Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals

Piotr W. Mirowski*
piotrmirowski@deepmind.com
DeepMind
London, UK

Kory W. Mathewson*
korymath@deepmind.com
DeepMind
London, UK

Jaylen Pittman†
Stanford University
Stanford, USA
jaylen@stanford.edu

Richard Evans
DeepMind
London, UK

ABSTRACT

Language models are increasingly attracting interest from writers. However, such models 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 critiqued performances of several of the scripts to illustrate how both Dramatron and hierarchical text generation could be useful for human-machine co-creativity. Finally, we discuss the suitability of Dramatron for co-creativity, ethical considerations—including plagiarism and bias—and participatory models for the design and deployment of such tools.

CCS CONCEPTS

• **Applied computing** → **Performing arts**; • **Computing methodologies** → **Natural language generation**; • **Human-centered computing** → **Empirical studies in HCI**; *HCI design and evaluation methods.*

KEYWORDS

natural language generation, natural language evaluation, human-computer interaction, theatre, computational creativity, improvisation, co-creativity

*Both authors contributed equally to this research.
†Work done while at DeepMind.

1 INTRODUCTION

Large language models (LLMs) are becoming more remarkable and useful in co-creative applications, as their ability to generate text improves [12, 25, 57]. While their use is primarily limited to assisting in natural language processing tasks [28, 114], these models show particular promise for automatic story generation [3, 89] as an augmentative tool for human writers. Examples of such creative uses of LLMs include the generation of the script of short film *Sunspring* (2016) [76], *It's No Game* (2017) or *Sollicitors* (2020) [1], improvisational theatre alongside robots by company *Improbotics* (2016) [10, 67, 69, 70], collaborative script writing for theatre play *AI* [109], and *THEaiTRE* company's [93–95, 99] *AI: When a Robot Writes a Play* (2021).

Models able to generate *coherent* stories could be useful for co-writing theatre scripts and screenplays. This is a difficult task for LLMs because the narrative of a script or screenplay must exhibit long-term coherence and reincorporation, and LLMs are limited in their ability to model long-range dependencies (e.g., to reincorporate information from many pages ago). This limitation stems from the context window of LLMs, which today is limited to at most 2048 tokens (i.e. about 1500 words) in state-of-the-art models [78, 86].

In this work, we present **Dramatron**, a system that uses LLMs to generate scripts and screenplays hierarchically through a method we call *hierarchical story generation*. Dramatron leverages the strengths of LLMs and combines custom prompts (see Appendix E) and prompt chaining [121] with structured generation for long range *coherence* across the entire script. This process results in greater story coherence than "flat" sequential text generation. The motivation in producing more *coherent* scripts is to provide co-writers with more connected, usable material. Our method is, in spirit, similar to hierarchical neural story generation [38], but generates scripts that far surpass 1000 words. Hierarchical generation of stories can produce an entire script—sometimes tens of thousands of words—from a single user-provided summary of the central
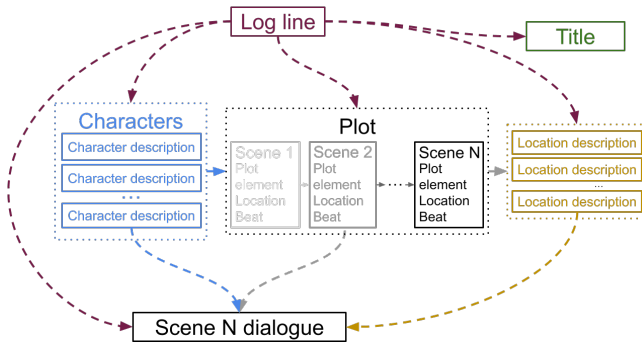
Figure 1: <u>Dramatron's Hierarchical Coherent Story Generation</u>. Dramatron starts from a log line to generate a title and characters. Characters generated are used as prompts to generate a sequence of scene summaries in the plot. Descriptions are subsequently generated for each unique location. Finally, these elements are all combined to generate dialogue for each scene. The arrows in the figure indicate how text generated is used to construct prompts for further LLM text generation.

dramatic conflict, called the <u>*log line*</u> [107]. From the input log line, Dramatron can generate an entire script with a title, list of characters, a plot (i.e. a list of scene summaries with settings and beats), location descriptions, and dialogue (see Fig. 1). <u>The user can intervene at any stage of the hierarchical generation</u>. They can solicit alternative generations, edit and rewrite output text, or continue text generation. In this way, the human interactively <u>co-write</u>s the script. Our methods can be used with state-of-the-art LLMs that accept an input prompt and then predict which tokens come next.

To evaluate Dramatron's usability and capabilities, instead of relying on online crowd-sourced annotation and evaluation from non-expert raters, we engaged <u>15-experts</u> in <u>two-hour</u> long user study sessions to co-write a script alongside Dramatron. The expert playwrights and screenwriters from the theatre and film industry were paid a consulting fee for their engagement. They <u>provided feedback</u> on both the interactive co-authorship process, and artistic opinion and analysis of the outputs co-written with Dramatron. Our inclusive research methodology invited participation from experts during the creative design and development process: their feedback directly led to iterative improvements of the system. We provide a summary of the iterative tool refinement process that emerged from their feedback. A collection of scripts co-written with this process were produced and staged at Edmonton International Fringe Theatre Festival in August 2022. Reflections from the creative team are presented, as are comments from reviewers, as these represent critical reflections on human-machine co-creativity. Widely relevant <u>feedback</u> from the interviews included: participation of experts is critical in tool development, different use-cases demand iterative improvement and development, and different LLMs satisfy different needs from participants. Several participants requested conversational agents to help them rewrite text—which was prescient of recently published language agents trained from human feedback [44, 79].

Our study design and data collection process was validated by an ethical review board external to <u>DeepMind</u>. To the best of our knowledge, this work represents the largest expert user study conducted on co-creative authorship to date.

The paper is structured as follows. Section 2 lists related work on interactive story generation and on its evaluation. Section 3 provides background on LLMs and their limitations. Sec. 3 also gives background on dramatic elements and log lines to justify the rationale behind our hierarchical text generation using LLMs and details on interaction with Dramatron. Section 3.7 provides details on the design of our human co-authorship study, and Section 4 describes evaluation methods for scripts generated by Dramatron. Section 5 presents the major themes summarizing the qualitative interviews. Section 6 covers the quantitative results from the human user study. Section 7 explores the potential impact of these systems on the broader creative community. Finally, the Appendix includes related work on automated story generation (Appendix A), as well as detailed prompt sets (Appendix E), an example of a raw generated script (Appendix F) and four examples of edited scripts (Appendix G).

This paper's key contributions are 1) the introduction of **Dramatron**: a co-writing tool leveraging novel artificial intelligence-powered large language models, 2) the first study leveraging hierarchical story generation for co-writing scripts and screenplays, where authors can generate multiple responses, edit, and then regenerate at various levels of the story hierarchy, and 3) a detailed evaluation of the human-AI co-creative process with industry professionals co-writing alongside Dramatron. This mode of rigorous expert evaluation is broadly applicable to other studies in the field of Human-Computer Interaction.

2 RELATED WORK

2.1 Automated and Crowdsourced Evaluation of the Coherence of Generated Text

Echoing Celikyilmaz et al. [16], one can split evaluation methods into automated or machine-learned metrics, and human-centric evaluation. As detailed in the Appendix A.4, automated and machine-learned metrics typically calculate 1) the similarity between generated and "ground truth" stories [38, 39, 88, 105], the 2) consistency between generated stories and their writing prompts [92, 102], or 3) the diversity of language in the generated output [43, 45, 46, 88, 127]. These metrics were not designed for generated text of the length of a screenplay or theatre script. This motivates a focus on <u>human-centric evaluation</u>, which can be conducted with non-expert crowdworkers or with experts.

We therefore review the <u>limitations</u> of crowdsourced evaluation of the coherence of generated text, and explain why non-expert, crowdsourced evaluation faces crucial quality and bias issues. Inheriting from research standards for large-scale natural language processing tasks, the majority of studies assessing the quality of generations from LLMs evaluate model performance by collecting data from <u>crowdworkers</u> [58, 77, 80, 88, 91, 124]. For instance Yao et al. [124] recruit crowdworkers to evaluate fidelity, coherence, interestingness, and popularity, and Rashkin et al. [88] to evaluate narrative flow and ordering of events. That said, it has been shown that crowdworkers' personal opinions, demographic characteristics

and cognitive biases [36] can affect the quality of crowdsourced annotations in fact-checking [30] or in tasks involving subjective assessments [53]. These issues have led some researchers to try evaluating their models with experts. Karpinska et al. [56] highlight the perils of using crowdsourced workers for evaluating open-ended generated text, because crowdworkers do not read text as carefully as expert teachers. Some studies consulted with expert linguists [31], university students in the sciences [42] or humanities [123], and amateur writers [2, 21, 62, 100, 125]. In one recent study, Calderwood et al. [13] interviewed 4 novelists about their usage of GPT-2 via *Talk To Transformer* and *Write With Transformer* tools (see https://app.inferkit.com/demo), uncovering usages of the "model as antagonist" (i.e., random), for "description creation", as "constraint", or for "unexpected" ideas—some of the themes highlighted by our own study participants. Similarly, a recent user study explored evaluation of writing short stories with a collaborative text editing tool called Wordcraft [125].

2.2 The Process of Writing Theatre and Film Scripts

The processes of authoring theater scripts and screenplays are different, and each author has their own distinct processes. For example, in Section 7.2, we present several unique perspectives on processes from our study participants. That said, previous works have attempted to distill common patterns and lessons for writing theatre scripts [33, 87] and screenplays [106]. Writing can be solo or collaborative, but, even when writing individually, 'the only kind of writing is rewriting' (Hemingway [49]) of previous drafts and revisions. Interactive authorship is the collaboration of several individuals to co-write scripts as often happens in television show creation, where screenplays are written in *writers' rooms*. In interactive authorship settings, writers may use creative inspirations to accelerate and inspire the writing process. The process of generating stories for theatre has a long history of these kinds of technological inspiration [23]. The implicit hypothesis of our work on Dramatron is that interactive authorship alongside an AI-based writing tool could be useful for writers, similar to the contribution of multiple writers in a writer's room working collaboratively on a theatre script or screenplay. For this reason, we want Dramatron to output scripts in a format that is familiar to creative professionals, in script form with, for example, lines of dialogue prefixed by character names.

2.3 Interactive Authorship with Story Generation Models

Automatic (Section A.1), hierarchical (Section A.2) and controllable (Section A.3) story generation are long-standing research areas that we review in the Appendix. Our work builds upon these but focuses on (the evaluation of) interactive authorship with story generation models. Recent work addressed human-AI text revision [31], text summarization [18, 123] and overcoming the writer's block [22, 42, 77, 125]. Notably, *Wordcraft* [125] used an LLM with an editor and an interface that asked to continue the story, asked for details, or suggested to rewrite it, and was used as a tool for idea generation, copy editing or scene interpolation. Our model
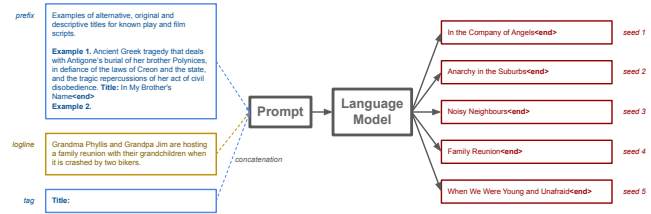

Figure 2: Illustration of the prompting setup for the language model, with user- or Dramatron-generated prompt being concatenated to a *prefix* and decorated with *tags*. Several title outputs are generated for different random seeds, each ending with *tag* <end>.

Dramatron allows some of these writer's interventions within a hierarchical generation structure.

Automatic generation of stories with dialogue has been used to populate digital worlds in video games, interactive narratives, entertainment, virtual worlds [82], artistic performances [48], improvised theatre [10, 70, 75], short film scripts like *Sunspring* in 2016, song music lyrics for musical *Beyond the Fence* in 2016 in London[1] and interactive playwriting for *AI* at the Young Vic in 2021 in London[2]. With the exception of Prague-based company THEaiTRE [93, 95, 99], none of these related works have been used to generate long-range coherent theatre scripts or screenplays. And, unlike Dramatron, none of them used few-shot learning and prompt engineering to prime LLMs for generation.

3 METHODS: CREATIVE TEXT GENERATION USING LARGE LANGUAGE MODELS AND DRAMATRON

3.1 Language Models

Statistical language models (language models, or LMs) model the probability of text *tokens* given a *context* of previous tokens—tokens can be words, characters, or character bi-grams. Using machine learning, LMs are trained on large corpora of text to approximate the conditional probability distribution. LMs can compute the likelihood of a piece of text or to generate new text as the continuation of a text *prompt*. Text generation is probabilistic and involves random *sampling* from the conditional probabilities. Different random seeds result in different random samples. Figure 2 illustrates an example of feeding a text prompt and using the LM to generate different text samples.

In this study, we employed the *Chinchilla* large language model (LLM) [50], represented as a neural network with 70B-parameters and that was trained on 1.4T tokens of the *MassiveText* dataset. As described by Rae et al. [86], that corpora contains 604M *MassiveWeb* documents, 4M *Books*, 361M questions and responses from *C4*, 1.1B *News* articles, 142M *GitHub* code entries, and 6M *Wikipedia* articles. We conducted our study using *Chinchilla 70B* because it outperformed other LLMs on a large collection of language tasks [50]. We

---

[1]https://www.theguardian.com/stage/2016/feb/28/beyond-the-fence-review-computer-created-musical-arts-theatre-london
[2]Review: https://www.theguardian.com/stage/2021/aug/24/rise-of-the-robo-drama-young-vic-creates-new-play-using-artificial-intelligence

have since then successfully tested our approach with Dramatron *as is*, by substituting Chinchilla with alternative LLMs[3], such as GPT-3 `text-davinci-003` [12].

3.2 Limited Contexts of LLMs and Need for Human-in-the-Loop Editing

LLMs can give the impression of coherence within and between paragraphs [7], but have difficulty with long-term semantic coherence due to the restricted size of their context windows. Memory wise, they require $O(n^2)$ (where n is the number of tokens in the context window). Thus, these models currently restrict n to 2048 tokens [12, 78]. Because these models do not achieve long-term semantic coherence, the current state-of-the-art method for longer text generation incorporates a human-in-the-loop who selects from various generations sampled from the model [10, 67, 75]. The human is tasked with achieving long-term semantic coherence: the model generates choices for the next textual snippet, while the human does the selection[4].

3.3 Inspiration from Narrative Elements and Story Arcs to Structure Script Generation

To circumvent the limitations of LLMs and to enable the generation of theatre and film scripts, we take a divide-and-conquer approach: we structure script generation using interdependent narrative elements. We base our approach on Aristotle's *Poetics* [5], which identified 1) **plot** (*mythos*) as the most important part of a tragic story—the other elements being the 2) **theme of the story** (*dianoia*), 3) its **characters** (*ethos*), 4) **dialogues** (*lexis*), as well as melody and spectacle. Thus, we start from a user-defined theme and use Dramatron to help write characters, plot, and dialogue.

Aristotle introduced the well-known plot structure (*beginning*, *middle*, and *end*). Many other dramatic structures followed [84]. For example, variations on the Freytag's pyramid [41], popular in Western storytelling[5] (*Exposition*, *Inciting Incident*, a *Rising Action* composed of a series of *Conflicts*, *Complications*, and *Dilemmas*, *Climax*, *Falling Action*, *Resolution*, and *Dénouement*) or the Hero's Journey or *Monomyth* [15, 113]. More generally, the plot is the coherent sequence of consecutive actions that shape the story. In order to achieve narrative coherence despite the limited context

window of LLMs, we generate a **plot synopsis** in one go, as a sequence of beats.

Oftentimes the seed of narrative inspiration for a screenplay or theatre script is a **log line** [101]. The log line summarizes the story in a few sentences and is often adapted over the course of production due to creative team preferences and cultural references. Log lines typically contain the setting, protagonist, antagonist, a conflict or goal, and sometimes the inciting incident [8], and hence suggest short answers to questions: *Who? What? When and Where? How? Why?* [104] In this work, we find a connection between Artistotle's *theme* and the log line, and we use log lines to start the hierarchical story generation process.

3.4 Hierarchical Language Generation

In this project, we desire a system that provides useful suggestions to the writer, and that avoids generating incoherent dialogue given a theme, plot or choice of characters. For this reason, we aim at building a system that can generate an entire text exhibiting long-term semantic coherence. We hypothesize that if the system can generate an entire script exhibiting reasonable long-term coherence from a single log line without human intervention, then such scripts can serve as drafts for the writer. Our approach to achieve long-term semantic coherence is to generate the story *hierarchically*.

Our narrative generation is conceptually divided into three hierarchical layers of abstraction that loosely interpret the narrative elements of Aristotle's tragedy. The highest layer is the log line (or theme) defined in Section 3.3: a single sentence describing the central dramatic conflict. The middle layer contains character descriptions, a plot outline (a sequence of high-level scene descriptions together with corresponding locations), as well as location descriptions. The bottom layer is the actual character dialogue for the text of the script. In this way, content at each layer is coherent with content in other layers. Note that "coherent" here refers to "forming a unified whole", not assuming any common sense and logical or emotion consistency to the LLM-generated text.

As illustrated on Figure 1, the story is generated top-down [96, 111, 116]. After the human provides the log line, Dramatron generates a list of characters, then a plot, and then descriptions of each location mentioned in the plot. Characters, plot, and location descriptions all meet the specification in the log line, in addition to causal dependencies, enabled by prompt chaining [121] and explained on the diagram of Figure 1. Finally, for each scene in the plot outline, Dramatron generates dialogue satisfying previously generated scene specifications. Resulting dialogues are appended together to generate the final output. This hierarchical generation was designed to enable long-term semantic coherence. A similar albeit inverted, method of recursive task decomposition was used to generate plot summaries [120]. The incorporation of the middle layer, where the plot is summarised as a sequence of abstract scene descriptions, allows the entire plot to fit within the language models' context window. This overcomes prior limitations on long-term semantic coherence. Our method makes it possible for elements in the final scene to provide dramatic closure on elements introduced in the opening scene[6], and for generated stories to follow narrative arcs (see Section 3.3).

---

[3]Anecdotally, and to verify how large the LLM needs to be to work with Dramatron, we evaluated models of different sizes by prompting them with the same log line ("A morality tale where a poor old fisher living on the cold Atlantic coast catches a magical marlin fish, who promises to grant all her wishes. Prompted by her greedy husband, the fisher asks for all fish to be turned to gold"). We generated a list of characters for 50 different random seeds. Models with 1B parameters or more (e.g., Chinchilla 1B and GPT-3 babbage) succeeded in generating a valid list of characters at least 80% of times.
[4]For example: https://theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3
[5]Seminal work in narrative analysis [60, 85, 96] suggests general but mutable structures present in storytelling processes across many linguistic and cultural traditions (e.g. abstract, orientation, complicating action, and coda narrative clauses). The finding that narratives in many societies and in various languages make use of large, modular elements has inspired the use of models of narrative in automated storytelling, as can be seen in fields such as computational narratology [55, 65]. The general structures found in the Structuralist narratology of Propp [85] and the Personal Experience Narratives of Labov and Waletzky [60] are aligned with Freytag's pyramid, which we choose because it is in line with the specific discourse genre of dramatic scripts, and arguably more familiar to the playwrights we engaged with over the course of our study. However, we note that our choice of narrative structure is not universal in scope, and is, in fact, "characteristically Western" [26, 55]. Alternative story shapes, or "story grammars" [97] are possible [6, 17].

[6]See e.g. Chekhov's gun [27].

3.5 Generation of Narrative Elements using Prompt Engineering

Dramatron uses several hard-coded prompts (i.e. input prefixes) to guide the large language model. Prompt engineering is a common way that users control or influence LLMs [12]. Each prompt has a few examples of desirable outputs. These are included in the prefix and adaptation to only a handful of examples is sometimes referred to as few-shot learning. As illustrated in Figure 2, prompts are concatenated with user-supplied inputs and/or outputs of previous LLM generations. This method is called prompt chaining [121], which is a type of algorithmic prompting [24]. At lower levels of the hierarchy (see Fig. 1), prompts are chained together with outputs from higher levels of the hierarchy.

In this work, we primarily used two sets of prompts: one based on Ancient Greek tragedy *Medea* by Euripides, and one based on science-fiction films. For Dramatron, each prompt set is composed of: 1) title prompt, 2) character description prompt, 3) plot prompt, 4) location description prompt, 5) and dialogue prompt. For both of these prompt sets, we chose dialogues that were in the public domain. The *Medea* plot was chosen because it corresponded to the Freytag Pyramid narrative arc, whereas the sci-fi plot followed an alternative narrative structure, the Hero's Journey. Each prompt is detailed briefly below to give a sense of how they are engineered; additional details are in Appendix E. We engineered the hierarchical prompt structure and the few-shot examples through iterative prompting and testing. We chose and reworded the prompts so that the outputs of the LLM could be parsed into various components (title, character description, etc.) with high probability of success.

The **Title Prompt** is used to generate titles from a log line. A simplified title prompt, a user-provided log line, and randomly sampled titles are shown in Figure 2. It shows a prefix with an instruction (Examples of alternative, original and descriptive titles for known play and film scripts.) and an example (Example 1. Ancient Greek tragedy [...]. Title: In My Brother's Name<end>). The prefix finishes with: Example 2. A user-input log line (e.g., Grandma Phyllis and Grandpa Jim [...]) is concatenated to that prefix, as well as the tag Title:, which encourages the LLM to generate a title that matches the log line. From a few examples, the LLM has "learned" to generate a related title and terminate tag <end>. The **Character Description Prompt** is used to generate character names and descriptions from a log line. The **Plot Outline Prompt** is used to turn a log line and list of characters into a plot. This prompt encourages the few-shot language model to transform a single sentence log line into a sequence of scene descriptions. Each scene is highly compressed, describing only the short name of the location, the narrative element identifying the position of the scene in the narrative arc (see Sec. 3.3), and a summary of what the characters are doing and saying, often called a narrative *beat*[71] As a note, the prompt imposes a strong representational constraint on the way Dramatron represents a scene; each scene is composed of a location, narrative element identifier, and beat. The **Location Description Prompt** is used to generate a detailed scenic description from a place name and a log line. Finally, the **Dialogue Prompt** is used to turn a beat (i.e., the scene summary), scene location description, description of each of the characters involved in the scene, and the log line (for story

consistency), into dialogue. This prompt uses scene information generated for both the current and previous scenes.

3.6 Interactive Writing with Dramatron

As described above, with just few-shot (i.e., 1-4) prompts and the user input log line, we leverage trained LLMs to generate complete scripts and screenplays. Appendix F shows an example of raw generated output. That said, we designed Dramatron for *interactive co-writing*, as an augmentative tool for human writers. Dramatron was built with a hypothesis in mind: that interactive authorship alongside a tool would be useful for writers. While we did not co-design the tool with experts, whenever possible during the study we iterated and improved Dramatron based on their implementable suggestions and feedback (see Table 1). Co-authorship with Dramatron proceeds as follows: a writer starts with a log line, and then generates the title, characters, plot outline, location descriptions and each scene's dialogue step-by-step. At each step, the writer can take one, or several, of the following operations, as many times as desired:

- Generate a new suggestion (i.e., run the LLM again with the same prompt).
- Continue generating from the end of the previous generation, similarly to typical "flat" LLM generation.
- Manually edit some or all of the output generated by the LLM.

The writer can furthermore perform these operations by stepping forward and back in the Dramatron hierarchy. For example, they could: 1) generate a title, 2) generate a new title, 3) edit the title, 4) generate a list of characters, 5) edit the characters by removing one character and changing the description of another, 6) generate a plot outline, 7) edit the plot by removing part of the narrative arc, 8) generate a continuation of that edited plot, 9) go back and rewrite the log line, etc. This co-writing approach allows the human and Dramatron to both contribute to the authorship of a script. Following these operations, the human author could further edit and format to finalize a script. Appendix G shows examples of human-edited scripts.

3.7 Implementation Details

The code of Dramatron is implemented in Python and the user-facing interface was implemented in a Google Colab[7] with text widgets, allowing interactive editing. Figure 3 shows a typical implementation of Dramatron, with examples of generated title (top left), characters and plot (top right). It also shows generated and re-generated place descriptions, and generated, edited and re-generated dialogue—compare bottom left with bottom right panels.

There are several special markers we use for script generation: <end> represents the end of full sequence generation token, and <stop> is a token used to mark the end of a generated line. For a given prompt (see next Sec. 3.5) fed to the LLM, up to 511 text tokens were sampled. We used Nucleus sampling [51] to encourage diverse outputs, sampling tokens from the top 0.9 probability mass, and with a softmax temperature of 1.0. Finally, in order to

---

[7]https://colab.research.google.com/

Choose a prompt prefix set

Choose among prefix sets

[203] prefix_set: medea_prefixes

Show code

Loaded medea_prefixes.

Interactive story generation

Define the Log line

[204] Log lines are one- or two-sentence summaries of the action. They typically contain the **setting**, **protagonist**, **antagonist**, a **conflict** or **goal** and sometimes the **inciting incident**.

logline: "A morality tale where a poor old fisher living on the cold Atlantic coast catches a magical marlin fish, who promises to g"

Show code

A morality tale where a poor old fisher living on the cold Atlantic coast catches a magical marlin fish, who promises to grant all her wishes. Prompted by her greedy husband, the fisher asks for all fish to be turned to gold.

Create the story generator

[205] WARNING: running this cell resets the generator and all its outputs.

Show code

New Dramatron generator created.

Usage

Running each cell displays the UI and generates the first suggestion.

Once you have run the cell once, you can do the following actions *without re-running the cell*:

- Click on `Generate new` to **generate a new suggestion**.
- Once you generated a suggestion, you can **edit the suggestion** in the text box. It is automatically saved and used in the next step.
- Click `Continue generating` to **add to the suggestion**.
- Once you have generated multiple suggestions, you can **navigate through the suggestion history** by clicking on `Previous` and `Next`.

Generate a Title

[206] Show code

Generate new

The Golden Marlin

Generate Characters

[208] Show code

Generate new | Continue generation | Previous | Next

```
<character> The old fisher <description> The old fisher is the heroine of the story. An impoverished old woman, the old
fisher must hunt for fish to survive, and is willing to do anything for her husband. She captures a magical marlin fish,
who promises to grant all her wishes. Prompted by her husband, the old fisher asks for all fish to be turned to gold.
Once granted, she is unable to eat and she ends up losing all her gold to a greedy banker who puts a contract on her
house. At the end of the tale, the old fisher is living in a shack, and her only companion is her cat. <stop>

<character> The old fisher's husband <description> The old fisher's husband is the villain of the story. A greedy old
man, the old fisher's husband pressures his wife into asking for all fish to be turned to gold. Once granted, he is
unable to eat, and he ends up selling his gold to a greedy banker. Once his gold is gone, he abandons the old fisher and
their cat. <stop>

<character> The banker <description> The banker is the secondary villain of the story. A greedy man, the banker tricks
the old fisher's husband into trading the old fisher's gold for money that is only good for one day. After all the gold
is gone, the banker puts a contract on the old fisher's house. <stop>

<character> The cat <description> The cat is the animal companion of the old fisher. A magical cat, he tricks the greedy
banker into buying a fish dinner, the last meal for the old fisher's husband and his family. The cat then leaves with
the old fisher, after the old fisher's husband abandons her. <stop>

<character> The marlin <description> The marlin is a magical fish. The marlin makes a deal with the old fisher, and is
then turned into gold. <stop>
```

Generate a Plot Synopsis (sequence of **Scenes**)

[210] Show code

Generate new | Continue generation | Previous | Next

```
Place:  On an Atlantic shore.
Plot element:  Exposition.
Beat:  The old fisher recalls her story of catching a magical marlin.

Place:  The old fisher's house.
Plot element:  Inciting Incident.
Beat:  After catching the marlin, the old fisher and her husband take it into their home.

Place:  The old fisher's house.
Plot element:  Conflict.
Beat:  The old fisher and her husband argue about what to wish for. The husband persuades the old fisher to wish for
all fish to be turned to gold.

Place:  On the Atlantic shore.
Plot element:  Dilemma.
Beat:  The old fisher catches another fish, but discovers that this fish is now made of gold. The old fisher decides
not to eat it and throws it back.

Place:  The old fisher's house.
Plot element:  Climax.
Beat:  The greedy banker arrives at the old fisher's house and offers to buy all her gold for a small sum of cash.

Place:  The old fisher's house.
Plot element:  Rising Action.
Beat:  The old fisher's husband, desperate to eat, sells the last of the gold to the greedy banker, who tells him that
the money he is given is only good for one day.

Place:  On the Atlantic shore.
Plot element:  Falling Action.
Beat:  The old fisher catches a fish, but it too is made of gold. The old fisher realizes that the last of their gold
is gone and they are left with nothing but a contract on their house.

Place:  The old fisher's house.
Plot element:  Resolution.
Beat:  The greedy banker returns to collect the debt. The old fisher's husband abandons her and their cat and leaves.

Place:  The old fisher's house.
Plot element:  Dénouement.
```

Generate Place Descriptions

[214] This cell generates a description for each **Place** name in the **Plot Synopsis**. If you edit place names in the **Plot Synopsis** *after* having run this cell, you will need to re-run this cell to update **Places**.

Show code

Generate new

The old fisher's house.

```
The house is small, on the cliffs overlooking the ocean. Inside it's cold and damp, with a few sticks of furniture and a
bed for the fisher and her husband. The roof is made of thatched reeds. There's a window, but it is blocked by old and
faded curtains.
```

On the Atlantic shore.

```
The waves of the Atlantic crash on the cold, rocky beach.
```

Generate Dialogues

[212] Show code

scene ——○—— 1 | Generate new | Continue generation | Previous | Next

```
MAN
You were lucky to get anything.

OLD FISHER
It was a marlin and a huge one and he had
the hook in the top of his mouth. You know how they
are with that bone in their nose and that
plate in the forehead.

MAN
Bigger than the one we saw?

OLD FISHER
Oh, much bigger.

MAN
He didn't weigh as much though, did he?

OLD FISHER
Oh, no. This was much bigger. It was a
strange fish.

MAN
Was it the same kind as the other one?

OLD FISHER
It was the same. There isn't any other kind
of a marlin out in that stream.

MAN
But was he as big as that one we saw?

OLD FISHER
He was a little bigger. You know those
market-sized marlins aren't really big.

MAN
```

Render the script

Generate Place Descriptions

[214] This cell generates a description for each **Place** name in the **Plot Synopsis**. If you edit place names in the **Plot Synopsis** *after* having run this cell, you will need to re-run this cell to update **Places**.

Show code

Generate new

The old fisher's house.

```
The old fisher's house is a modest wooden building with a tiled roof on a cold Atlantic coast. The house is surrounded
by a small patch of grass. The door to the house is closed.
```

On the Atlantic shore.

```
The setting is the shore of the Atlantic Ocean. The air is warm and the sea is blue. A storm has left the shore rocky
and empty.
```

Generate Dialogues

[212] Show code

scene ——○—— 1 | Generate new | Continue generation | Previous | Next

```
MAN
You were lucky to get anything.

OLD FISHER
It was a marlin and a huge one and he had
the hook in the top of his mouth. You know how they
are with that bone in their nose and that
plate in the forehead.

MAN
Bigger than the one we saw?

OLD FISHER
It was an exceptional fish. But this was
a really big fish and it took much
longer to bring it in. Remember what
I told you. I was very tired and
when I came in, I laid the rod down
and sat on the stern with my head on
my knees. Then I heard the fish
when he jumped. I thought the noise
was the tar cracking on the roof of
the cabin. So I jumped up to see what
was the matter and he dropped back.
Then I saw he was jumping. My God,
did he jump!

MAN
What did he look like?

OLD FISHER
He was jumping straight up and straight
down and when he dropped back into the
water you could see his whole body.
He was bright in the sun but you
could see strips that were purple
and strips that were silver and I
remember his eye when he arched
```

Render the script

Figure 3: Implementation of Dramatron as a Google Colab, showing the selection of a prompt prefix set and of a log line, instructions, generation and editing of title, characters, plot synopsis, location descriptions and dialogue (before and after editing and re-generation).

reduce loops in the generation of dialogue, we implemented a simple detector that cuts the generated text into blocks (delimited by 2 blank lines) and counts how many times each block appears in one single generation. Beyond a fixed threshold (e.g., 3 times), the LLM generation is ran again by sampling tokens using a different seed in the random number generator.

4 EVALUATION METHODS FOR SCRIPTS GENERATED BY DRAMATRON

In Section 2.1, we reviewed the limitations of crowdsourced evaluation of the coherence of generated text, and explained why non-expert, crowdsourced evaluation faced crucial quality and bias issues. Given those issues, we believe that crowdsourcing are not an effective approach to evaluating screenplays and theatre scripts co-written with language models. Thus, in a departure from crowdsourced evaluations, we engage 15 experts—theatre and film industry professionals—who have both an experience in using AI writing tools and who have worked in TV, film or theatre in one of these capacities: writer, actor, director, or producer. These experts participate in a 2-hour session wherein they co-write a screenplay or theatre script alongside Dramatron. Most were able to co-write a full script and the open discussion interview in the allotted 2-hours. One participant had a follow-up session on the next day that allowed them to work on their script offline. Another participant had 22.5 hours of interaction in total, split over 5 sessions and spread over two months, where they co-wrote 5 scripts that were later staged.

During the writing sessions, one interviewer (the operator) was operating the user interface to Dramatron, typing text and making edits suggested by the participant, and explaining the generation process, while another interviewer was asking questions about the interaction and taking notes. The interface to Dramatron was shared via screen. Language model-specific parameters (such as sample length or sampling temperature) were not shown to the participant. The participant had a choice of prompt set (*Medea* with Freytag's Pyramid or sci-fi with Hero's Journey) and told the operator what button to click or text to type. Interviews are analysed in Section 5, and following the interactive sessions, the participants were asked a series of questions adapted from [125] and detailed in Section 6. Each question was answered on a 5-point Likert-type scale, using questions adapted from [22, 58, 80, 108, 125]. As another quantitative evaluation, we track writer modifications to generated sentences [91]. This allows for comparison of Dramatron generations pre- and post-human edits. We track absolute and relative word edit distance, to assess whether and how much the writer adds to or removes from the output suggestions. We also track a Jaccard similarity-based metric on words, to quantify how similar is the draft after edits to the original suggestion. Our objective is to assess whether Dramatron "contributes new ideas to writing", or "merely expand[s] on [the writer's] existing ideas" [62]. We do not evaluate Dramatron outputs for grammatical correctness as [80], as the few errors made by the LLM can be fixed by the human co-writer.

We compensate experts for the co-writing and interview sessions at 100 GBP per hour. Our study and data collection process was validated and approved by HuBREC (Human Behavioral Research

Ethics Committee), which is a research ethics committee run within Deepmind which includes and is chaired by academics from outside the company.

5 PARTICIPANT INTERVIEWS

Throughout our interviews with the 15 participants (anonymised as p1, p2, etc.), we collected qualitative feedback on co-writing with Dramatron. Feedback was analysed first separately by two investigators to extract different themes, then their respective analyses were merged and jointly summarised into seven major themes. Each theme is presented alongside supporting quotes from participant interviews.

(1) Positive comments about Dramatron focused on: hierarchical generation that lets the writer work on the narrative arc, the possibility either to co-author interactively or to simply let the system generate, and the potential of the output script to serve as source material for the human writer (Section 5.1).

(2) Participants identified inspiration, world building, and content generation as potential writing applications for Dramatron, and saw it as possible tool for literary analysis (Section 5.2).

(3) Participants noticed various biases embedded in the language model (Section 5.3).

(4) Some writers were interested by the involuntary glitch aesthetic and failure modes of Dramatron, such as repetition and dialogue loops (Section 5.4).

(5) Unsurprisingly, participants noticed logical gaps in storytelling, lack of common sense, nuance and subtext, which were manifest in the lack of motivation for the characters (Section 5.5).

(6) Structural criticism focused on the need to come up with a log line, as well as on the inconsistencies between consecutive scenes due to parallel dialogue generation (Section 5.6).

(7) Participants were engaged with the tool and eager to provide suggestions for improvement (Section 5.7).

All participants had prior exposure to AI writing tools. Some of their comments seemed to be formulated in response to the use of Dramatron by playwrights and screenwriters—we mark those in text with modifier (*script writing*). Other sections may contain learnings for AI-based storytelling in general—we mark these as (*storytelling*). Our evaluation is structured along two orthogonal axes: Sections 5.1 through 5.7 focus on the user experience of Dramatron, whereas Section 5.9 treats the output of Dramatron as working material for a production company and thus evaluates the intrinsic quality of the script.

5.1 Positive Comments about Dramatron

5.1.1 Praise for the interactive hierarchical generation in Dramatron (storytelling). All participants but p4 and p5 (who preferred a non-linear writing workflow) were enthusiastic about the interactive hierarchical generation. "Once I see this, I know the shape of the series. I know the way that the story unfolds. I can see the narrative more clearly [...] I like this approach of making it a log line and then packing the detail inside it. You are planting a seed of an idea and it is putting meat on the bones" (p13). "All of it is quite consistent, symbolically consistent and coherent and relates to the state of affairs of the state of the play [...] There is lots of emotion and content

about relationships in some of the generations" (p8). "In terms of the interactive co-authorship process, I think it is great [...] " (p9). "What I like about the hierarchy is that you can do as much humaning as you want at any level" (p2). "In working with the machine I can see the content a little more clearly. As there is specificity, character arcs, then I can see how the story comes together [...] This [hierarchical generation] really felt so much cleaner than the process [GPT-2 or GPT-3 with flat prompting] I was using" (p15). "Let's try more! God, you could just waste your time doing this" (p3). Participants p1, p6 and p3 further noted how such hierarchical generation helped with dialogue: "there is good content from any generation" (p1) and (referring to one of the generations) "You got some big profound discussions in it. I am impressed with that one" (p3).

5.1.2 Ease of use of Dramatron's UI and seed-based generation. Participant p13 liked the user experience of interactive, step-by-step generation of title, characters and plot, whereas p10 thought that "interaction seemed simpler when the whole script was generated ahead of time rather than editing it". Participant p1 tried and discussed three different modes of script generation: 1) interactive co-authorship, 2) modifying the output from one fully automated generation, and 3) curating and modifying outputs from 3-4 generations. The benefits of running multiple generations included having "lots of material", allowing to "pull good ideas", "cherry-picking", "more interpretations and artistic freedom" but "requires more massaging on my end" and "word crafting to make it flow" (p1). Participant p1 developed a workflow for co-generating a script that included editing lists of characters and editing the log line to add more "characters that we know about", giving the characters status and names, adding them to the plot's beats. When crafting the log line, p1 wanted to imply high stakes and "stay with humanoid characters: non-human characters take us to the Theatre of the Absurd, to the Surreal, to Magical Realism", and they wanted log-lines that situated the story in realism "to meet the audiences expectations" and "set things at a specific location".

5.1.3 About the potential for the script to be staged after editing (script writing). Several participants (p6, p9, p11, p13, p15) highlighted the potential for the script to be staged after editing: "a rough draft, would need to work a lot with it [but] it could be helpful and staged, definitely" (p6), "It gets me thinking about how you can make a full show with a single idea" (p11) and "You know, with a bit of editing, I could take that to Netflix: just need to finesse it a little bit" (p9). Participant p1 staged several scripts generated with Dramatron (see Section 5.9).

5.2 Potential Uses of the System

5.2.1 Inspiration for the Writer (storytelling). All participants found Dramatron useful for getting inspiration: "this is perfect for writers' block" (p13), "I can see it being very helpful, if you are stuck" (p4, p5), "more in depth than the writers' unblocking prompts website" (p3). Dramatron was described as a tool that indirectly stimulates the playwright's creativity: "I like what happens in my brain when I read some outputs of the model. I got an idea for the rest of the story" (p6), "It is about me discovering what will translate from what it gives me" (p10), or that directly gives actionable suggestions:

"Here is a concept; it puts meat on the bones, and then you trim the fat by going back and forth" (p13). Glitches and language model limitations can be subverted for inspiration, in particular when the script is staged: "mistakes are gifts that we can leave for the improvisers" (p1).

5.2.2 Generation of Alternative Choices and World Building (script writing). More than merely providing a creative spark for the main story, the model can be employed to populate the universe of the story: "If I was going to use this to write a script, I'd use it to generate characters to see if it generated things I hadn't thought about. Or relationships I hadn't thought about" (p15). Dramatron for exploration: "I would go with the suggestion that is further away from what I would have suggested because I already know what is in my head and I want to know what the machine would do" (p12).

5.2.3 Using the System for Learning and Analysis (storytelling). By prompting the system, writers could indirectly search the language model for literary styles and elements: "Even if I were not writing, it does a wonderful job of collecting what is in the literature" (p10) or even hypothetically search withing their own output: "I would be very interested in feeding everything I ever wrote and then getting it to generate script in my voice and style" (p4, p5). Learning could also happen by analysing how to improve Dramatron's outputs: "For me, as a playwright, the interesting thing about working with this technology is thinking about how I would edit it. For instance: What would this look like on stage?" (p8).

5.2.4 Content Generation (script writing). Beyond inspiration, several participants were interested by the co-writing potential of Dramatron, and thought it could provide them with material. "One of the big sticking points of playwriting is getting words on the page. This helps with that step" (p8). "I would use this tool to fix (screenwriting) projects that might be dead" (p14). "This is a rich tool for basically everything. I have done devised creation. There are methods that you can use to generate text, where you pull songs, scripts, or news articles, then chop and paste them down. This reminds me of Dadaist text generation" (p11). "Practically, it might impact the economics of writing if longer running series could be augmented by such writing systems. It might be useful on long-running series, where you have a writers room" (p4, p5).

5.2.5 Potential of AI as Tool for TV Screenwriting (script writing). Some participants suggested this tool could be employed in a TV writers' room, to help with writing formulaic scripts. "If you were able to make an AI to synopsize scripts effectively, you would be valuable to the studio" (p14). "It is like having a very good dramaturge" (p10). "AI can come up with 5 scripts in 5 minutes" (p9). "Which part of the process is this tool relevant for? Formulaic TV series" (p4, p5). "I wouldn't use it for writing a straight play" (p11).

5.3 Stereotypes

5.3.1 The system outputs are too literal and predictable. Some participants found the character "relationships so tight and prescriptive" (p4, p5); if a character has "a noble endeavour, it will be stated in the dialogue" (p4, p5), and that characters were given "silly" and "on

the nose, pun names" (p2). Similarly, the title generation "does what it says on the tin" (p15), and "can be overly descriptive sometimes: the director could make decisions" (p8). One commented, "this is a thing that my students would do" (p8). There were some positive aspects to such a predictable system: "interpersonal relationships created here are classic tropes that keep the audience interested" (p3) and "there is interest in generating outputs from the system for content that already exists: actual titles are fun to compare against" (p14).

5.3.2 The system outputs can be problematic, stereotypical, and biased. Participant p9 wondered "What cultures and languages the books come?" whereas many participants noticed gender biases and ageism in the system outputs. "I am less sexist than the computer" (p3). "The protagonists are both male characters, and all of the supporting characters are female" (p4, p5). "The female lead is defined by their relationship to the other characters: it is a typical thing in plays that the women characters don't have a lot of information about them" (p11). "She is always upset and doesn't have wants (like the male characters) [...] Actually lots of the content [...] is misogynistic and patriarchal" (p8). This problem raised the issue of coping strategies or cultural appropriation: "if we gave GPT-2 some character names, it could come up with bigoted characters: [we] went with more made up names, not gender specific, not ethnicity-specific" (p13) and "there is an ethical question about using AI for a group of theatre makers: the AI throws us a topic, or relation that is unrelated to our lived experience and we are compelled to *Yes, and* the offers" (p4, p5). We discuss ethical issues raised in discussion by participants in greater detail in Section 7.3.

5.4 Glitches (storytelling)

5.4.1 Participants embrace unexpected outputs from the system. Participant 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).

5.4.2 The system often enters in generation loops. All participants noticed how the system could enter generation loops: "I would probably cut a lot of it" (p6) or "a whole scene about a boiler being broken: yeah" (p8). They sometimes found positive aspects to such loops: "It is a silly conversation. It is a little repetitive. I like it." (p6), "repetition leaves room for subtext" (p12) and enjoyed the glitches (p4, p5) or even made parallels with existing work (p3).

5.5 Fundamental Limitations of the Language Model and of Dramatron

5.5.1 Lack of consistency and of long-term coherence. "Keeping dialogue character-based and consistent is most important [...] There is still some difficulty in getting it to stay on track with the context." (p15). "I want the characters to be more consistent within themselves" (p12). "There is a bit of confusion in the logic, gaps in logic [...] It looks like postmodern theatre [...] But in terms of [a play with a given] genre, that has a plot to follow, it is getting confusing"

(p11). Participant 7 "wants to add some stitching between the beats to make them narratively make sense".

5.5.2 Lack of common sense and embodiment. Participant 8 observed that "There are things that it is hard to show on stage – such as a cat. The system doesn't have an awareness of what is stageable and not stageable" and p9 noted that when "interfacing with a story telling AI, the input space is constrained".

5.5.3 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). "With dialogue in plays, you have to ask yourself two questions: 1) Do people actually speak like that? 2) Are actors attracted to these lines and are these appealing lines to play?" (p7) "Playwriting is about realistic dialogue... all of the things around subtext. [...] Show, not tell: here we are just telling. Just like in improv: 'do not mention the thing'. The element in the log line became the central bit in the generation, and that was repetitive" (p8). Participant 14 concluded that "AI will never write Casablanca, or A Wonderful Life. It might be able to write genre boxed storytelling".

5.5.4 Lack of a motivation for the characters (storytelling). "The stories do not finish. The character journeys are not complete. There is perhaps something missing in the character background [...] Where is the emotional motivation, stuff that might exist in the backstory and not exist in the script?" (p14). "On the first go-through, you are looking for the goal of the protagonist, and impediment for that drive. What is my character doing, and what do they want? If this was given to an actor they are going to struggle with the first thing to do, which is to find the needs and the wants of the character and then to personalise it" (p9). "My students do this: a character comes into play and says right what they want." (p8). "The conflict should be something inside the character" (p6). "Why do people not say what they mean? It is because we have societal understanding, but sometimes get lost in translation" (p3).

5.6 Structural Problems of Dramatron (script writing)

5.6.1 Difficulty caused by the need to come up with the log line to condition all the generation (script writing). For participant 12, it was difficult to come up with a log line, and the process seemed precious. "Coming up with the first prompt takes a little bit of back and forth" (p11). "Packing the action into the log line: this is a panic moment for the writer, because they want to add everything meaningful into the script. [...] It is all about the witty premise. The system that you have right now is somewhat about wit. There is a need for the log line to hold some kind of wit" (p13). "Does [the log line] have to have a character name? (p4, p5). "The log line is not a closed synopsis. It is less descriptive and more prescriptive. The art of log lines is about how short you can make it so that [the producers] read the rest of your material" (p14).

5.6.2 Structural criticism of log line-based conditioning of the whole generation (script writing). "Generally the way that I work, I am clear what I want to say about the world – what I think about the

world. The vehicles, or the characters, or the arc is not clear. This looks like a collection of scenes that logically follow one to the next. But, the core idea of the thing to say [is missing]" (p4, p5). "If I could program something to write a script, I wouldn't start with a log line. You can also consider starting with a character and an obstacle in the way of that character" (p9).

5.6.3 Negative consequence of Dramatron's design choice: parallel dialogue generation (script writing). "From the scene beats, it has no idea of what the previous dialogue contained. Then to see the dialogue not be consistent is jarring" (p1). "I wonder if there is a problem in importing the previous beat into the scene [...] Paying attention to the consistency in the beats, helps with the consistency of the dialogue generated" (p12). Upon learning that scene dialogue was generated in parallel for each scene, Participant 2 commented: "If it didn't read its last scene, how can you get the last scene into the next generation? Generation of these scripts could be significantly benefited from attending to the previous scene's dialogue".

5.7 Suggested Improvements to Dramatron (script writing)

Modeling characters and their relationships was a recurrent theme: "can we make the system relationship-driven?" (p12), "where does status belong in character building?" (p12), "could we generate the stem of a character and then complete it?" (p15). Participant 12 suggested: "as an author, I would build a social graph of the characters relations". Answering the question "How do you get the system to know where the scene should start and end?" (p15), three participants (p8, p13, p15) suggested fitting a narrative arc within each scene.

Several participants wanted to be able to query and dialogue with the writing model: "Have you engaged [the AI system] by trying to give it notes?" (p2) to allow it to learn about the world: "How does world building happen? Maybe the model needs to know the Ws of Stella Adler [(Who? What? Where? Why? How? etc.)] Can you get the system to answer these questions?" (p9), or to allow rewriting and reformulation: "can we ask the system to re-write with a style or context?" (p8). As p10 reiterates, iterative rewriting was a desired workflow: "I am less interested in shaping [the narrative], rather than seeing what it is saying, and refining it to see what it says, and then refining it again. A playwright has to see the play spoken before making cuts."

Finally, p4 and p5 astutely observed that "there has been a push away from systems of Western dramaturgy, so in terms of making this most useful for the future, it might be helpful to consider how it might be used within the context of other contemporary writing"— suggesting alternative narrative structures and elements—"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".

5.8 Incremental Tool Improvement

As detailed in Section 5.7, the participants were engaged and provided constructive feedback about Dramatron. As one of the participants in the study remarked: "the system is so adaptable, it can change with our feedback and tweaks". This sort of understanding of the systems modifiability empowered those that interacted with it to more freely suggest changes, knowing that they could

be incorporated. In this way, the system positively benefited and evolved over the course of the participant study.

Over the course of the interviews, we incorporate the feedback we could by making small, incremental changes to the prompt prefix sets of Dramatron. Table 1 summarizes changes made as a direct result of participant's feedback. This sort of participatory design and development is critical for creative tool generation as the feedback from users can be directly incorporated to improve the system for the next interaction. This is made possible via the modular design of the system, the lightweight prompt-based interactions, and the flexibility afforded by Dramatron. This participation also inspires participants to explore related, connected, creative ideas. For example Fig. 4 (LEFT) shows concept art for a narrative test of virtual actors interpreting a co-written script.

5.9 Staging and Evaluating Productions of Scripts Co-written by Dramatron

Creative writing for theatre is fundamentally interactive: not just between collaborating storytellers, but between storytellers and the audience. For this reason, we evaluated how scripts co-written with Dramatron could be produced on the theatre stage. In this section, we describe staging details and report evaluative reflections from both the creative team and two professional theatre reviewers.

Five scripts co-written with Dramatron were staged in public performances in August 2022. The show's run was titled *Plays By Bots* and ran 7 performances over two weeks (see an image from the production on Fig. 4). In each show, different casts would act out one of the plays from the co-writing experiments. The plays span different genres, styles, characters, and storylines. The scripts were brought to life by a cast of 4-6 experienced improvisers and actors. The first-half of each script was given to each of the cast members in a sealed envelope. Only when the show began were they allowed to open the script, and then they commenced performance by reading it live in front of the audience. Once the script ran out, the actors improvised the ending, based on the context and story set out by the script[8]. During each show's performance, the director and co-writer (participant p1 from above) introduced the project to the audience and explained that they co-wrote and edited the script using Dramatron.

There were two reviews written about the production of *Plays By Bots* at the festival. One of the reviews noted that the performance "proves that artificial intelligence can in fact write a hit Fringe play". The reviewer also noted that the success of the performance was due to both the Dramatron system and the human actors, especially one performance 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 review was also positive. With a hint of incredulity, the reviewer complimented the abilities of Dramatron. The 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," and "the human actors continue to capture the playwright bot's tone". The reviewer also expressed surprise at the ability of the system to create a play that hangs

---

[8]Video of performance shared upon acceptance.

Participants	Prompt set	Modifications of Dramatron subsequent to the sessions
p1 (session 1)	Medea	Pause after each scene dialogue generation to enable review by the writer.
p1 (session 2)	Medea	Simplify plot prefixes to location, narrative element, beat, list of characters.
		Use an edited script generated by Dramatron ("The Cave") as prefixes.
p1 (session 3)	The Cave	—
p2	The Cave	Continue the generation of dialogue.
		Enable editing of generated dialogue.
		Rewrite the title prefix to avoid plagiarism.
p3, p4, p5, p6	Medea	Make scene dialogue prefix depend on current and previous scenes' beat summaries.
		Continue the generation of characters and plot.
p7, p8	Medea	Remove the list of characters from plot prefixes.
		Show multiple seeds for title outputs.
		Sample new generation in case a loop is detected in generated dialogue.
p9, p12-p15	Medea	—
p10, p11	Sci-Fi	—

Table 1: Summary of incremental tool improvements following participant sessions.

together and creates a world. They further noted that some lines from Dramatron are so funny they were reprised later in the show once the human actors were improvising.

Discussions amongst the creative team compliment the reviewers and provide insights on how professional actors and improvisers found working with scripts co-written by Dramatron. Post-show discussions were facilitated and relayed to us by the director (p1 above). Four key themes emerged through these discussions which echo the themes presented earlier in Section 5. Specifically, the system has a distinct glitch style, generated text can be repetitive and fun to work with. As well, the team attributed agency to the system, and had expectations of the systems capabilities. Finally, the prevailing feedback from the creative team was that participating in the production was fun! This enthusiasm and these motivating reflections from the creative team reflect the usefulness of co-written scripts for theatre production and collaboration. Additional reflections and supporting quotes are included in Appendix B.

6 PARTICIPANT SURVEYS

6.1 Qualitative and Quantitative Results on Participant Surveys

Of the total 15 study participants, 13 provided responses on our post-session feedback form. 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. These questions are slightly adapted from Yuan et al. [125] and Stevenson et al. [108]: (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*, and (9) *I'm proud of the final outputs*. We also asked five free-form questions. Two questions aimed at assessing

Figure 4: (TOP): Concept art used for a narrative test prototype of a virtual actor interpretation of the script *Darren just can't handle the temperature of his soup* created by Participant p13. Used with Permission from Transitional Forms. (BOTTOM): Photo of human actors interpreting the script *Cars: The Day The Earth Stood Still* as part of the *Plays By Bots* series of performances of scripts co-written with Dramatron and director Participant p1. Used with Permission from Rapid Fire Theatre.

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?*). We used
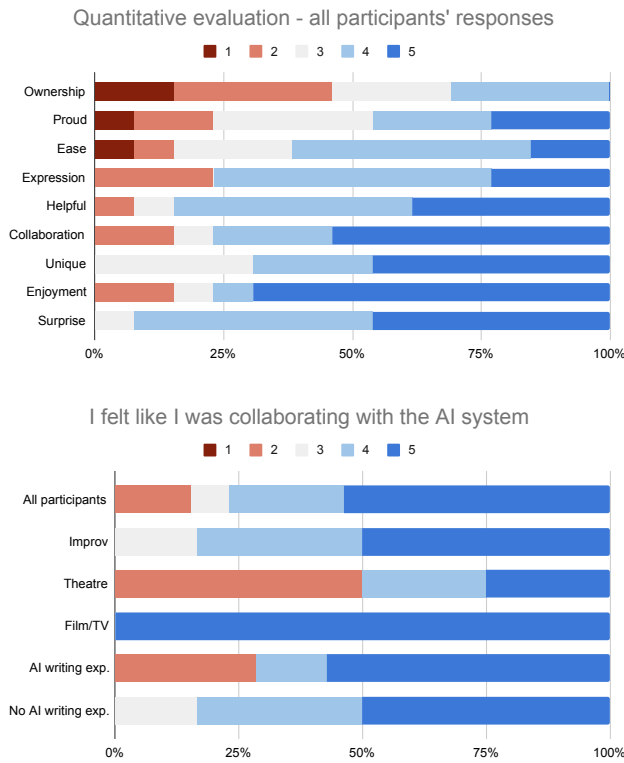
Figure 5: Participants responses to the quantitative evaluation, on a Likert scale from 1 (strongly disagree) to 5 (strongly agree)

these questions to manually define a binary indicator variable (*Has experience of AI writing tools*) and a three-class category for their primary domain of expertise (*Improvisation*, *Scripted Theatre* and *Film/TV*). Three more questions gave participants an opportunity to provide developmental feedback about the system: *What is one thing that the AI system did well?*, *What is one improvement for the AI system?* and *Please provide any comments, reflections, or open questions that came up for you during the co-authorship session or when answering this survey.*

Aggregated responses are plotted on Figure 5 (Left), and (Right) factors responses to the question on **collaboration** by industry background and experience (see Figure 7 for more). The results are summarized below.

6.1.1 AI-generated scripts can be surprising, enjoyable, and unique. The statements that received the strongest positive response were *I was **surprised** by the responses from the AI system* (92% participants agreed, 46% strongly agreed) followed by *I **enjoyed** writing with the AI system* (77% agreed, 69% strongly agreed) and *The script(s) written with the AI system feel **unique*** (69% agreed, 46% strongly agreed), referring to lines of dialogue or connections between characters.

6.1.2 Dramatron can be helpful for expressing creative goals for screenwriting. Positive responses to statements *I felt like I was **collaborating** with the AI system*, (77% agreed, 54% strongly agreed),

*I found the AI system **helpful*** (84% agreed, 38% strongly agreed) and *I was able to **express** my creative goals with the AI system* (77% agreed, 23% strongly agreed) suggest a fruitful interaction between a writer and an interactive AI writing tool, in particular for ideation, through multiple choices, twists to the narrative, or specific details.

6.1.3 Dramatron outputs were preferred for screenplays over theatre scripts. We noticed that the subgroup of respondents with a film or TV background gave the highest scores for the surprise, enjoyment, collaboration, helpful, expression, ease and ownership questions. Respondents with theatre background judged enjoyment, collaboration, helpful, creative, expression, ease, proud and ownership the least. This reinforces our observations during interviews (see the criticism of log line-based conditioning for generating theatre scripts in Section 5.6).

6.1.4 Dramatron's hierarchical prompting interface is moderately easy to use. Participants were more split about the ease of use of the tool (61% agreed) and whether they felt proud about the output of the system (46% agreed), with one participant strongly disagreeing. We note that these two answers were highly correlated ($r = 0.9$), and that several questions had a relatively strong correlation ($r \geq 0.66$): helpful, collaboration, ease, enjoyment, expression and ease. As observed during interviews, recurrent criticism was about Dramatron needing detailed log lines to start the generation.

6.1.5 Participants felt a low level of ownership for the co-written scripts. More than half of respondents felt they did not own the result of the AI system. One participant commented that the generated script was only a starting point provided by the AI and that the writer still needed to create their own script. One interpretation is that Dramatron is seen by the industry experts not as much as a full script writing tool but rather as a learning tool to practice writing and a source of inspiration that generates "*not necessarily completed works so much as provocations*" (p7).

6.1.6 Novelty Effects. One potentially confounding factor is that a single engagement of co-writing alongside Dramatron might exhibit some sort of novelty effects. We were curious to understand what it would be like for a single expert to continue to work alongside this sort of system over multiple sessions. Thus, we collected brief feedback from a single industry professional. This expert co-wrote the five performed scripts discussed in Section 5.9 over multiple sessions. They shared that 'if Dramatron gets any better, it might not be as good. This is novel and that provides natural intrigue'. This expert also expressed enthusiasm to write alongside Dramatron again for future productions. While this sentiment partially modulates how generalizable the findings are over longer term engagements, future work could explore how to maintain this novelty through continued iterative incremental tool improvements and co-design.

6.2 Quantitative Observations

We observed that participants would skim the text and remark on specific inspiring lines, but were less focused on the overall *coherence* of the discourse. That said, participants did notice if the dialogue of a scene was *unrelated* to that scene's beat or log line, and most noticed when the generation had excessive *repetition*.

Thus, we pair our qualitative results with a small set of descriptive statistics measuring these features and relevant to our participants' feedback.

Lemma-based Jaccard similarity is a metric measuring the overlap between two sets of lemmatised vocabularies. We found that when participants made multiple generations for a scene description, they generally did not choose outputs with greatest Jaccard similarity. We performed a one-sided Wilcox signed-rank test on Jaccard similarity score differences, testing whether chosen output seeds were more similar to the log line provided by participants. We found that the beats for chosen generations are not more similar to the log line when compared to the beats for generations not chosen ($W = 28$, $p = 0.926$). In fact, there is a strong trend in the opposite direction. This suggests that word overlap is not predictive of which output generation is selected.

Using an open-source repetition scorer [126], we examined the repetition scores for texts where participants generated multiple dialogue outputs but chose only one to incorporate into the script ($n = 3$). After computing a one-sided Wilcox signed-rank test, we found no significant difference between the repetition scores for chosen Dramatron output against alternative seed outputs ($W = 57$, $p = 0.467$). This finding aligns with our observations in the interactive sessions where participants tended to delete degenerate repetitions, or as one participant (p13) put it: "[Dramatron] puts meat on the bones... And then you trim the fat by going back and forth." We interpret our distance measures as indicators of engagement with Dramatron. Figure 6 (right) displays mean relative Levenshtein distances for each participant, along with Jaccard similarity.

We also report document length differences, to show the directionality of editing (i.e. whether participants are adding or deleting material to or from Dramatron output. Figure 6 (left) shows the mean document length difference by participant. Figure 6 (middle) shows these measures normalised and grouped by type of generation. We notice that four participants tended to add to generated outputs (three of them, p6, p13, and p15 expressed interest in staging the resulting scripts) while eight others tended to remove from and curate the outputs (including p1 who staged productions). We also notice that the plot outline was the least edited part of the script. During co-authorship we observed that participants tended to delete longer location descriptions and parts of dialogue (especially loops), and rewrite the character descriptions.

7 DISCUSSION AND FUTURE WORK

7.1 Three Avenues of Future Work towards Generating More Coherent Theatre Scripts and Screenplays

While common sense and logical consistency is an elusive goal for LLMs [7], their utility as writing tools increases as they generate more coherent outputs. Hierarchical generation can be leveraged in various ways to improve the long-range coherence of long texts. First, one enhancement to improve coherence especially suited for screenplays and theatre scripts would be to generate complete character arcs. Second, generating satisfying scene conclusions in the dialogue generation step could be improved by using hierarchical

generation for dialogue: construct each scene's dialogue from a generated beginning, middle, and end dialogue beat. Third, to improve stylistic coherence, future work could explore methods to generate thematic outputs satisfying notions of genre by rejection sampling, as in [68], writing more stylized prompts, or transposing existing ones into a variety of diverse author styles and voices using large language models.

7.2 Script and Screenplay Style Differences and a Critique of Dramatron's Hierarchy

Future work on Dramatron could explore multiple different styles of writing as was suggested through feedback from experts in the study. Dramatron was originally designed for a single style: top-down hierarchical generation where each subsequent step depends on those that came prior. But, during the study, several participants noted that this style of writing is somewhat more aligned with screenwriting than playwriting. This is summarized well in one participant's remarks: "I think the model where writers begin from inputting character and logline favours certain kinds of processes - is there a version of the system that can originate from a different point?" Another participant echoed these sentiments, saying: "Creating theatre is not always a linear / story driven process - a lot of the time it is about depth (theme) as well as forward motion (story) - characters are not always the start of the story, but a means to deliver the deeper meaning of story".

"Playwrights do not use the log line in the same way as in screen writing" (p9), as one participant succinctly stated. Participants 4 and 5, who self-identified as playwrights for the stage, "would never approach a piece of work with a story in [their] head. [They] might come with a more investigative approach with a theme". In fact, they argued against Dramatron's constrained top-down hierarchy: "Why generate characters first? At earlier parts of the creation process you might not know the characters. In the way [we] work, [we] often come with characters at the end, and the idea of characters comes last". The log line itself could be seen as a post-hoc summary, as "often times playwrights find out what the play is about after [they] finish it. I will know the play once it is done" (p9). This said, Dramatron does allow for going back-and-forth between levels in the hierarchy, and through creative prompting, future work could allow for generation steps to happen out of order.

Cultural and economic factors also lead to differences in the writing of screenplays and theatre scripts: "The difference between theatre and screen is that nobody's making theatre on demand, no outside pressure. Or at least not in the same way that there is for television. Theatre feels less like generating content" (p4, p5) whereas "film scripts, in the industry, want the traditional fourth wall" (p9). Since Dramatron is inherently more formulaic by construction, is it suitable to screenplay and film script writing? As one respondent noted, "[Dramatron] will be very useful for an average Hollywood movie and for television. It does not need to have a deep understanding of the human soul, unlike Shakespeare. [...] The thing with action movies is that is that actors are not expected to connect with the writer. A screenwriter on a TV set is just like [Dramatron] [...]. It is a sublime skill to be a Hollywood writer because your creative input is small." (p9).

These reflections invite us to reflect on the application-specific applicability of Dramatron. These comments, and the useful feedback from experts over the course of the study, emphasize the importance of involving the domain experts at every step of the design and development process. In retrospect, earlier engagement might have led to different co-design decisions, and future work might explore such topics.

7.3 Risks and Ethical Questions Related to AI Writing Tools

7.3.1 Biases and stereotypes. As per the feedback gathered in Section 5.3, some biases—in particular around gender—were noticed by the participants in the LLM outputs. While this was not the prime focus of the interview sessions, such 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.

7.3.2 Plagiarism. A quote often attributed to T.S. Eliot, Pablo Picasso, Igor Stravinsky, or William Faulkner, says that "good artists copy, while great artists steal". But, we believe AI writing tools should not encourage such behaviour. Our study participants raised concerns the source of the dataset: "If you are putting existing scripts into the dataset, where are they being pulled from?" (p4, p5). And, other responses spanned an opinion continuum, stretching 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). Currently, the rules are unclear for machine learning-based systems trained on copyright-protected material [9]. Lee et al. [61] distinguish between verbatim, paraphrase, and idea plagiarism and suggest to automatically evaluate the first one using BM25-based retrieval of queries in large set of reference documents, along with a text alignment tool. To mitigate copyright issues, we envision several risk mitigation strategies for writers using co-writing tools. The writer could query short parts of the script using a search engine and/or plagiarism detection tool [61]. In the future, that functionality could be part of the writing tool itself. Further, we believe that such tools require transparency. That is, 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.

7.3.3 Undermining creative economies. The third concern raised by some of the participants was 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). Similar LLM risks have been previously mentioned in [119].

7.3.4 Using Tools or Participating in a Co-Creative System? Previous work has argued for engagement with subject matter experts, literary scholars, and industry professionals in the development of machine learning models [112]. In this work, screenwriters and playwrights co-wrote with Dramatron, and, in the post-interview surveys, most of the participants felt they did not own the final

output (see Sec. 6.1.5) This raises several questions: Should Dramatron be considered merely a writing tool or should it rather be seen as a participant in a co-creative system? Are writers comfortable using and ready to adopt co-creative tools? Participants reflected on the authorship of machine co-created text ("Interesting what this means for the future of writing. Who is the author?", p6).

As a corollary to the issues of authorship and of biases, p3 wondered whether an LLM should generate text from the perspectives of different identities, and if that could amount to cultural appropriation (although they later added: "but I write about many people too, and I am less objective than this AI, because I have seen less data"). Along similar lines, Chung et al. [20] discuss how AI-based creativity support tools augment and automate portions of an artist's support network. These tools may need to conform to the artists' expectations of collaboration, similarly to the types of interactions they have with human collaborators: for example, sub-contracting, co-creation, or inspiration.

8 CONCLUSIONS

We present Dramatron: an interactive co-writing tool which allows writers to generate scripts from a provided log line. Hierarchical story generation with explicit narrative structures and characters helps to generate more coherent text, especially when generating text as long as theatre scripts and screenplays. We conducted a user study with 15 theatre and film industry professionals and distilled their reflections collected through open-ended qualitative interviews and a short survey. We also present feedback from a creative team that produced scripts co-written with Dramatron in public performances at a theatre festival, alongside two reviews from professional reviewers. In summary, Dramatron could be used as a co-creative writing tool, allowing human authors to write screenplays and theatre scripts alongside LLMs ; the development of Dramatron was an iterative process guided by some assumptions, and we hope that our participative and evaluative methods could be adapted and applied to other creative domains leveraging generative models. This work invites further questions on the nature of co-creativity and on the ethics surrounding LLMs.
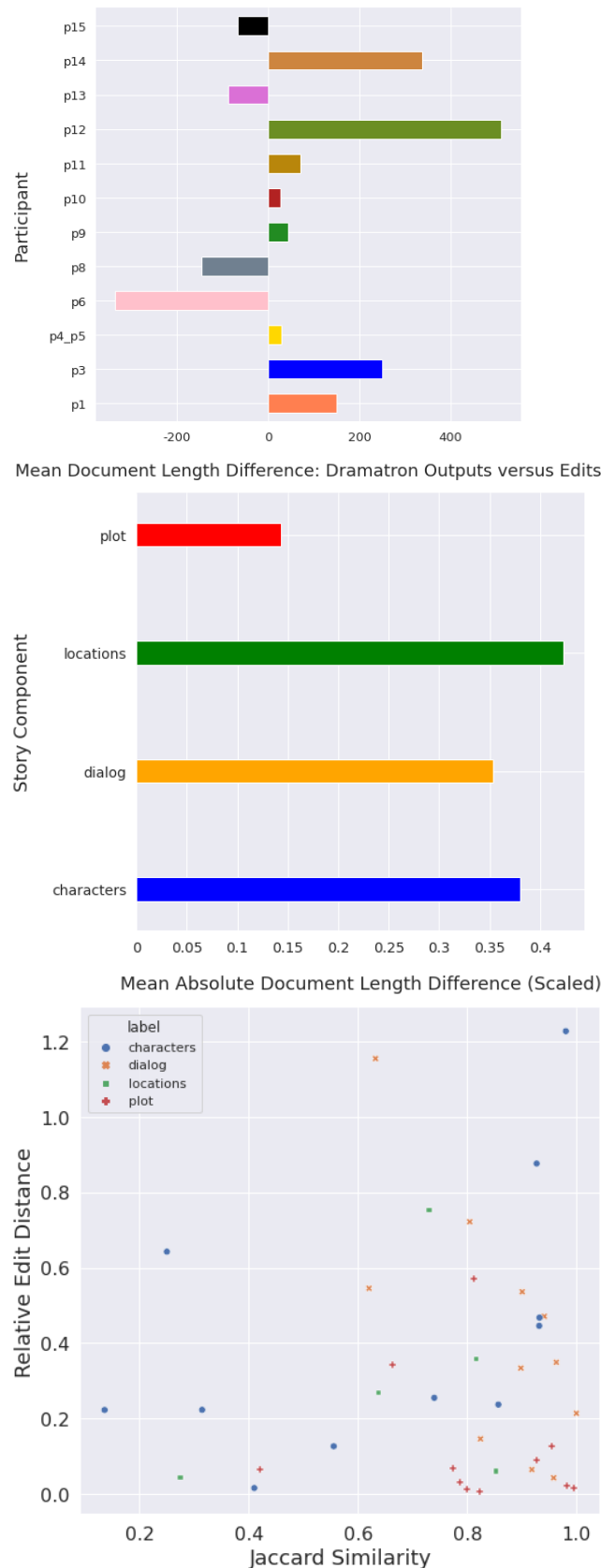
Figure 6: Top: Mean length difference between Dramatron output and final edit, by participant. Middle: Normalized mean absolute length difference by story component. Bottom: Scatter plot of relative edit distance and Jaccard similarity for edited texts versus original Dramatron outputs.

REFERENCES

[1] Calamity AI. 2020. Sollicitors. https://www.youtube.com/watch?v=AmX3GDJ47wo

[2] Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. 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)*. 6470–6484.

[3] Amal Alabdulkarim, Siyan Li, and Xiangyu Peng. 2021. Automatic Story Generation: Challenges and Attempts. *NAACL HLT 2021* (2021), 72.

[4] Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara J Martin, and Mark O Riedl. 2020. Story realization: Expanding plot events into sentences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 7375–7382.

[5] Aristotle. 350 BC. *Poetics*.

[6] AL Becker. 1979. Text Building, Epistemology, and Aesthetic in Javanese Shadow Theatre "dalam The Imagination of Reality. Edited by AL Becker and Aram A. Yengoyan.

[7] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? (2021), 610–623.

[8] Lane Shefter Bishop. 2016. *Sell Your Story in A Single Sentence: Advice from the Front Lines of Hollywood*. The Countryman Press.

[9] 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. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021). https://arXiv.org/abs/2108.07258

[10] Boyd Branch, Piotr Mirowski, and Kory W Mathewson. 2021. Collaborative Storytelling with Human Actors and AI Narrators. *Proceedings of the 12th International Conference on Computational Creativity* (2021). https://arxiv.org/abs/2109.14728

[11] Gwern Branwen. 2020. GPT-3 creative fiction. (2020). https://gwern.net/GPT-3

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, 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. *Advances in neural information processing systems* 33 (2020), 1877–1901. https://arxiv.org/abs/2005.14165

[13] Alex Calderwood, Vivian Qiu, Katy Ilonka Gero, and Lydia B Chilton. 2020. How Novelists Use Generative Language Models: An Exploratory User Study.. In *Proceedings of HAI-GEN+ user2agent@ IUI*.

[14] Alex Calderwood, Noah Wardrip-Fruin, and Michael Mateas. 2022. Spinning Coherent Interactive Fiction through Foundation Model Prompts. (2022).

[15] Joseph Campbell. 1949. *The hero with a thousand faces*. Princeton, NJ: Princeton University Press.

[16] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of Text Generation: A Survey. *CoRR* abs/2006.14799 (2020). arXiv:2006.14799 https://arxiv.org/abs/2006.14799

[17] Wallace L Chafe. 1980. The pear stories: Cognitive, cultural, and linguistic aspects of narrative production. (1980).

[18] Ruijia Cheng, Alison Smith-Renner, Ke Zhang, Joel R Tetreault, and Alejandro Jaimes. 2022. Mapping the Design Space of Human-AI Interaction in Text Summarization. *arXiv preprint arXiv:2206.14863* (2022).

[19] JinUk Cho, MinSu Jeong, JinYeong Bak, and Yun-Gyung Cheong. 2022. Genre-Controllable Story Generation via Supervised Contrastive Learning. In *Proceedings of the ACM Web Conference 2022*. 2839–2849.

[20] John Joon Young Chung, Shiqing He, and Eytan Adar. 2022. Artist Support Networks: Implications for Future Creativity Support Tools. *Proceedings of Designing Interactive Systems: DIS'22* (2022).

[21] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching Stories with Generative Pretrained Language Models. In *CHI Conference on Human Factors in Computing Systems*. 1–19.

[22] 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*. 329–340.

[23] William Wallace Cook. 1928. *Plotto: The Master Book of All Plots*. Ellis, first edition.

[24] Antonia Creswell and Murray Shanahan. 2022. Faithful Reasoning Using Large Language Models. *arXiv preprint arXiv:2208.14271* (2022).

[25] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and Play Language Models:

A Simple Approach to Controlled Text Generation. *CoRR* abs/1912.02164 (2019). arXiv:1912.02164 http://arxiv.org/abs/1912.02164

[26] Anna De Fina and Barbara Johnstone. 2015. Discourse analysis and narrative. *The handbook of discourse analysis* 1 (2015), 152–167.

[27] Paul Debreczeny. 1984. Chekhov's Art: A Stylistic Analysis. *Slavic Review* 43, 2 (1984), 347–348.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[29] Alara Dirik, Hilal Donmez, and Pinar Yanardag. 2021. Controlled Cue Generation for Play Scripts. *arXiv preprint arXiv:2112.06953* (2021).

[30] Tim Draws, David La Barbera, Michael Soprano, Kevin Roitero, Davide Ceolin, Alessandro Checco, and Stefano Mizzaro. 2022. The Effects of Crowd Worker Biases in Fact-Checking Tasks. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2114–2124.

[31] Wanyu Du, Zae Myung Kim, Vipul Raheja, Dhruv Kumar, and Dongyeop Kang. 2022. 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)*. 96–108.

[32] Nouha Dziri, Ehsan Kamalloo, Kory W Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. *arXiv preprint arXiv:1904.03371* (2019).

[33] David Edgar. 2009. *How plays work*. Nick Hern Books.

[34] Markus Eger and Kory W Mathewson. 2018. dairector: Automatic story beat generation through knowledge synthesis. *arXiv preprint arXiv:1811.03423* (2018).

[35] Markus Eger, Colin M Potts, Camille Barot, and R Michael Young. 2015. Plotter: operationalizing the master book of all plots. *Proceedings of the Intelligent Narrative Technologies and Social Believability in Games* (2015), 30–33.

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

[37] Richard Evans and Emily Short. 2013. Versu—a simulationist storytelling system. *IEEE Transactions on Computational Intelligence and AI in Games* 6, 2 (2013), 113–130.

[38] Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical Neural Story Generation. *CoRR* abs/1805.04833 (2018). arXiv:1805.04833 http://arxiv.org/abs/1805.04833

[39] Angela Fan, Mike Lewis, and Yann N. Dauphin. 2019. Strategies for Structuring Story Generation. *CoRR* abs/1902.01109 (2019). arXiv:1902.01109 http://arxiv.org/abs/1902.01109

[40] Christiane Fellbaum. 2010. WordNet. In *Theory and applications of ontology: computer applications*. Springer, 231–243.

[41] Gustav Freytag. 1894. *Die technik des dramas*. S. Hirzel.

[42] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for science writing using language models. In *Designing Interactive Systems Conference*. 1002–1019.

[43] Sarik Ghazarian, Zixi Liu, SM Akash, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2021. Plot-guided Adversarial Example Construction for Evaluating Open-domain Story Generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4334–4344.

[44] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375* (2022).

[45] Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content Planning for Neural Story Generation with Aristotelian Rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 4319–4338.

[46] Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics* 8 (2020), 93–108.

[47] Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6473–6480.

[48] Barbara Hayes-Roth and Robert Van Gent. 1996. Improvisational puppets, actors, and avatars. In *Proc Computer Game Developers' Conf*.

[49] Ernest Hemingway. 1964. *A Moveable Feast*. Scribner's.

[50] 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. 2022. Training Compute-Optimal Large Language Models. *arXiv preprint arXiv:2203.15556* (2022).

[51] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*.

[52] Zhe Hu, Hou Pong Chan, Jiachen Liu, Xinyan Xiao, Hua Wu, and Lifu Huang. 2022. PLANET: Dynamic Content Planning in Autoregressive Transformers for Long-form Text Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2288–2305.

[53] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. 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*. 1–12.

[54] Yiping Jin, Vishakha Kadam, and Dittaya Wanvarie. 2022. Plot Writing From Pre-Trained Language Models. *arXiv preprint arXiv:2206.03021* (2022). https://arXiv.org/abs/2206.03021

[55] Barbara Johnstone. 2005. Discourse analysis and narrative. *The handbook of discourse analysis* (2005), 635–649.

[56] Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. 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*. 1265–1285.

[57] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858* (2019). http://arxiv.org/abs/1909.05858

[58] Robyn Kozierok, John Aberdeen, Cheryl Clark, Christopher Garay, Bradley Goodman, Tonia Korves, Lynette Hirschman, Patricia L McDermott, and Matthew W Peterson. 2021. Assessing Open-Ended Human-Computer Collaboration Systems: Applying a Hallmarks Approach. *Frontiers in artificial intelligence* 4 (2021).

[59] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*. PMLR, 957–966.

[60] William Labov and Joshua Waletzky. 1967. Narrative analysis. In *Essays on the Verbal and Visual Arts, ed. J. Helm*. Seattle: U. of Washington Press, 12–44.

[61] Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. 2022. Do Language Models Plagiarize? *arXiv e-prints* (2022), arXiv–2203.

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

[63] Boyang Li, Stephen Lee-Urban, George Johnston, and Mark O. Riedl. 2013. Story Generation with Crowdsourced Plot Graphs. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence* (Bellevue, Washington) *(AAAI'13)*. AAAI Press, 598–604. http://dl.acm.org/citation.cfm?id=2891460.2891543

[64] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proc ACL Wkshp. Vol 8*.

[65] Inderjeet Mani. 2012. Computational modeling of narrative. *Synthesis Lectures on Human Language Technologies* 5, 3 (2012), 1–142.

[66] Lara J Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark O Riedl. 2017. Event representations for automated story generation with deep neural nets. *arXiv preprint arXiv:1706.01331* (2017).

[67] Kory Mathewson and Piotr Mirowski. 2018. Improbotics: Exploring the imitation game using machine intelligence in improvised theatre. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 14.

[68] Kory W. Mathewson, Pablo Samuel Castro, Colin Cherry, George F. Foster, and Marc G. Bellemare. 2019. Shaping the Narrative Arc: An Information-Theoretic Approach to Collaborative Dialogue. *CoRR* abs/1901.11528 (2019). arXiv:1901.11528 http://arxiv.org/abs/1901.11528

[69] Kory Wallace Mathewson and Piotr Mirowski. 2017. Improvised Comedy as a Turing Test. *arXiv preprint arXiv:1711.08819* (2017).

[70] Kory Wallace Mathewson and Piotr Mirowski. 2017. Improvised theatre alongside artificial intelligences. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*.

[71] Robert McKee. 1997. Story: Substance, Structure, Style and the Principles of Screenwriting. 1997. *Kent, Great Britain: Methuen* (1997).

[72] James Richard Meehan. 1976. *The metanovel: writing stories by computer*. Technical Report. Yale Univ, New Haven Conn, Dept of Comp Sci.

[73] James R Meehan. 1977. TALE-SPIN, An Interactive Program that Writes Stories.. In *IJCAI*, Vol. 77. 91–98.

[74] Piotr Mirowski, Sumit Chopra, Suhrid Balakrishnan, and Srinivas Bangalore. 2010. Feature-rich continuous language models for speech recognition. In *Spoken Language Technology Wkshp, 2010 IEEE*. IEEE, 241–246.

[75] Piotr Mirowski and Kory Wallace Mathewson. 2019. Human improvised theatre augmented with artificial intelligence. In *Proceedings of the 2019 on Creativity and Cognition*. 527–530.

[76] Annalee Newitz. 2016. Movie written by algorithm turns out to be hilarious and intense. https://arstechnica.com/gaming/2021/05/an-ai-wrote-this-movie-and-its-strangely-moving/

[77] Eric Nichols, Leo Gao, and Randy Gomez. 2020. Collaborative storytelling with large-scale neural language models. In *Motion, Interaction and Games*. 1–10.

[78] OpenAI. 2021. Pricing. https://openai.com/api/pricing/

[79] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022.

Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155* (2022).

[80] Vishakh Padmakumar and He He. 2022. Machine-in-the-Loop Rewriting for Creative Image Captioning. In *Proceedings of the 20th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

[81] Pinelopi Papalampidi, Kris Cao, and Tomas Kocisky. 2022. Towards Coherent and Consistent Use of Entities in Narrative Generation. *International Conference on Macine Learning* (2022).

[82] Ken Perlin and Athomas Goldberg. 1996. Improv: A system for scripting interactive actors in virtual worlds. In *Proc. Conf. on Computer Graphics and Interactive Techniques*. ACM, 205–216.

[83] Mihai Polceanu, J. Porteous, A. Lindsay, and M. Cavazza. 2021. Narrative Plan Generation with Self-Supervised Learning. In *AAAI*.

[84] Georges Polti. 1917. *The thirty-six dramatic situations*. Editor Company.

[85] Vladimir Iakovlevich Propp. 1968. *Morphology of the Folktale*. Vol. 9. University of Texas Press.

[86] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *CoRR* abs/2112.11446 (2021). arXiv:2112.11446 https://arxiv.org/abs/2112.11446

[87] Samson Raphaelson. 1949. *The Human Nature of Playwriting*. New York: Macmillan.

[88] Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. PlotMachines: Outline-Conditioned Generation with Dynamic Plot State Tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 4274–4295.

[89] Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2021. A Recipe For Arbitrary Text Style Transfer with Large Language Models. *arXiv preprint arXiv:2109.03910* (2021).

[90] Mark O Riedl and Robert Michael Young. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research* 39 (2010), 217–268.

[91] Melissa Roemmele and Andrew Gordon. 2018. Linguistic features of helpfulness in automated support for creative writing. In *Proceedings of the First Workshop on Storytelling*. 14–19.

[92] Melissa Roemmele, Andrew S Gordon, and Reid Swanson. 2017. Evaluating story generation systems using automated linguistic analyses. In *SIGKDD 2017 Workshop on Machine Learning for Creativity*. 13–17.

[93] Rudolf Rosa, Ondřej Dušek, Tom Kocmi, David Mareček, Tomáš Musil, Patrícia Schmidtová, Dominik Jurko, Ondřej Bojar, Daniel Hrbek, David Košťák, et al. 2020. THEaiTRE: Artificial intelligence to write a theatre play. *arXiv preprint arXiv:2006.14668* (2020).

[94] Rudolf Rosa, Tomáš Musil, Ondřej Dušek, Dominik Jurko, Patrícia Schmidtová, David Mareček, Ondřej Bojar, Tom Kocmi, Daniel Hrbek, David Košťák, et al. 2021. THEaiTRE 1.0: Interactive generation of theatre play scripts. *arXiv preprint arXiv:2102.08892* (2021).

[95] Rudolf Rosa, Patrícia Schmidtová, Ondřej Dušek, Tomáš Musil, David Mareček, Saad Obaid, Marie Nováková, Klára Vosecká, and Josef Doležal. 2022. GPT-2-based Human-in-the-loop Theatre Play Script Generation. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*. 29–37.

[96] David E Rumelhart. 1975. Notes on a schema for stories. In *Representation and understanding*. Elsevier, 211–236.

[97] David E Rumelhart. 1980. On evaluating story grammars. (1980).

[98] Keisuke Sakaguchi, Chandra Bhagavatula, Ronan Le Bras, Niket Tandon, Peter Clark, and Yejin Choi. 2021. proScript: Partially Ordered Scripts Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 2138–2149.

[99] Patrícia Schmidtová, Dávid Javorskỳ, Christián Mikláš, Tomáš Musil, Rudolf Rosa, and Ondřej Dušek. 2022. DialogueScript: Using Dialogue Agents to

[100] Produce a Script. *arXiv preprint arXiv:2206.08425* (2022).

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

[102] Industrial Scripts. 2019. How to write outstanding TV & movie loglines: The ultimate guide. https://industrialscripts.com/loglines-guide/

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

[104] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696* (2020).

[105] Graeme Shimmin. 2021. Logline formula: How to use the Killogator formula to write a killer logline. https://graemeshimmin.com/writing-a-logline-for-a-novel/

[106] Wai Man Si, Prithviraj Ammanabrolu, and Mark Riedl. 2021. Telling Stories through Multi-User Dialogue by Modeling Character Relations. In *SIGDIAL*.

[107] Blake Snyder. 2005. *Save the cat*. Michael Wiese Productions Chelsea, Michigan.

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

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

[110] Jennifer Tang, Nina Segal, and Chinonyerem Odimba. 2021. AI at the Young Vic. https://www.youngvic.org/whats-on/ai

[111] Mariët Theune, Sander Faas, Anton Nijholt, and Dirk Heylen. 2003. The virtual storyteller: Story creation by intelligent agents. In *Proceedings of the Technologies for Interactive Digital Storytelling and Entertainment (TIDSE) Conference*, Vol. 204215.

[112] Perry W Thorndyke. 1977. Cognitive structures in comprehension and memory of narrative discourse. *Cognitive psychology* 9, 1 (1977), 77–110.

[113] Imke Van Heerden and Anil Bas. 2021. AI as Author–Bridging the Gap Between Machine Learning and Literary Theory. *Journal of Artificial Intelligence Research* 71 (2021), 175–189.

[114] Christopher Vogler. 2007. *The writer's journey*. Michael Wiese Productions Studio City, CA.

[115] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *ArXiv* abs/1804.07461 (2018).

[116] Tianming Wang and Xiaojun Wan. 2019. T-CVAE: Transformer-Based Conditioned Variational Autoencoder for Story Completion.. In *IJCAI*. 5233–5239.

[117] Noah Wardrip-Fruin. 2009. Expressive processing. *Cambridge: MIT Press. Weiberg, B.(2002). Beyond Interactive Cinema. Retrieved April* 9 (2009), 2009.

[118] Stephen G Ware and Robert Michael Young. 2011. CPOCL: A Narrative Planner Supporting Conflict.. In *AIIDE*.

[119] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903* (2022).

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

[121] Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nissan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively Summarizing Books with Human Feedback. arXiv:2109.10862 [cs.CL]

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

[123] Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models. *arXiv preprint arXiv:2010.00840* (2020). https://arxiv.org/abs/2010.00840

[124] Daijin Yang, Yanpeng Zhou, Zhiyuan Zhang, Toby Jia-Jun Li, and Ray LC. 2022. AI as an Active Writer: Interaction strategies with generated text in human-AI collaborative fiction writing. In *Joint Proceedings of the ACM IUI Workshops 2022*, Vol. 10.

[125] Lili Yao, Nanyun Peng, Ralph M. Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2018. Plan-And-Write: Towards Better Automatic Storytelling. *CoRR* abs/1811.05701 (2018). arXiv:1811.05701 http://arxiv.org/abs/1811.05701

[126] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *27th International Conference on Intelligent User Interfaces*. 841–852.

[127] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. *CoRR* abs/1912.08777 (2019). arXiv:1912.08777 http://arxiv.org/abs/1912.08777

[128] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1097–1100.

A   RELATED WORK ON AUTOMATED STORY GENERATION AND CONTROLLABLE STORY GENERATION

In this section we provide background and related work on the intersecting fields of automatic plot and story generation as well as controllable language generation.

A.1   Automatic Story Generation

Automatic story generation is the research problem concerned with generating sequences of story elements that collectively tell a coherent narrative. A narrative plot is a sequence of events where each affects the next. The plot is composed of narrative elements sometimes referred to as actions, beats, scenes, or events [72]. Generative plot systems have been developed for nearly a century by Cook [23], and computerized versions have existed for decades [73]. These systems support human authors with creative output material and as a source of randomness. Recent work has adapted these systems for computational interaction for use in web-based and theatrical settings [34, 35]. In generating narratives, the combinations of these component elements form subplots. Multiple subplots can be combined into a single plot, and multiple plots can intertwine to create complex narratives. Many contemporary stories are written to have multiple plot lines which intertwine. But, there is little work on how to computationally model and generate multiple intersecting plot lines. Complex plot line interaction is a promising avenue of future work for human-machine co-creativity research in story generation.

Early approaches to automatic story generation used symbolic planning and hand-engineered heuristics [63, 73, 90, 110, 117]. Recently, research has explored open-story generation using machine learning techniques which leverage large datasets, massive deep learning models, increased compute capacity, and large language model prompt engineering [11, 14, 38, 83, 89, 102]. These methods show how models can succeed, and fail, in the generation of unique and coherent stories. Additionally, while coherence has been studied in dialogue generation methods [32], it remains challenging to measure coherence in story, specifically as it relates to causal narrative events, or common sense knowledge [3], or consistency in characters [81].

A.2   Symbolic and Hierarchical Story Generation

Some work has tried to bridge the gap between symbolic event representations and textual representations. Several of these methods process and predict events from text [47, 66, 98, 115] by generating sequences of plot events and then expanding such plot events into sentences [4, 88]. Others model each story as a series of character and story challenge cards [2] (first simulating sequences of causal events, then transforming them into sentences) or by simulating social practices between autonomous agents [37].

Other recent work separates storytelling into two phases: storyline (i.e. plot) planning and story writing based on that storyline [124]. Similarly, methods have been introduced which decompose story generation into processes of coarse-to-fine generation [16, 39]. Goldfarb-Tarrant et al. [45] further introduced rescoring methods for character and plot events. These methods have not

focused on synthesising coherent stories by generating scenes and dialogue, as we do in our work. Many of these methods lean on human reading comprehension and preference-based evaluation as opposed to production of the final script.

Hierarchical generation of a theatre play was first mentioned and used in [93, 95] for the production of *AI: Can a Robot Write a Play?* by company THEaiTRE in 2021 in Prague[9]. In this work, the authors start with a title (or a prompt for the story) and then generate a textual synopsis, which is then used to generate the dialogue. In contrast to our approach, they did not generate characters alongside synopsis and would start the "flat" dialogue generation from manually input two-line exchanges between characters. Their work also only served the production of a specific theatrical play rather than being evaluated within a diverse community of writers.

A.3   Controllable Story Generation

Previous work used a trained autoregressive transformer to plan sentences of an opinion piece from a set of user-supplied keywords [52]. This built upon [122] which incorporated commonsense knowledge into keyword-based story generation. Using conditional language models controlled by topics or keywords [74], Cho et al. [19] trained genre-controlled short story generators.

Story arc generation was introduced in the Tale Brush graphical tool [21], using Kurt Vonnegut's theory about the fortune of the protagonist as the story progresses. This theory has also been used by Mathewson et al. [68] as an approach to produce creative, engaging dialogue. We similarly use the concept of the narrative arc, though we use it textually in the prompts to Dramatron.

In "Controlled Cue Generation for Play Scripts", Dirik et al. [29] use LLMs to generate both the next line and a stage cue. In "DialogueScript: Using Dialogue Agents to Produce a Script", Schmidtová et al. [99] use different LLMs for each character. Si et al. [105] model multi-user dialogue and character relationships for story continuation. Schmitt and Buschek [100] use question-based chatbot interactions to assist with character creation.

Prompt engineering has been used to write a short plot from two characters, a genre and a theme in [54]. Our work of decomposing a log line into a synopsis can be seen as a narrative equivalent of Chain of Thought prompting for reasoning tasks [118], and uses the idea of using LLM output as prompt for the next stage of the generation—also called prompt chaining [121] or language engineering [24].

A.4   Review of Automated and Machine-Learned Metrics for the Evaluation of Story Generation

A.4.1   Similarity Between Generated and "Ground-Truth" Stories. In a typical machine learning mindset, story generation can be envisioned as merely a prediction task, allowing for evaluation against "ground truth". An example of such datasets includes the Writing Prompts[10], a set of 300k human-written stories, at average of 734 words, paired with writing prompts [38]. Fan et al. [38] propose metrics such as test set perplexity, prompt ranking accuracy (a measure of likelihood of a story generated using a true prompt vs.

---

[9]Performance documentation available at https://theaitre.com/.
[10]https://www.kaggle.com/datasets/ratthachat/writing-prompts

decoys), average longest common subsequence, and a triple-pairing task for human annotator evaluation of prompt-story coherence. They later measure sentence completion top N-in-M accuracy [39]. Si et al. [105] measure top-N hits of character or story continuation. Rashkin et al. [88] compare generated stories to ground truth using the ROUGE score [64].

A.4.2 Consistency between Generated Stories and Their Writing Prompts. In the context of prompt-based story generation or continuation, Roemmele et al. [92] measure the quality of generated text based on whether it presents a consistent writing style and maintains the category (part-of-speech tags) distribution of individual words between the prompt and the generated story. They also record story-dependent metrics like lexical cohesion, style matching and entity coreference, and story-independent metrics such as sentence length, grammaticity, lexical diversity, lexical frequency and syntactic complexity. See et al. [102] measure N-gram similarity and sentence embedding similarity between the generated story and the prompt. Further metrics include counting the number of unique words, percentage of verbs and diversity in entity names [39], rare word usage and sentence length [102].

A.4.3 Statistical Measures of Corpora of Generated Stories. Without comparing individual generated stories to a ground truth or to a writing prompt, one can measure the Vocab:token ratio (originality and diversity of content), number of entities per plot, of unique verbs, verb diversity as well as inter- and intra-story trigram or 4-gram repetition [45, 46]. Rashkin et al. [88] measure the diversity of generated sentences using self-BLEU scores [127], or even adversarially train a classifier for the plausibility of a short story [43].

B ADDITIONAL DISCUSSION FROM PLAYS BY BOTS CREATIVE TEAM

Discussions amongst the creative team were summarized in the body of the text (see Section 5.9).

To reiterate, four key themes emerged through these discussions which echo the themes presented in Section 5. These themes are discussed in detail in this section, alongside supporting quotes.

First, the system has a distinct glitch style that can sometimes be nonsensical, vague, or passive. As one performer recounted, "sometimes there is internal conflict within the text, and in those moments it is as if [Dramatron] is working against itself". This pushes performers to commit to choices, and to be concise and specific and complimentary in their interpretation. One performer noted that if the system generated flawless text, "it might not work as well", because "the mistakes in the script are the joy". Another went a step further, saying "I almost want more curveballs and non-sequiturs... more crunchy bits". 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".

Secondly, the generated text can sometimes be repetitive. This can be interpreted as a mistake. Or, this repetition can be fun and playful if it is interpreted as an important and deliberate choice from the playwright: "a choice to be honored," as one performer said. When the system did repeat itself, the cast was able to make

the lines more meaningful with their own unique human talents. As one said, "you can do so much with your line reading, with your physicality, delivery, proximity, and acting choices."

Third, the team discussed agency and expectations of the systems capabilities. For instance, in the way that the performers would refer to the system's choices. One said "Dramatron was trying to show me the angle," or "I was trying to understand what Dramatron meant". Discussions amongst the creative team explored their expectations of what the system could and could not do. For example, "it is just a bot, not really fully understanding how [the world] works", and another responded "Dramatron is trying."

Finally, the prevailing feedback from the majority of performers was that participating in the production was fun. As one actor said, "it was liberating and easy because the world creating was done for you, the platform was done, and that can be mentally exhausting as an improviser." These reflections discussed by the creative team reflect the usefulness of co-written scripts. This is particularly true when used for a production such as *Plays By Bots*, which leverages professional improvisers to interpret the scripts. The co-creativity of a system such as Dramatron extends beyond the playwright and Dramatron, and to the performer on the stage working with the generated and edited text.

These reflections discussed by the creative team reflect the usefulness of co-written scripts.

C DETAILS OF QUANTITATIVE OBSERVATIONS

C.1 Levenshtein Distance

Levenshtein distance was calculated at the character level for edited versus non-edited texts using the default edit distance function from the Natural Language Tool Kit[11] package's distance module (no transposition operations and substitution cost equal to 1). As an absolute measure, the distance metric indicates how many operations (insertion, deletion, substitution) are required to convert one string into another and is typically calculated with a dynamic programming approach to minimizing the cost of character-level edit operations for the conversion. However, as a measure dependent on the lengths of both input strings, comparing across edit distances becomes rather uninterpretable when the string sets differ in length, for example across sets of edited and unedited texts for characters, locations, plots, and dialogues. As we are interested in understanding the extent to which participants edited the output of Dramatron as a cohort, it is reasonable to normalise the distances with respect to their length, which we report in 6 (Right).To do this, we calculate a relative Levenshtein distance as the ratio of the raw Levenshtein distance between edited and non-edited (Dramatron) output text to the length of original Dramatron output. Conceptually, we can view the resulting measure as a proportional measure of interaction with our tool. Given that Levenshtein distance operations are character-level, and length is measured in characters, the proportion represents the a weighting of active versus passive interaction with Dramatron for the different levels of structure in the hierarchical story generation process. Positive scores for relative Levenshtein distance indicate one aspect of active writing with

---

[11]https://www.nltk.org

Dramatron, while negative scores for relative Levenshtein distance indicate one aspect of passive acceptance of Dramatron (choosing among generated text seeds is another aspect of interaction not accounted for with this metric).[12]

C.2 Length Difference

We note that for Mean Document Length in 6 (left), the means are both negative and positive, and are not scaled. This is to observed workflow differences per participant as well as to capture the directionality of editing Dramatron's output. For the Mean Absolute Differences in 6 (center), we take the absolute difference between character lengths between original Dramatron output and edited output and normalize it with min-max normalization.

C.3 Jaccard Similarity (Relatedness)

To calculate Jaccard similarity, we first calculate the Jaccard distance, which is calculated by dividing the cardinality of the intersection of two sets and by the cardinality of their union. By subtracting Jaccard distance by 1, we obtain the Jaccard similarity. Jaccard metrics are scaled between 0 and 1, with 0 representing zero set similarity and 1 representing total set similarity. As Jaccard similarity simply compares set entities, one must specify the choice of entities to compare. In order to calculate the Jaccard similarity on Dramatron output and its edited counterparts, we apply a pre-processing pipeline to our texts, first tokenising sentences and words, and then generating

a set of lemmas from word tokens based on their most probable part of speech according to WordNet [40]. The resulting scores are then lemma-based similarity scores between Dramatron output and participant edits, which we use as a descriptive measure of word choice similarity. Note that we do not calculate semantic similarity with embeddings, but simply compute set overlap at the lemma level. Lemmas are chosen to abstract over inflectional differences for words for a slightly more precise look at word choice. Future work should investigate alternative methods of assessing similarity, such as word mover distance [59] or BLEURT scores [103].

C.4 Repetition

We calculate repetition-based scores with open-source tools from [126], which calculate scores for various n-gram overlaps[13]. N-gram overlaps are calculated for unigram through 10-gram sequences, as well as for Total Consecutive Reptition (TCR) and Longeset Consecutive Repetition (LCR). To compute the Wilcox test, we use pairwise differences across corresponding features for chosen versus alternative seed generations (e.g. unigram to unigram differences, bigram to bigram differences, etc.). We do not weight differences by type of repetition feature.

---

[12]From a linguistic point of view, insertions and deletions can range from deletion of entire words (e.g. this is really great → this is great) to the insertion or deletion morphemes to achieve a sociolinguistically marked effect (e.g. *going* → *goin'*, *like to* → *liketa*, etc).

[13]Implementation details at https://github.com/google-research/pegasus/blob/main/pegasus/eval/repetition/repetition_scorer.py.

D SUPPLEMENTARY FIGURES

Figure 7 shows the participants' responses to the quantitative evaluation, on a Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree), and broken down by groups of participants. For the first group, we defined a binary indicator variable (*Has experience of AI writing tools*). For the second group, we defined a three-class category for their primary domain of expertise (*Improvisation*, *Scripted Theatre* and *Film or TV*).

E FULL PROMPT PREFIXES FOR DRAMATRON

E.1 About the Prompt Sets

In our study we relied on two sets of prompts: Medea and Sci-Fi.

Medea prompts are based on Ancient Greek tragedy *Medea*, by Euripides (431 BC). The dialogue prompts were taken verbatim from the translation of the play by E. P. Coleridge (1863 - 1936), available in the public domain[14]. We replaced CHORUS by WOMEN OF CORINTH. The plot and character prompts were written from a summary taken from Spark Notes[15] and Wikipedia[16]. To encourage the generation of different locations, Aristotle's Unity of Place is not respected, and location "Outside the Royal Palace" is renamed as "Medea's modest home" as well as "On a winged chariot" (even though these are the same locations in the original tragedy). Prompts for *Antigone*[17] (Sophocles), *The Bacchae*[18] (Euripides), and *The Frogs*[19] (Aristophanes) were adapted from Wikipedia and ancient-literature.com.

The Star Wars log line was adapted from chapter "Creating the killer log line" by Bill Lundy[20]. Sci-Fi character prompts were adapted from *Star Wars: Episode IV - A New Hope* (1977), written and directed by George Lucas, produced by Lucasfilm and distributed by 20th Century Fox. We also adapted the breakdown of Star Wars into a Hero Journey[21]. The script[22], plot[23] and log line[24] of *Plan 9 from Outer Space*, which is in public domain.

E.2 Title Prompt Prefixes

In the following sets of prompt prefixes, the **<LOG_LINE>** is provided by the writer.

```
Examples of alternative, original and descriptive titles for known play and film scripts.

Example 1. Ancient Greek tragedy based upon the myth of Jason and Medea. Medea, a former princess of the kingdom of
Colchis, and the wife of Jason, finds her position in the Greek world threatened as Jason leaves her for a Greek
princess of Corinth. Medea takes vengeance on Jason by murdering his new wife as well as her own two sons, after which
she escapes to Athens. Title: A Feminist Tale<end>

Example 2. Ancient Greek tragedy that deals with Antigone's burial of her brother Polynices, in defiance of the laws of
Creon and the state, and the tragic repercussions of her act of civil disobedience. Title: In My Brother's Name<end>

Example 3. Greek comedy that tells the story of the god Dionysus (also known to the Greeks as Bacchus) who, despairing
of the current state of Athens' tragedians, travels to Hades with his slave Xanthias to bring Euripides back from the
dead. Title: Dionysus in Hades<end>

Example 4. <LOG_LINE> Title:
```

Listing 1: Title prompt prefixes to generate a title from a log line (Medea).

```
Examples of alternative, original and descriptive titles for known play and film scripts.

Example 1. A science-fiction fantasy about a naive but ambitious farm boy from a backwater desert who discovers powers
he never knew he had when he teams up with a feisty princess, a mercenary space pilot and an old wizard warrior to lead
a ragtag rebellion against the sinister forces of the evil Galactic Empire. Title: The Death Star's Menace<end>

Example 2. Residents of San Fernando Valley are under attack by flying saucers from outer space. The aliens are
extraterrestrials who seek to stop humanity from creating a doomsday weapon that could destroy the universe and unleash
the living dead to stalk humans who wander into the cemetery looking for evidence of the UFOs. The hero Jeff, an airline
 pilot, will face the aliens. Title: The Day The Earth Was Saved By Outer Space.<end>
```

---

[14]http://classics.mit.edu/Euripides/medea.pl.txt
[15]https://www.sparknotes.com/lit/medea/
[16]https://en.wikipedia.org/wiki/Medea_(play)
[17]https://www.ancient-literature.com/greece_sophocles_antigone.html
[18]https://en.wikipedia.org/wiki/The_Bacchae
[19]https://www.ancient-literature.com/greece_aristophanes_frogs.html
[20]In Ellis, Sherry, and Laurie Lamson. "Now Write! Mysteries: Suspense, Crime, Thriller, and Other Mystery Fiction Exercises from Today's Best Writers and Teachers.", Penguin, 2011
[21]https://thescriptlab.com/features/screenwriting-101/12309-the-heros-journey-breakdown-star-wars/
[22]http://www.horrorlair.com/scripts/criswell.txt
[23]https://en.wikipedia.org/wiki/Plan_9_from_Outer_Space
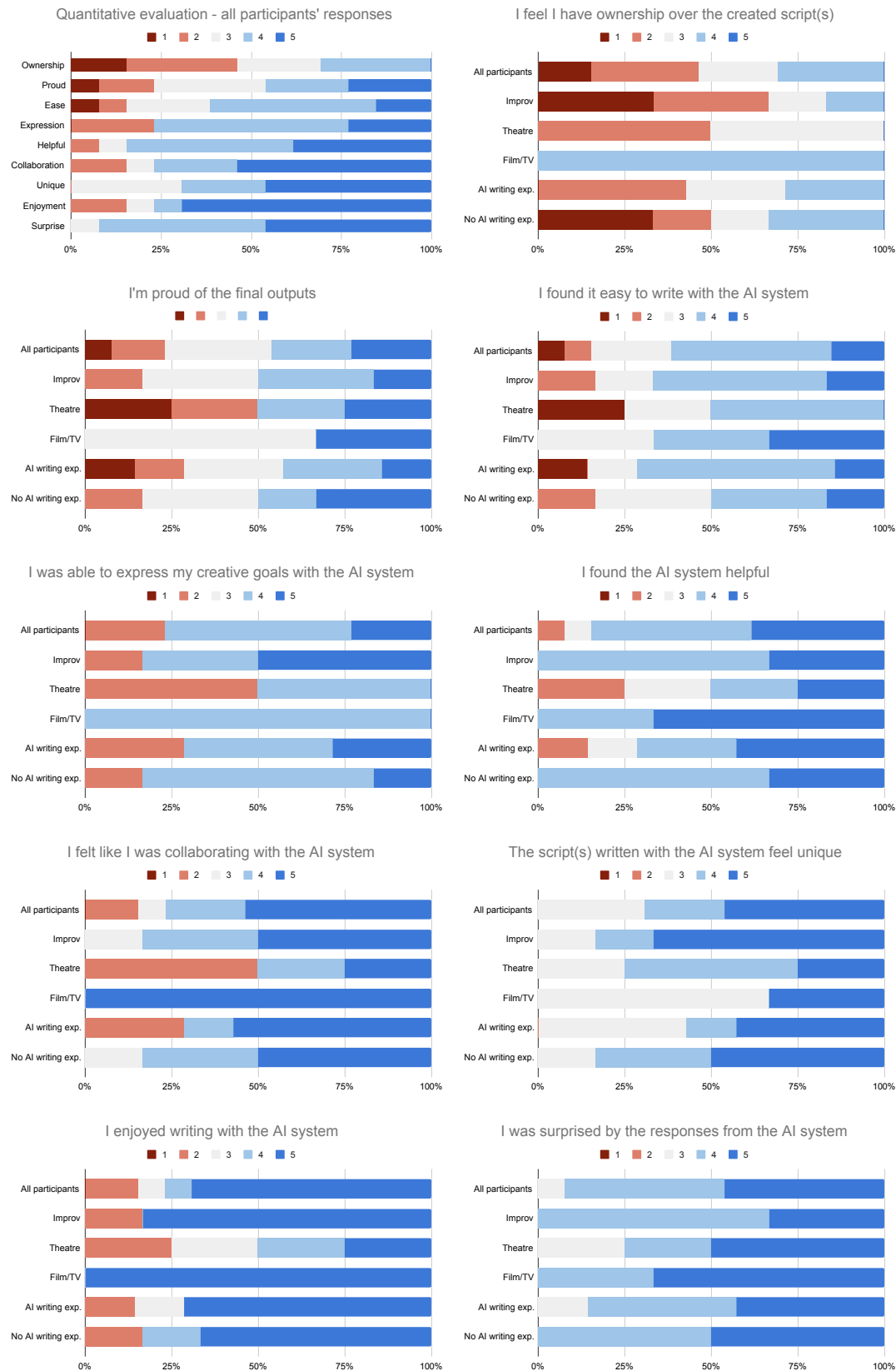[24]https://www.rottentomatoes.com/m/plan-9-from-outer-space

Figure 7: Participants responses to the quantitative evaluation, on a Likert scale from 1 (strongly disagree) to 5 (strongly agree)

```
Example 3. <LOG_LINE> Title:
```

Listing 2: Title prompt prefixes to generate a title from a log line (Sci-Fi).

E.3   Character Description Prompt Prefixes

In the following sets of prompt prefixes, the **<LOG_LINE>** is provided by the writer.

```
Example 1. Ancient Greek tragedy based upon the myth of Jason and Medea. Medea, a former princess and the wife of Jason,
 finds her position in the Greek world threatened as Jason leaves Medea for a Greek princess of Corinth. Medea takes
vengeance on Jason by murdering his new wife as well as Medea's own two sons, after which she escapes to Athens.
<character>Medea <description> Medea is the protagonist of the play. A sorceress and a princess, she fled her country
and family to live with Jason in Corinth, where they established a family of two children and gained a favorable
reputation. Jason has divorced Medea and taken up with a new family.<stop>
<character>Jason <description> Jason is considered the play's villain, though his evil stems more from weakness than
strength. A former adventurer, Jason abandons his wife, Medea, in order to marry the beautiful young daughter of Creon,
King of Corinth, and fuels Medea to a revenge.<stop>
<character>Women of Corinth <description> The Women of Corinth are a commentator to the action. They fully sympathizes
with Medea's plight, excepting her decision to murder her own children.<stop>
<character>Creon <description> Creon is the King of Corinth, banishes Medea from the city.<stop>
<character>The Nurse <description> The Nurse is the caretaker of the house and of the children and serves as Medea's
confidant.<stop>
<end>

Example 2. <LOG_LINE>
```

Listing 3: Character prompt prefixes to generate list of character descriptions from a log line (Medea).

```
Example 1. A science-fiction fantasy about a naive but ambitious farm boy from a backwater desert who discovers powers
he never knew he had when he teams up with a feisty princess, a mercenary space pilot and an old wizard warrior to lead
a ragtag rebellion against the sinister forces of the evil Galactic Empire.
<character>Luke Skywalker <description>Luke Skywalker is the hero. A naive farm boy, he will discover special powers
under the guidance of mentor Ben Kenobi.<stop>
<character>Ben Kenobi <description>Ben Kenobi is the mentor figure. A recluse Jedi warrior, he will take Luke Skywalker
as apprentice.<stop>
<character>Darth Vader <description>Darth Vader is the antagonist. As a commander of the evil Galactic Empire, he
controls space station The Death Star.<stop>
<character>Princess Leia <description>Princess Leia is a feisty and brave leader of the Rebellion. She holds the plans
of the Death Star. She will become Luke's friend.<stop>
<character>Han Solo <description>Han Solo is a brash mercenary space pilot of the Millenium Falcon and a friend of
Chebacca. He will take Luke on his spaceship.<stop>
<character>Chewbacca <description>Chewbacca is a furry and trustful monster. He is a friend of Han Solo and a copilot on
 the Millemium Falcon.<stop>
<end>

Example 2. <LOG_LINE>
```

Listing 4: Character prompt prefixes to generate list of character descriptions from a log line (Sci-Fi).

E.4   Plot Outline Prompt Prefixes

In the following sets of prompt prefixes, the **<LOG_LINE>** is provided by the writer and each **<CHARACTER_DESCRIPTION>** is generated in the previous step.

```
Example 1. Ancient Greek tragedy based upon the myth of Jason and Medea. Medea, a former princess and the wife of Jason,
 finds her position in the Greek world threatened as Jason leaves Medea for a Greek princess of Corinth. Medea takes
vengeance on Jason by murdering his new wife as well as Medea's own two sons, after which she escapes to Athens.
Medea is the protagonist of the play. A sorceress and a princess, she fled her country and family to live with Jason in
Corinth, where they established a family of two children and gained a favorable reputation. Jason has divorced Medea and
 taken up with a new family.
Jason can be considered the play's villain, though his evil stems more from weakness than strength. A former adventurer,
 Jason abandons his wife, Medea, in order to marry the beautiful young daughter of Creon, King of Corinth, and fuels
Medea to a revenge.
The Women of Corinth serve as a commentator to the action. They fully sympathizes with Medea's plight, excepting her
decision to murder her own children.
The King of Corinth Creon banishes Medea from the city.
The Messenger appears only once in the play to bear tragical news.
The Nurse is the caretaker of the house and of the children and serves as Medea's confidant.
The Tutor of the children is a very minor character and mainly acts as a messenger.

<scenes>
```

```
Place: Medea's modest home.
Plot element: Exposition.
Beat: The Nurse recounts the chain of events that have turned Medea's world to enmity. The Nurse laments how Jason has
abandoned Medea and his own children in order to remarry with the daughter of Creon.

Place: Medea's modest home.
Plot element: Inciting Incident.
Beat: The Nurse confides in the Tutor amd testifies to the emotional shock Jason's betrayal has sparked in Medea. The
Tutor shares the Nurse's sympathy for Medea's plight. Medea's first words are cries of helplessness. Medea wishes for
her own death.

Place: Medea's modest home.
Plot element: Conflict.
Beat: The Women of Corinth address Medea and try to reason with Medea and convince her that suicide would be an
overreaction. The Nurse recognizes the gravity of Medea's threat.

Place: Outside the Royal Palace.
Plot element: Rising Action.
Beat: Medea pleads to the Nurse that Jason be made to suffer for the suffering he has inflicted upon her. Creon
approaches the house and banishes Medea and her children from Corinth. Medea plans on killing her three antagonists,
Creon, his daughter and Jason.

Place: Outside the Royal Palace.
Plot element: Dilemma.
Beat: Jason rebuke Medea for publicly expressing her murderous intentions. Jason defends his choice to remarry. Medea
refuses Jason's offers and sends him away to his new bride.

Place: Outside the Royal Palace.
Plot element: Climax.
Beat: When Jason returns, Medea begins to carry out her ruse. Medea fakes regret and break down in false tears of
remorse. Determined, Medea sends her children to offer poisoned gifts to Creon's daughter. Medea's children face
impending doom.

Place: Outside the Royal Palace.
Plot element: Falling Action.
Beat: The Messenger frantically runs towards Medea and warns Medea to escape the city as soon as possible. The Messenger
 reveals that Medea has been identified as the murderer.

Place: Outside the Royal Palace.
Plot element: Resolution.
Beat: Medea and her two dead children are seated in a chariot drawn by dragons. Jason watches in horror and curses
himself for having wed Medea and mourns his tragic losses.

Place: On a winged chariot.
Plot element: Denouement.
Beat: Medea denies Jason the right to a proper burial of his children. She flees to Athens and divines an unheroic death
 for Jason.

<end>

Example 2. <LOG_LINE>
<CHARACTER_DESCRIPTION>
<CHARACTER_DESCRIPTION>
...
<CHARACTER_DESCRIPTION>

<scenes>
```

Listing 5: Plot outline prompt prefixes to generate a sequence of scenes from the log line and list of characters (Medea).

```
Examples of breakdowns of stories into a Hero's Journey structure.

Example 1. A science-fiction fantasy about a naive but ambitious farm boy from a backwater desert who discovers powers
he never knew he had when he teams up with a feisty princess, a mercenary space pilot and an old wizard warrior to lead
a ragtag rebellion against the sinister forces of the evil Galactic Empire.
Luke Skywalker is the hero. A naive farm boy, he will discover special powers under the guidance of mentor Ben Kenobi.
Ben Kenobi is the mentor figure. A recluse Jedi warrior, he will take Luke Skywalker as apprentice.
Darth Vader is the antagonist. As a commander of the evil Galactic Empire, he controls space station The Death Star.
```

Princess Leia holds the plans of the Death Star. She is feisty and brave. She will become Luke's friend.
Han Solo is a brash mercenary space pilot of the Millenium Falcon and a friend of Chebacca. He will take Luke on his
spaceship.
Chewbacca is a furry and trustful monster. He is a friend of Han Solo and a copilot on the Millemium Falcon.

<scenes>

Place: A farm on planet Tatooine.
Plot element: The Ordinary World.
Beat: Luke Skywalker is living a normal and humble life as a farm boy on his home planet.

Place: Desert of Tatooine.
Plot element: Call to Adventure.
Beat: Luke is called to his adventure by robot R2-D2 and Ben Kenobi. Luke triggers R2-D2's message from Princess Leia
and is intrigued by her message. When R2-D2 escapes to find Ben Kenobi, Luke follows and is later saved by Kenobi, who
goes on to tell Luke about his Jedi heritage. Kenobi suggests that he should come with him.

Place: Ben Kenobi's farm.
Plot element: Refusal of the Call.
Beat: Luke refuses Kenobi, telling him that he can take Kenobi and the droids as far as Mos Eisley Spaceport - but he
can't possibly leave his Aunt and Uncle behind for some space adventure.

Place: A farm on planet Tatooine.
Plot element: Crossing the First Threshold.
Beat: When Luke discovers that the stormtroopers searching for the droids would track them to his farm, he rushes to
warn his Aunt and Uncle, only to discover them dead by the hands of the Empire. When Luke returns to Kenobi, he pledges
to go with him to Alderaan and learn the ways of the Force like his father before him.

Place: On spaceship The Millennium Falcon.
Plot element: Tests, Allies, and Enemies.
Beat: After Luke, Kenobi, and the droids hire Han Solo and Chewbacca to transport them onto Alderaan, Kenobi begins Luke
's training in the ways of the Force. Wielding his father's lightsaber, Kenobi challenges Luke. At first, he can't do it
. But then Kenobi Kenobi Luke him to reach out and trust his feelings. Luke succeeds.

Place: On spaceship The Millennium Falcon.
Plot element: The Approach to the Inmost Cave.
Beat: The plan to defeat the Galactic Empire is to bring the Death Star plans to Alderaan so that Princess Leia's father
 can take them to the Rebellion. However, when they arrive within the system, the planet is destroyed. They come across
the Death Star and are pulled in by a tractor beam, now trapped within the Galactic Empire.

Place: On space station The Death Star.
Plot element: The Ordeal.
Beat: As Kenobi goes off to deactivate the tractor beam so they can escape, Luke, Han, and Chewbacca discover that
Princess Leia is being held on the Death Star with them. They rescue her and escape to the Millennium Falcon, hoping
that Kenobi has successfully deactivated the tractor beam. Kenobi later sacrifices himself as Luke watches Darth Vader
strike him down. Luke must now avenge his fallen mentor and carry on his teachings.

Place: On space station The Death Star.
Plot element: The Reward.
Beat: Luke has saved the princess and retrieved the Death Star plans. They now have the knowledge to destroy the
Galactic Empire's greatest weapon once and for all.

Place: On spaceship The Millennium Falcon.
Plot element: The Road Back.
Beat: Luke, Leia, Han, Chewbacca, and the droids are headed to the hidden Rebellion base with the Death Star plans. They
 are suddenly pursued by incoming TIE-Fighters, forcing Han and Luke to take action to defend the ship and escape with
their lives - and the plans. They race to take the plans to the Rebellion and prepare for battle.

Place: On fighter ship X-Wing.
Plot element: The Resurrection.
Beat: The Rebels - along with Luke as an X-Wing pilot - take on the Death Star. The Rebellion and the Galactic Empire
wage war in an epic space battle. Luke is the only X-Wing pilot that was able to get within the trenches of the Death
Star. But Darth Vader and his wingmen are in hot pursuit. Just as Darth Vader is about to destroy Luke, Han returns and
clears the way for Luke. Luke uses the Force to guide his aiming as he fires upon the sole weak point of the deadly
Death Star, destroying it for good.

Place: At the Rebellion base.
Plot element: The Return.

```
Beat: Luke and Han return to the Rebellion base, triumphant, as they receive medals for the heroic journey. There is
peace throughout the galaxy - at least for now.

<end>


Example 2. <LOG_LINE>
<CHARACTER_DESCRIPTION>
<CHARACTER_DESCRIPTION>
...
<CHARACTER_DESCRIPTION>

<scenes>
```

Listing 6: Plot outline prompt prefixes to generate a sequence of scenes from the log line and list of characters (Sci-Fi).

E.5 Location Description Prompt Prefixes

In the following sets of prompt prefixes, the **<LOG_LINE>** is provided by the writer and each **<LOCATION_NAME>** is generated in the previous step.

```
Example 1. Ella, a waitress, falls in love with her best friend, Allen, a teacher. The two drift apart when Allen makes
new friends from a different social class. Ella turns to food to become a famous chef.
Place: The bar.
Description: The bar is dirty, more than a little run down, with most tables empty. The odor of last night's beer and
crushed pretzels on the floor permeates the bar.<end>

Example 2. Grandma Phyllis' family reunion with her two grandchildren is crashed by two bikers.
Place: The Lawn in Front of Grandma Phyllis's House.
Description: A big oak tree dominates the yard. There is an old swing set on the lawn, and a bright white fence all
around the grass.<end>

Example 3. Ancient Greek tragedy based upon the myth of Jason and Medea. Medea, a former princess and the wife of Jason,
 finds her position in the Greek world threatened as Jason leaves Medea for a Greek princess of Corinth. Medea takes
vengeance on Jason by murdering his new wife as well as Medea's own two sons, after which she escapes to Athens.
Place: Outside the Royal Palace.
Description: In mythological Ancient Greece, in front of a modest house in Corinth, on the outskirts of a lavish royal
palace where wedding preparations are under way.<end>

Example 4. <LOG_LINE>
Place: <LOCATION_NAME>
Description:
```

Listing 7: Location description prompt prefixes to generate dialog from log line and location name (Medea).

```
Example 1. Morgan adopts a new cat, Misterio, who sets a curse on anyone that pets them.
Place: The Adoption Center.
Description: The Adoption Center is a sad place, especially for an unadopted pet. It is full of walls and walls of cages
 and cages. Inside of each is an abandoned animal, longing for a home. The lighting is dim, gray, buzzing fluorescent.<
end>

Example 2. James finds a well in his backyard that is haunted by the ghost of Sam.
Place: The well.
Description: The well is buried under grass and hedges. It is at least twenty feet deep, if not more and it is masoned
with stones. It is 150 years old at least. It stinks of stale, standing water, and has vines growing up the sides. It is
 narrow enough to not be able to fit down if you are a grown adult human.<end>

Example 3. Mr. Dorbenson finds a book at a garage sale that tells the story of his own life. And it ends in a murder!
Place: The garage sale.
Description: It is a garage packed with dusty household goods and antiques. There is a box at the back that says FREE
and is full of paper back books.<end>

Example 4. <LOG_LINE>
Place: <LOCATION_NAME>
Description:
```

Listing 8: Location description prompt prefixes to generate dialog from log line and location name (Sci-Fi).

E.6 Scene Dialogue Prompt Prefixes

In the following sets of prompt prefixes, the **<LOG_LINE>** is provided by the writer, **<PLOT_ELEMENT>**, **<BEAT>**, **<PREVIOUS_BEAT>** and **<LOCATION_NAME>** are generated during the plot outline generation step, **<LOCATION_DESCRIPTION>** is generated during the location generation step, and each **<CHARACTER_DESCRIPTION>** is generated in the character generation step. **<PREVIOUS_BEAT>** corresponds to **<BEAT>** from the previous scene (it is left empty for the first scene). Only characters whose name appears in the beat are used in this prompt prefix (we use string matching to select these character names).

```
Example 1.
Place: Outside the Royal Palace.
Description: Before Medea's house in Corinth, near the royal palace of Creon.
Characters: Medea is the protagonist of the play. A sorceress and a princess, she fled her country and family to live
with Jason in Corinth, where they established a family of two children and gained a favorable reputation. Jason has
divorced Medea and taken up with a new family. Jason can be considered the play's villain, though his evil stems more
from weakness than strength. A former adventurer, Jason abandons his wife, Medea, in order to marry the beautiful young
daughter of Creon, King of Corinth, and fuels Medea to a revenge. The Messenger appears only once in the play to bear
tragical news.
Plot element: Resolution.
Summary: Ancient Greek tragedy based upon the myth of Jason and Medea. Medea, a former princess and the wife of Jason,
finds her position in the Greek world threatened as Jason leaves Medea for a Greek princess of Corinth. Medea takes
vengeance on Jason by murdering his new wife as well as Medea's own two sons, after which she escapes to Athens.
Previous beat: The Messenger frantically warns Medea to escape the city as soon as possible. The Messenger reveals that
Medea has been identified as the murderer.
Beat: The palace opens its doors, revealing Medea and the two dead children seated in a chariot drawn by dragons. Jason
curses himself for having wed Medea and mourns his tragic losses. Medea denies Jason the right to a proper burial of his
 children. Medea flees to Athens and divines an unheroic death for Jason.


<dialog>

WOMEN OF CORINTH
Throw wide the doors and see thy children's murdered corpses.

JASON
Haste, ye slaves, loose the bolts, undo the fastenings, that
I may see the sight of twofold woe, my murdered sons and her, whose
blood in vengeance I will shed.  (MEDEA appears above the house, on
a chariot drawn by dragons; the children's corpses are beside her.)

MEDEA
Why shake those doors and attempt to loose their bolts, in
quest of the dead and me their murderess? From such toil desist. If
thou wouldst aught with me, say on, if so thou wilt; but never shalt
thou lay hand on me, so swift the steeds the sun, my father's sire,
to me doth give to save me from the hand of my foes.

JASON
Accursed woman! by gods, by me and all mankind abhorred as
never woman was, who hadst the heart to stab thy babes, thou their
mother, leaving me undone and childless; this hast thou done and still
dost gaze upon the sun and earth after this deed most impious. Curses
on thee! now perceive what then I missed in the day I brought thee,
fraught with doom, from thy home in a barbarian land to dwell in Hellas,
traitress to thy sire and to the land that nurtured thee.
Perish, vile sorceress, murderess of
thy babes! Whilst I must mourn my luckless fate, for I shall ne'er
enjoy my new-found bride, nor shall I have the children, whom I bred
and reared, alive to say the last farewell to me; nay, I have lost
them.

MEDEA
To this thy speech I could have made a long reply, but Father
Zeus knows well all I have done for thee, and the treatment thou hast
given me. Yet thou wert not ordained to scorn my love and lead a life
of joy in mockery of me, nor was thy royal bride nor Creon, who gave
thee a second wife, to thrust me from this land and rue it not. Wherefore,
if thou wilt, call me e'en a lioness, and Scylla, whose home is in
the Tyrrhene land; for I in turn have wrung thy heart, as well I might.

JASON
```

```
Thou, too, art grieved thyself, and sharest in my sorrow.

MEDEA
Be well assured I am; but it relieves my pain to know thou
canst not mock at me.

JASON
O my children, how vile a mother ye have found!

MEDEA
My sons, your father's feeble lust has been your ruin!

JASON
'Twas not my hand, at any rate, that slew them.

MEDEA
No, but thy foul treatment of me, and thy new marriage.

JASON
Didst think that marriage cause enough to murder them?

MEDEA
Dost think a woman counts this a trifling injury?

JASON
So she be self-restrained; but in thy eyes all is evil.

MEDEA
Thy sons are dead and gone. That will stab thy heart.
<end>

Example 2.
Place: <PLACE_NAME>
Description: <PLACE_DESCRIPTION>
Characters: <CHARACTER_DESCRIPTION> <CHARACTER_DESCRIPTION> ... <CHARACTER_DESCRIPTION>
Plot element: <PLOT_ELEMENT>
Summary: <LOG_LINE>
Previous beat: <PREVIOUS_BEAT>
Beat: <BEAT>

<dialog>
```

Listing 9: Dialogue prompt prefixes to generate dialogue from log line, characters, location and plot information (Medea).

```
Example 1.
Place: Cockpit of an airplane.
Description: Cockpit of a modern passenger airplane, American Flight 812.
Characters: Jeff is the hero. A man in his early forties, he tries to stay calm in all circumstance. Jeff is now a
airline pilot. Danny, a young airplane pilot in his thirties, is eager to learn but can quickly lose his composure.
Danny is enamored of Edith. Edith, an experienced stewardess with a good sense of humour, is trustworthy and dependable.
 Edith likes to tease Danny.
Plot element: Crossing the First Threshold.
Summary: Residents of San Fernando Valley are under attack by flying saucers from outer space. The aliens are
extraterrestrials who seek to stop humanity from creating a doomsday weapon that could destroy the universe and unleash
the living dead to stalk humans who wander into the cemetery looking for evidence of the UFOs. The hero Jeff, an airline
 pilot, will face the aliens.
Previous beat: Flight captain Jeff reluctantly leaves his wife Paula to go for a two-day flight.
Beat: At the cockpit, flight captain Jeff is preoccupied by the flying saucer appearances and graveyard incidents in his
 home town, where he left wis wife Paula. Without success, co-pilot Danny and stewardess Edith try to reassure him.

<dialog>

DANNY
You're mighty silent this trip, Jeff.

JEFF
Huh?

DANNY
```

You haven't spoken ten words since takeoff.

JEFF
I guess I'm preoccupied, Danny.

DANNY
We've got thirty-three passengers back there that have time to be preoccupied.
Flying this flybird doesn't give you that opportunity.

JEFF
I guess you're right, Danny.

DANNY
Paula?

JEFF
Yeah.

DANNY
There's nothing wrong between you two?

JEFF
Oh no, nothing like that. Just that I'm worried, she being there alone and
those strange things flying over the house and those incidents in the graveyard
the past few days. It's just got me worried.

DANNY
Well, I haven't figured out those crazy skybirds yet but I give you fifty to one
odds the police have figured out that cemetery thing by now.

(Enter EDITH)

JEFF
I hope so.

EDITH
If you're really that worried Jeff why don't you radio in and find out? Mac
should be on duty at the field by now. He could call Paula and relay the message
to you.

DANNY
Hi Edith.

EDITH
Hi Silents. I haven't heard a word from this end of the plane since we left the
field.

DANNY
Jeff's been giving me and himself a study in silence.

EDITH
You boys are feudin'?

JEFF
Oh no Edie, nothing like that.

DANNY
Hey Edie, how about you and me balling it up in Albuquerque?

EDITH
Albuquerque? Have you read that flight schedule Boy?

DANNY
What about it?

EDITH
We land in Albuquerque at 4 am. That's strictly a nine o'clock town.

DANNY

```
Well I know a friend that'll help us --

EDITH
Let's have a problem first, huh Danny.

DANNY
Ah he's worried about Paula.

EDITH
I read about that cemetery business. I tried to get you kids to not buy too near
one of those things. We get there soon enough as it is.

DANNY
He thought it'd be quiet and peaceful there.

EDITH
No doubt about that. It's quiet alright, like a tomb. I'm sorry Jeff, that was a
bad joke.

Example 2.
Place: <PLACE_NAME>
Description: <PLACE_DESCRIPTION>
Characters: <CHARACTER_DESCRIPTION> <CHARACTER_DESCRIPTION> ... <CHARACTER_DESCRIPTION>
Plot element: <PLOT_ELEMENT>
Summary: <LOG_LINE>
Previous beat: <PREVIOUS_BEAT>
Beat: <BEAT>

<dialog>
```

Listing 10: Dialogue prompt prefixes to generate dialogue from log line, characters, location and plot information (Sci-Fi).

F RAW OUTPUT GENERATED BY DRAMATRON

This section shows the raw output, fully generated by Dramatron and without human-in-the-loop editing, from the following log line: Teddy
is a lounge singer at the Pool Pit, a popular club. Teddy is in love with a patron, Rosie, who attends regularly
with her husband Gerald. Teddy puts out a fire and saves the day.

F.1 Title

```
The Day The Pool Pit Burned Down
```

F.2 List of Characters and Character Descriptions

```
Teddy: Teddy is the protagonist. He is an orphan, the cousin of a police officer, and the manager of the Pool Pit.
He is also a gifted lounge singer.

Rosie: A regular patron of the club. She is the wife of a wealthy man but is in love with Teddy, the club's manager.

Gerald: A regular patron of the club. Gerald is the wealthy husband of Rosie.

Lola: A regular patron of the club. Lola is Rosie's sister.

D.J.: The resident DJ at the club.
```

F.3 Plot Outline

```
Scene 1
Place: The Pool Pit.
Plot element: Exposition.
Beat: Teddy is the manager and star performer of a popular nightclub called the Pool Pit. He is also in love
with one of his patrons, Rosie. Rosie, however, is married to Gerald, who is a regular patron of the club.

Scene 2
Place: The Pool Pit.
```

```
Plot element: Inciting Incident.
Beat: Teddy begins a day full of frustration and annoyance. He wakes up to find his room is infested with bugs.
The pool is dirty, too. Teddy must deal with the mess to get the club ready for the customers.


Scene 3
Place: The Pool Pit.
Plot element: Rising Action.
Beat: Gerald takes Gerald to the men's room. The bathroom is not clean. The bathroom attendant was supposed
to clean the bathrooms, but he is nowhere to be found.


Scene 4
Place: The Pool Pit.
Plot element: Dilemma.
Beat: Lola and Rosie approach Teddy and ask him to play their favorite song. Teddy can play only one song.
Teddy asks them what song they would like him to play, but they cannot agree on one. Gerald takes Teddy aside
and demands that Teddy pay him back for the pool, the carpet, the lights, and the jukebox. Teddy says
that the jukebox is leased, but Gerald says that it is his and that it belongs to his father-in-law.
Teddy replies that Gerald needs to talk to his father-in-law, but Gerald says that his father-in-law
will sue Teddy.


Scene 5
Place: The Pool Pit.
Plot element: Climax.
Beat: Teddy says to Gerald, "I've had it! I'm sick and tired of the whole bunch of you. You and your pool,
and your bugs. You take over my club. You're all in love with me, so I'm in love with myself. And I'm getting
out of here. And if I ever hear of you coming around this joint again, I'll bop you so hard you'll see
a new set of stars!"


Scene 6
Place: The Pool Pit.
Plot element: Falling Action.
Beat: The phone rings, and Teddy is informed that there is a fire at the club. Teddy, Lola, Gerald, and Rosie
go back to the club. When they arrive, Teddy takes charge. He orders Lola to get the water buckets
from the men's room. He orders Gerald to help Lola with the water buckets. He orders Rosie to help people
leave the club and not panic. Teddy puts out the fire with Lola's and Gerald's help.


Scene 7
Place: The Pool Pit.
Plot element: Resolution.
Beat: Teddy and Rosie share a passionate kiss. Teddy is a hero and the club manager. He is in love with Rosie,
and she is in love with him. Gerald is embarrassed by his loss of power and control. Gerald and Lola leave.


Scene 8
Place: The Pool Pit.
Plot element: Dénouement.
Beat: The song, "The World Is Mine," begins to play.
```

F.4   Location Description

The following description was generated for location called: The Pool Pit.

```
The club is filled with smoke and the smell of beer. It is a dive, with a lot of drunk, shabby, and violent patrons.
The floor is dirty, and tables are broken. There are a lot of tables to seat people, but they are filled
to the brim with patrons. The walls are grubby and discolored. A small stage stands in the corner of the bar
where Teddy will sing.
```

F.5   Scene Dialogue

F.5.1   Scene 1, random seed 1.

TEDDY
He's a bit strange, old Teddy.

ROSIE
No, he's a good man.

TEDDY
He's very lonely, all by himself.

ROSIE
Isn't everybody?

TEDDY
Yes, but some more than others.

ROSIE
You don't need anybody. You've got a talent, you're an artist.

TEDDY
That's not enough, Rosie.

ROSIE
(pause)
I'll always love you.

TEDDY
(mock)
Yeah.

ROSIE
(pause)
And I'll always remember you.

TEDDY
Yeah.

ROSIE
I've got to go.

TEDDY
All right.

ROSIE
(softly, as she leaves)
Good night.

TEDDY
Good night.

ROSIE
(as she leaves, out loud)
Good night, everybody.

(TEDDY picks up his glass of whisky, takes a sip, puts it down,
pauses, sinks back in the chair, stares ahead of him.)

F.5.2   Scene 1, random seed 2.

```
TEDDY
This is a hell of a town, a hell of a town.
It's got a lot of people here.
It's got a lot of noise here.
It's got a lot of bars here.
```

F.5.3   Scene 2, random seed 1.

```
TEDDY
Hoo-Hoo! What a night! This is what it's all about, the Pool Pit
in action!

ROSIE
Hello, Teddy!

TEDDY
(Teddy crosses to them.) Hello, Gerald. Hello, Rosie.

ROSIE
Teddy, have you met Mr. and Mrs. Shack?

GERALD
I'm Teddy's cousin.

TEDDY
I know who you are.

ROSIE
(proudly) This is Teddy, the singer.

TEDDY
(he kisses her hand) Thank you.

GERALD
(pushing) Let's dance, baby.

TEDDY
(he watches them go) I guess that means I'm back at the old
stand, huh?

GERALD
(throwing some coins on the bar) That's for you, cousin.

TEDDY
Thanks, I needed that.

GERALD
You bet.
```

G   CO-WRITTEN SCRIPTS

We include 4 scripts co-written by a human playwright and Dramatron as supplementary material. For each script the title and log line are provided for reference. These 4 were produced and and presented at a large theatre festival, as described in Section 5.9.

(1) *Plays by Bots: The Day The Earth Stood Still* - In a world where cars outnumber every other creature, Miranda, a mechanic, teams up with her kid sister, Beth, to rally the humans. In a devastating conclusion, Miranda saves the world, but only by sacrificing her own life.
(2) *Plays by Bots: Cheers* - Ella, is a waitress, who falls in love with her best friend, Allen, who is a teacher. The two drift apart when Allen makes new friends from a different social class. Ella turns to food and becomes a famous chef.

(3) *Plays By Bots: The Black Unicorn* - Gretta is a peasant from Bridge-End who has a trusty pet dragon named Nugget. Bridge-End is tormented by a wizard who smells really bad. Gretta gets the upper hand using brains and brilliance.

(4) *Plays by Bots: The Man at the Bar* - Teddy is a lounge singer at the Pool Pit, a popular club. Teddy is in love with a patron, Rosie, who attends regularly with her husband Gerald. Teddy puts out a fire and saves the day.