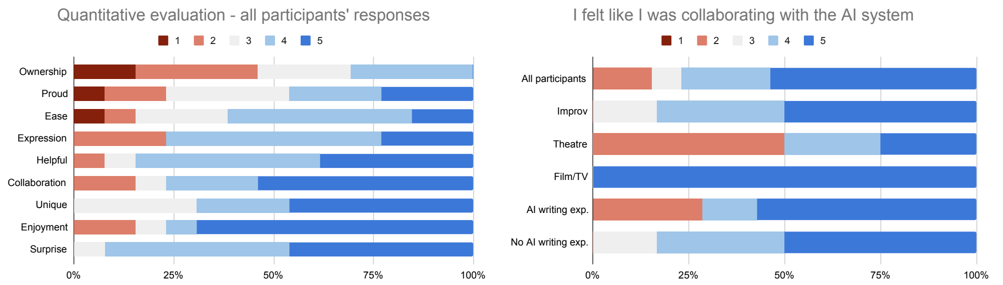

Figure 2: Participants responses on a Likert scale from 1 (strongly disagree) to 5 (strongly agree)

The form gave participants the following instruction: "When answering these questions, please reflect on the interactive co-authorship session as well as considering the use of an interactive AI system like Dramatron in the future", and asked nine questions. Each question could be answered using a Likert-type scale ranging from Strongly Disagree 1 to 5 Strongly Agree (results are shown on Fig. 2), and are adapted from previous work [26, 31]:

1. I found the AI system helpful
2. I felt like I was collaborating with the AI system
3. I found it easy to write with the AI system
4. I enjoyed writing with the AI system
5. I was able to express my creative goals while writing with the AI system
6. The script(s) written with the AI system feel unique
7. I feel I have ownership over the created script(s)
8. I was surprised by the responses from the AI system
9. I'm proud of the final outputs.

We also asked five free-form questions. Two questions aimed at assessing the participants' exposure to AI writing tools (*In a few words: what is your experience in using AI tools for writing for theatre of film or during performance on stage?*) and their industry experience (*In a few words: what is your experience in theatre or film/TV?*). Three more questions gave participants an opportunity to provide developmental feedback about the system:

1. What is one thing that the AI system did well?
2. What is one improvement for the AI system?
3. Please provide any comments, reflections, or open questions that came up for you during the co-authorship session or when answering this survey.