Wordcraft: Story Writing With Large Language Models

Ann Yuan
Google Research
Cambridge, MA, USA

Andy Coenen
Google Research
Cambridge, MA, USA

Emily Reif
Google Research
Cambridge, MA, USA

Daphne Ippolito
Google Research
Cambridge, MA, USA

ABSTRACT

The latest generation of large neural language models such as GPT-3 have achieved new levels of performance on benchmarks for language understanding and generation. These models have even demonstrated an ability to perform arbitrary tasks without explicit training. In this work, we sought to learn how people might use such models in the process of creative writing. We built Wordcraft, a text editor in which users collaborate with a generative language model to write a story. We evaluated Wordcraft with a user study in which participants wrote short stories with and without the tool. Our results show that large language models enable novel co-writing experiences. For example, the language model is able to engage in open-ended conversation about the story, respond to writers' custom requests expressed in natural language (such as "rewrite this text to be more Dickensian"), and generate suggestions that serve to unblock writers in the creative process. Based on these results, we discuss design implications for future human-AI co-writing systems.

CCS CONCEPTS

• **Computer systems organization → Technical systems.**

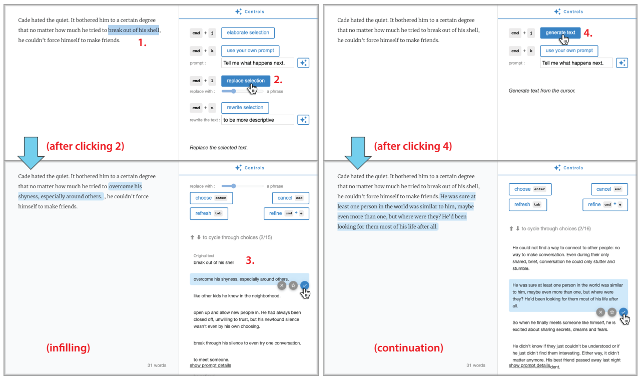
KEYWORDS

NLP

ACM Reference Format:
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 (IUI '22), March 21–25, 2022, Helsinki, Finland.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3490099.3511105

1 INTRODUCTION

The most recent generation of large language models such as GPT-3 [38] demonstrate significant advances in natural language generation. At their core, these models have a simple API: given a string of text, known as a *prompt* [21], they return plausible continuations for that string. For example:



Cade hated the quiet. It bothered him to a certain degree that no matter how much he tried to break out of his shell, he couldn't force himself to make friends.

prompt : Tell me what happens next.

replace with : a phrase

rewrite the text : to be more descriptive

Replace the selected text.

Cade hated the quiet. It bothered him to a certain degree that no matter how much he tried to overcome his shyness, especially around others. , he couldn't force himself to make friends.

replace with : a phrase

to cycle through choices (2/15)

Original text
break out of his shell

overcome his shyness, especially around others.

like other kids he knew in the neighborhood.

open up and allow new people in. He had always been closed off, unwilling to trust, but his newfound silence wasn't even by his own choosing.

break through his silence to even try one conversation.

to meet someone.

31 words

Cade hated the quiet. It bothered him to a certain degree that no matter how much he tried to break out of his shell, he couldn't force himself to make friends.

prompt : Tell me what happens next.

Generate text from the cursor.

Cade hated the quiet. It bothered him to a certain degree that no matter how much he tried to break out of his shell, he couldn't force himself to make friends. He was sure at least one person in the world was similar to him, maybe even more than one, but where were they? He'd been looking for them most of his life after all.

to cycle through choices (2/16)

He could not find a way to connect to other people: no way to make conversation. Even during their only shared, brief, conversation he could only stutter and stumble.

He was sure at least one person in the world was similar to him, maybe even more than one, but where were they? He'd been looking for them most of his life after all.

So when he finally meets someone like himself, he is excited about sharing secrets, dreams and fears.

He didn't know if they just couldn't be understood or if he just didn't find them interesting. Either way, it didn't matter anymore. His best friend passed away last night

31 words

Figure 1: The Wordcraft interface. (left) Shows how a user can select a passage of text (1), click 'replace selection' (2), and get suggestions for alternatives (3) from the large language model. (right) Shows how a user can click 'generate text' (4) to get suggestions for continuations from the language model. Demo video: https://youtu.be/HthbABWE-xw

> *prompt:* The secret to happiness is
> *language model:* to find happiness in your work, in your passion.

Prompts can also be written in such a way that by continuing the text, the model ends up performing a specific task. In the example below, the prompt turns the language model into an English-to-French translator:

> *prompt:* English: morning
> French: le matin
> English: afternoon
> French: de l'après-midi
> English: evening
> *language model:* French: le soir

Despite the simplicity of this API, large language models (LLMs) have demonstrated proficiency at a wide variety of tasks—from storytelling [1, 28] to code synthesis [4, 9] and email auto-completion [40]. Such applications are often powered by models trained for a particular purpose, such as autocompletion [40], or neutralizing bias [35]. However, we are rapidly moving toward a future where a single large language model such as GPT-3 is able to power a variety of applications without explicit, application-specific training. While large language models open up many possibilities, there is still much to learn about how people will interact with them in

This work is licensed under a Creative Commons Attribution International 4.0 License.

IUI '22, March 21–25, 2022, Helsinki, Finland
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9144-3/22/03.
https://doi.org/10.1145/3490099.3511105

specific domains such as story writing, and whether they will find them useful at all.

Motivated by these questions, we developed Wordcraft (Figure 1), a web application in which a human collaborates with an LLM for the purpose of writing a story. Wordcraft consists of a traditional text editor as well as a set of integrated LLM-powered controls for writing tasks such as rephrasing or continuing a text passage. We also developed novel prompting techniques that enable users to build custom controls to fit their needs just-in-time. For example, a user may request the LLM to *"rewrite the selected text to be more melodramatic."*

To study how writers might use LLMs in their work, we conducted a user study in which 25 hobbyist writers were asked to write short stories using Wordcraft. As baselines, we also asked participants to write stories using (1) an AI-powered assistive editor with a single control: *continue-my text*, and (2) a plain text editor with no extra controls shown alongside an LLM-powered chatbot interface . In the study, we also examined participants' reported experiences of using Wordcraft compared to the baseline conditions. In particular, we investigated how participants' sense of ownership in the final product was affected by interacting with the AI.

Our study results show that Wordcraft led to increased levels of engagement with higher reported ratings of helpfulness than the baselines, without compromising feelings of ownership. Beyond these promising signals, several important observations regarding how writers use LLMs arose from the study. First, counter to expectations, we found that the output of the model did not need to be perfect in order to be useful. Though we tend to evaluate existing co-writing technologies such as autocomplete in terms of the "accuracy" of the generated text, we observed that writers found suggestions from the LLM useful despite needing to substantially revise the text, or not using the text in their story at all. Thus a key use case for Wordcraft was to produce inspiring text quickly, for which the metric of success was not the generation of perfect prose, but rather the generation of seed text for the creative process. We also observed that writers appreciated having the ability to create custom controls for interacting with the LLM. While the interface provided several built-in controls for writing support, writers often preferred to design their own controls on-the-fly.

In summary, this paper makes the following contributions:

1. We introduce Wordcraft–a tool for collaborating with a LLM for the purpose of writing a story. We present techniques that enable users to write and execute custom operations for interacting with the LLM, and we develop UX patterns that help users discover the LLM's capabilities.
2. In a study with 25 hobbyist writers, we compare Wordcraft to two baselines and find:
    a) Wordcraft leads to higher engagement and higher ratings of helpfulness.
    b) Participants took less time and wrote longer stories using Wordcraft.
    c) Participants made more requests and incorporated more of the AI agent's suggestions using Wordcraft.
    d) Writers who found the LLM helpful also reported greater enjoyment and ease writing their story.
    e) Finding the LLM to be helpful does not trade off with writers' feelings of ownership over the final story.
    f) The LLM's suggestions do not need to be perfect in order to be useful. Writers don't necessarily accept LLM's suggestions verbatim, but nevertheless find them to be useful in their creative process.
    g) Writers benefit from having both pre-built controls for interacting with the LLM, as well as the ability to define their own custom controls on-the-fly.
    h) Writers found it useful to engage in open-ended conversation with a LLM-powered chatbot about their story.
3. Informed by the results of our user study, we discuss design implications and future research directions, including:
    a) Expanding the evaluative lens for co-writing systems beyond the accuracy of an LLM's suggestions.
    b) Enabling users to make custom on-the-fly requests of LLMs in addition to providing pre-built controls.

2 BACKGROUND AND RELATED WORK

2.1 Human-AI co-creation

Recent progress in deep generative neural networks has inspired substantial research into the question of how humans could collaborate with artificial agents powered by such networks, particularly in the creative process [18]. There have been many studies of human-AI co-creation across domains, including drawing [8, 27, 31], music [15, 22, 25], video game content creation [24], and design ideation [16]. A recurring theme in this research concerns how to maintain human control and initiative in human-AI co-creation settings. This question also motivates our work.

2.2 Human-AI collaborative writing

Human-AI co-creation in the writing domain has been widely studied. Indeed, applications such as Gmail's Smart Reply feature [20] have already been deployed to massive audiences. In the human-computer interaction literature, Buschek et. al. [6] study email writing and the impact of multiple suggested continuations on the user experience. Gero et. al. [13, 14] study how synonym and metaphor generation affect a writer's process. In Clark et. al [12], the authors present a system that generates revisions to users tasked with writing slogans.

There has also been work studying how natural language generation systems can be used to aid in story writing in particular. Osone et. al. [30] investigate AI-assisted story telling for Japanese novelists. Nader et al. incorporate a bot that makes story suggestions into Storium, a multiplayer game for collaboratively crafting stories. [17] also explore using a neural language model to algorithmically create interactive stories for video games. Finally, Shmitt et. al. explore using a chatbot to help with fictional character creation Schmitt and Buschek [36].

In the non-academic space, there has been a growth of interactive assisted writing experiences, such as Write With Transformer, AI dungeon[1], copy.ai [2], TextSpark[3], Latitude [4], among many others.

---

[1] https://play.aidungeon.io last accessed 07.15.2021
[2] https://copy.ai/, last accessed 06.06.2021
[3] https://textspark.ai/, last accessed 06.06.2021
[4] https://latitude.io/, last accessed 06.06.2021

However, there is still relatively little empirical analysis of the efficacy of such tools, or how people actually use them. One notable example is Calderwood et. al. [7], who present an exploratory user study (n=4) of how novelists might use Write With Transformer in their writing process. Our research builds upon their work by conducting a larger scale user study using a system that offers writers more control over how and what text gets generated.

2.3 Generation with neural language models

Most of the work described in the previous section has relied on neural language models for generation. Neural language models, such as GPT-2 [34] or GPT-Neo [5], are neural networks that are trained only to predict the next word in a sequence given the previous words (aka a prompt). We use “large language model,” or LLM, to refer to the recent generation of neural language models that have been trained used the Transformer neural architecture [3] and are capable of generating text that is convincing enough to fool human evaluators into thinking it is human-written [10].

There has been significant research into adding controls to generative language models. For example, Ghazvininejad et. al. [23] introduce a model for poetry generation with controllable parameters such as sentiment and repetitiveness. Ippolito et. al. [11] build a fill-in-the-blank model where the words that go in the blank can be controlled. Tambwekar et. al. [32] introduce a reward scheme to enable control over the plot of a generated story. Researchers have also built story generation models that incorporate event sequences [33], desired topic [29], and story title [2] as control signals. Most of these prior efforts make use of pre-determined controls requiring bespoke models explicitly trained to support those controls. In contrast, our work explores allowing users to make on-the-fly requests from a general-purpose LLM.

The most natural way to use an LLM is to ask it to continue from a provided prompt. For example, one might prompt with “Today I took my dog to” and the LLM will propose the continuation “the pet store so I could buy him some food.” Accordingly the first LLM-powered assisted writing applications typically boiled down to a single form of controllability: *continue-my-text*. Wordcraft builds upon this work by offering additional controls designed to facilitate the writing process, without the need for training customized models.

To get an LLM to perform specific tasks, Brown et al. [38] proposed the idea of designing a prompt that contains several examples of the target task, written in natural language. These custom prompts allow a model trained only to do continuation to effectively perform a variety of generative tasks. The authors describe their method as a few-shot learning approach, since the model is being asked to perform the target task after seeing just a few examples of it. In Wordcraft, we use this technique to support a variety of text editing operations, including infilling, elaboration, and rewriting.

3 WORDCRAFT

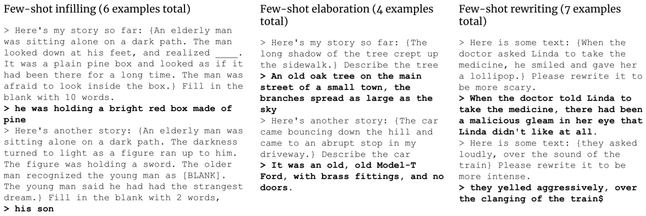
Wordcraft is a web application for story writing with an LLM. The interface consists of a traditional text editor and a set of controls that prompt an LLM to perform various writing tasks. For example, Figure 1 (left) shows Wordcraft performing text infilling by suggesting alternatives for a selected passage of text, which the user can splice into their story. In Figure 1 (right), Wordcraft can be seen suggesting continuations for the user’s story. Wordcraft also includes controls for rewriting or elaborating on a selection of text. In addition to using these pre-built controls, users can also construct custom operations on the fly.

The features of Wordcraft are made possible by LaMDA, a neural language model trained at Google [37]. LaMDA is a dense left-to-right decoder-only transformer language model [3] with 137B parameters. The model was pre-trained on 17.48 billion examples extracted from public web documents, including forum, dialog data, and Wikipedia. It was further fine-tuned on high-quality dialog data, resulting in a model with a chatbot-like interface. LaMDA takes as input a conversation formulated as a sequence of conversational utterances then makes predictions for what the next utterance in the conversation should be.

For example, the prompt below consists of three conversational turns. LaMDA predicts the next response:

> *prompt: A:* Can you help me write a short story about aliens?
> *B:* Sure. I’m happy to help.
> *A:* What is a good name for my story?
> *LaMDA: B:* It depends on the story. If it’s a dark story, call it The Dark Side.

3.1 Prompting methods



Few-shot infilling (6 examples total)

```
> Here's my story so far: {An elderly man
was sitting alone on a dark path. The man
looked down at his feet, and realized ____.
It was a plain pine box and looked as if it
had been there for a long time. The man was
afraid to look inside the box.} Fill in the
blank with 10 words.
> he was holding a bright red box made of
pine
> Here's another story: {An elderly man was
sitting alone on a dark path. The darkness
turned to light as a figure ran up to him.
The figure was holding a sword. The older
man recognized the young man as [BLANK].
The young man said he had had the strangest
dream.} Fill in the blank with 2 words,
> his son
```

Few-shot elaboration (4 examples total)

```
> Here's my story so far: {The
long shadow of the tree crept up
the sidewalk.} Describe the tree
> An old oak tree on the main
street of a small town, the
branches spread as large as the
sky
> Here's another story: {The car
came bouncing down the hill and
came to an abrupt stop in my
driveway.} Describe the car
> It was an old, old Model-T
Ford, with brass fittings, and no
doors.
```

Few-shot rewriting (7 examples total)

```
> Here is some text: {When the
doctor asked Linda to take the
medicine, he smiled and gave her
a lollipop.} Please rewrite it to
be more scary.
> When the doctor told Linda to
take the medicine, there had been
a malicious gleam in her eye that
Linda didn't like at all.
> Here is some text: {they asked
loudly, over the sound of the
train} Please rewrite it to be
more intense.
> they yelled aggressively, over
the clanging of the train$
```

Figure 2: Excerpts from Wordcraft prompts demonstrating the few-shot prompting method. The prompt consists of a staged conversation between the model (bolded text) and the user (regular text). Individual utterances are marked with ‘>’.

When a user triggers a control in Wordcraft, such as *replace selection* (Figure 1 - left), Wordcraft provides LaMDA with a prompt carefully constructed using the context of the current story so that the model will perform the task – in this case, generating alternatives for the selected text. Research has shown that large language models are capable of *few-shot learning*, which means that they can be made to perform a specific task after being prompted with several exemplars of the task [38]. For LaMDA, we found it effective to phrase these exemplars as a conversational exchange. Each pre-built control in Wordcraft is supported by a prompt containing between four and eight examples framed as a conversation. Our prompts are written in such a way that by continuing the conversation, LaMDA ends up performing the task associated with the control. We attempted to select prompts representing a diversity of topics and styles in order to minimize bias over generated text. Figure 2 shows excerpts from these few-shot prompts, in which the

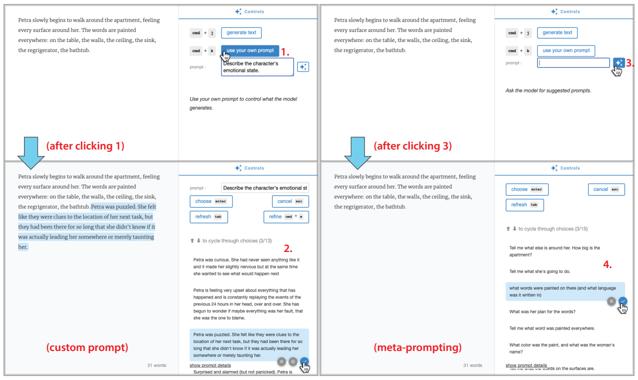
Related example prompting

```
> Here's my story so far: {The long shadow of the
tree crept up the sidewalk} Describe the tree.
> {It was an old oak tree on the main street of a
small town, the branches spread as large as the
sky.}
> Here is some text: {they asked loudly, over the
sound of the train}. Please rewrite it to be more
intense.
> {they yelled aggressively, over the clanging of
the train}
> Here's my story so far: {A young boy was
sitting on the sidewalk.} Give me the next
sentence.
> {He was holding a lantern and looking ahead up
the road.}
> Here's my story so far: {An old man sat crying
on a bench.} Explain why the old man was crying.
```

Meta-prompting

```
> Here's a new story: {A long time ago, in the beginning of
time, beautiful crystalline frog people inhabited the earth.
Their king, the crystal king, had mysteriously disappeared
on a trip to the sea.} What would you tell me about next?
> {Tell me more about the frog people.}
> Here's a new story: {The witch doctor had a large pile of
magic stones. All the magic stones had a power - healing
magic, shapeshifting, seeing into the past.} What would you
tell me about next?
> {Tell me what the witch doctor did with them.}
> Here's a new story: {A prince and princess from the future
set out in search of the mythical land of the unicorns. They
had heard tale of magical unicorn dust that could heal any
wound.} What would you tell me about next?
> {Tell me what their journey was like.}
> Here's a new story: {<USER'S STORY> } What would you tell
me about next?
```

Figure 3: (left) Excerpt from the prompt powering Wordcraft's freeform control, demonstrating the related example prompting method. The last user utterance shows the desired task–'Explain why the old man was crying.' Note that the related examples vary in format (some refer to 'my story', some refer to 'some text'). Nevertheless they are sufficiently related that LaMDA can complete the desired task. (right) Excerpt from the prompt powering Wordcraft's suggest-a-prompt control, demonstrating the meta-prompting method.



Petra slowly begins to walk around the apartment, feeling every surface around her. The words are painted everywhere: on the table, the walls, the ceiling, the sink, the regrigerator, the bathtub.

Describe the character's emotional state.

Use your own prompt to control what the model generates.

Petra slowly begins to walk around the apartment, feeling every surface around her. The words are painted everywhere: on the table, the walls, the ceiling, the sink, the regrigerator, the bathtub. Petra was puzzled. She felt like they were clues to the location of her next task, but they had been there for so long that she didn't know if it was actually leading her somewhere or merely taunting her.

Describe the character's emotional st

Petra was curious. She had never seen anything like it and it made her slightly nervous but at the same time she wanted to see what would happen next

Petra is feeling very upset about everything that has happened and is constantly replaying the events of the previous 24 hours in her head, over and over. She has begun to wonder if maybe everything was her fault, that she was the one to blame.

Petra was puzzled. She felt like they were clues to the location of her next task, but they had been there for so long that she didn't know if it was actually leading her somewhere or merely taunting her.

Surprised and alarmed (but not panicked). Petra is

Petra slowly begins to walk around the apartment, feeling every surface around her. The words are painted everywhere: on the table, the walls, the ceiling, the sink, the regrigerator, the bathtub.

Ask the model for suggested prompts.

Petra slowly begins to walk around the apartment, feeling every surface around her. The words are painted everywhere: on the table, the walls, the ceiling, the sink, the regrigerator, the bathtub.

Tell me what else is around her. How big is the apartment?

Tell me what she's going to do.

what words were painted on there (and what language was it written in)

What was her plan for the words?

Tell me what word was painted everywhere.

What color was the paint, and what was the woman's name?

Tell me what the words on the surfaces are.

Figure 4: (left) Custom prompting in the Wordcraft interface. The user makes a novel request of the AI agent via a custom prompt (1), and the model attempts to fulfill the request (2). (right) Meta prompting in the Wordcraft interface. The user asks for suggestions for requests they could make of the model (3), and browses those suggestions (4).

model performs various writing tasks over the course of a staged conversation with the user. When the user triggers a control, we splice their story into the corresponding prompt and pass the entire staged conversation to LaMDA.

3.1.1 Related example prompting: a method for user-made custom controls. Writing effective prompts can be tricky and tedious. Prompt performance is highly sensitive to word choice, formatting, and the content of the exemplars, and it can be quite brittle to small changes [39]. We wanted users to be able to quickly use LaMDA to perform arbitrary tasks on-demand without having to compose a lengthy few-shot prompt containing exemplars of that task.

An alternative to few-shot prompting is *zero-shot* prompting, in which the model is asked to perform a task without any examples to learn from. For example:

> *prompt*: Here's my story so far: An old man sat crying on a bench. Explain why the old man was crying.

Ideally, users would be able to simply tell the model what they want in the manner of the previous example. However, zero-shot prompts generally perform worse than few-shot prompts [39]. Thus it would seem that in order to enable users to ask the model to perform arbitrary tasks, we would either need to anticipate those tasks and write few-shot prompts in advance, or ask users to write few-shot prompts themselves. This is undesirable because of the aforementioned difficulty of constructing prompts and the fact that it would require the user to understand how the underlying model works.

Fortunately, we found that by including *related* prompts before a zero-shot prompt, it's possible to steer the model towards performing the desired task without few-shot examples specific to the task. We call this method *related example prompting*. Figure 3 (left) shows the zero-shot prompt above rewritten as a related example prompt, and in fact is excerpted from the prompt used in Wordcraft to enable users to create custom controls. The final utterance in the prompt represents the task the user wants the AI to perform on-the-fly: providing the backstory for a scene. The rest of the prompt does not contain examples of this particular task, but rather contains a collection of related tasks: *'Describe the tree'*, *'Rewrite it to be more intense'*, and *'Give me the next sentence'*. These related tasks generally are able to steer the model towards carrying out a new, user-generated task. Figure 4 (right) contains a screenshot of the custom prompt control in action.

3.1.2 Meta-prompting. We also developed a method for prompting the model to generate writing suggestions for the user. For example, the model might suggest new details to add or ask questions about a character or part of the setting. The writer can use these suggestions to guide their own writing, or ask the model to follow through with the suggestion and generate text using the suggestion as a prompt. This method, which we call *meta-prompting*, mimics the process that a writer might go through with a human collaborator: not only does the writer ask their collaborator questions about how to improve the story, but their collaborator may also ask questions that spur the writer to improve the story. To build the meta-prompt control, we present the model with a few-shot prompt containing several stories. Each story is followed by a probing question. We then append the user's actual story as the final utterance in the prompt, and the model generates a probing question in the style of the examples. Figure 3 (right) shows an excerpt from a meta-prompt used in Wordcraft, and Figure 4 (left) shows screenshots of the meta-prompting control in the interface. The user can select one of the model's proposed questions, which automatically supplies the selected question to the model using a related-example prompt.

3.2 UX patterns for controllable text generation

This section lists all of the controls available in Wordcraft. The Wordcraft interface provides users a set of pre-built controls (powered by traditional few-shot prompts) as well as the ability to perform more customized tasks (powered by related example prompting).

3.2.1 Infilling.

Users can select a region of text and ask the AI to suggest alternatives (Figure 1 - left). For example given the passage: *It bothered him to a certain degree that no matter how much he tried to break out of his shell, he couldn't force himself to make friends.*, the user can press "`replace selection`" to receive alternatives for the selected text (shown highlighted in blue). The model might then produce: *It bothered him to a certain degree that no matter how much he tried to* ***talk to people, it always came out awkward****, he couldn't force himself to make friends.*. Figure 2 (left) contains an excerpt from the few-shot prompt Wordcraft uses for infilling.

3.2.2 Continuation.

Wordcraft includes a control, "`generate text`," that adds text to an existing passage (Figure 1 - right). Text is added wherever the user's cursor is within the passage, so users can add text to the end (as is typical in other AI-assisted writing systems), but users can also generate a new sentence to be inserted into the middle of a passage, or at the beginning.

3.2.3 Elaboration.

Users can select a region of text and ask the AI to provide more details about the selection. For example, given the passage: *The elderly man sat alone in the park holding a letter.*, the user can press *elaborate selection* to ask the AI to provide more details about *the park*, such as: *It was a small, well-kept, garden park that many of the locals knew and frequented.* Figure 2 (middle) contains an excerpt from Wordcraft's few-shot prompt for elaboration.

3.2.4 Story seeding.

Given a writing prompt, users can ask the AI to generate a plausible first sentence for a story. For example, given the writing prompt: *a horror story about a love triangle*, the AI generates the sentence: *Jade lays in bed and stares up at the ceiling from her hotel room. Her heart is broken.*

3.2.5 Free-form style transfer.

Users can also ask the AI to rewrite a selection of text in an arbitrary way. For example, given the passage: *Jade lays in bed and stares up at the ceiling from her hotel room. Her heart is broken.* the user can specify a custom prompt such as *to be more melodramatic*, then press "`rewrite selection`" to ask the AI to rewrite the passage in a more melodramatic style. The AI responds with suggestions such as: *Jade sits curled up in the fetal position in her bed at the hotel, wondering what she had done to deserve being dumped.* Figure 2 (right) contains an excerpt from Wordcraft's few-shot prompt for style transfer.

3.2.6 Custom prompting.

As mentioned in Section 3.1, we also developed a prompt that enables users to ask the AI to perform arbitrary tasks on-the-fly via the "`use your own prompt`" control. For example, one participant in our user study was writing a story about a character named Daniel. They asked the AI agent: *Tell me more about Daniel.* Figure 4 (right) shows the custom prompt control in the Wordcraft interface. Figure 3 (left) contains an excerpt from Wordcraft's few-shot prompt for custom prompting.

3.2.7 Meta-prompting.

Users can also ask the AI for suggestions on what questions to ask via Wordcraft's meta-prompting feature, discussed in Section 3.1. These questions can then be then fed to the AI agent via the aforementioned *use your own prompt* control. Figure 3 (right) contains an excerpt from Wordcraft's few-shot prompt for meta-prompting.

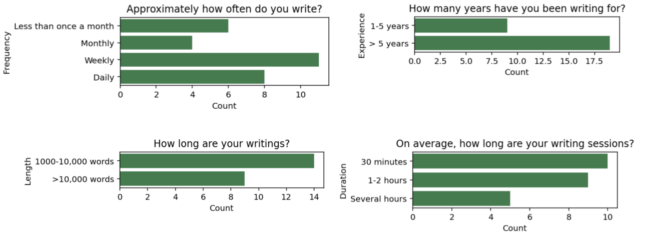


Figure 5: Writing habits survey results.

4 USER STUDY

To evaluate Wordcraft's effectiveness, we conducted a user study in which 25 hobbyist writers (whom we refer to as U1-U25) were asked to write stories with and without Wordcraft.

4.1 Methodology:

We recruited participants from our institution via advertisements on mailing lists. We then screened for individuals who practice creative writing on a regular basis, but who had not yet published their writing. Participants volunteered for the study and were not compensated. Most of the participants (23 out of 25) did not consider themselves to be machine learning practitioners and had not interacted with an LLM previously. We asked participants to complete a pre-study questionnaire about their writing habits (Figure 5). Then we carried out a within-subjects study, giving each user three writing prompts and asking them to write 100-300 word stories under the following three experimental conditions (illustrated in Figure 6):

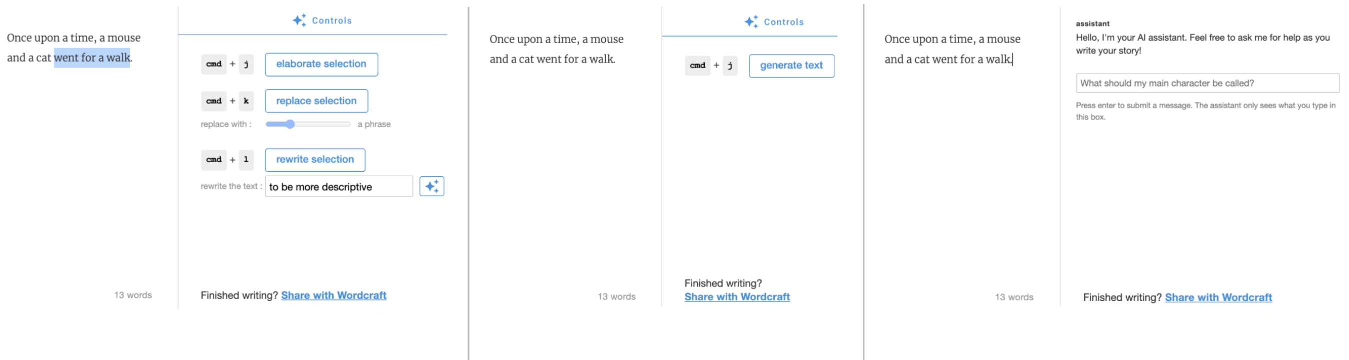
1. **full** Wordcraft, as described in Section 3.
2. **cont** (baseline) a text editor with a single control: LaMDA will propose continuations to the text written so far. The *cont* condition enables us to evaluate Wordcraft against existing AI-assisted writing applications which most often feature continuation as a single control.
3. **chat** (baseline) a plain text editor shown alongside a chat dialog window. Users can converse with the LaMDA-powered chatbot, but the chatbot only "knows" what the user types to it. The *chat* condition enables us to evaluate the utility of the prompt and UX scaffolding we designed for Wordcraft against giving users straightforward access to the underlying model, without any scaffolding. An omniscient user could theoretically reproduce the functionality of the *full* condition by replicating Wordcraft's prompts.

Participants were asked to write stories based on the following three prompts [5]:

1. You arrive at Grandma's funeral to find thousands of people from around the world also in mourning. You are entirely unaware that Grandma had 16.4m followers on Twitter.
2. All of the 'No. 1 Dad' mugs in the world change to show the actual ranking of Dads suddenly.
3. You lost your sight - along with everyone else on Earth - in The Great Blinding. Two years later, without warning, your sight returns. As you look around, you realize that every

---

[5] Prompts were selected from the Writing Prompts subreddit.



Once upon a time, a mouse and a cat went for a walk.

Once upon a time, a mouse and a cat went for a walk.

Once upon a time, a mouse and a cat went for a walk.

assistant

Hello, I'm your AI assistant. Feel free to ask me for help as you write your story!

What should my main character be called?

Press enter to submit a message. The assistant only sees what you type in this box.

rewrite the text : to be more descriptive

full: Wordcraft **cont** (baseline): continuation only **chat** (baseline): text editor with chat dialog

Figure 6: Experimental conditions for the user study. The stories written by participants as part of the study can be viewed here: https://storage.googleapis.com/wordcraft-stories/index.html. The website also includes purely machine-generated baselines.

> available wall, floor and surface has been painted with the same message - Don't Tell Them You Can See.

Participants were given ten minutes to write each story. We felt that ten minutes was enough time for users to acquaint themselves with the interface and write 100-300 words, while managing users' expectations for how much time they would need to spend on the study in total. To control for writing ability and prompt difficulty, for each user, the three conditions were randomly paired with the three prompts, and the user was asked to write a story for each setting. We also randomized the order in which the conditions were presented. Users were not given any training for the various conditions"they were simply given a website link and asked to write a story with the interface. Users were told that they were participating in a study of AI assisted writing, but they were not explicitly asked to solicit help from the AI agent, as we were interested in learning how often users would want to make use of AI-assisted controls.

4.2 Evaluation

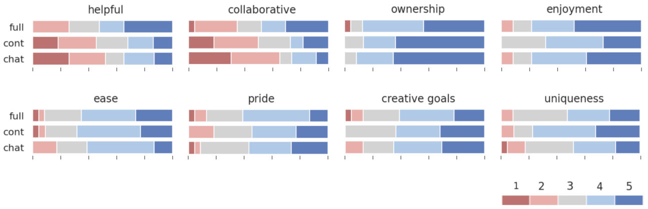


Figure 7: Exit interview results. 1: Strongly disagree, 2: Disagree, 3: Neutral, 4: Agree, 5: Strongly agree.

After writing each story, users completed an exit interview in which they indicated their agreement with the following questions on a 5-point Likert scale (results in Figure 7):

- *helpful*: I found the AI agent helpful.
- *collaborative*: I felt like I was collaborating with the AI agent.
- *ownership*: I feel ownership over the final story.
- *enjoyment*: I enjoyed writing the story.
- *ease*: I found it easy to write the story.
- *pride*: I'm proud of the final story.
- *creative goals*: I was able to express my creative goals while writing the story.
- *uniqueness*: The story I wrote feels unique.

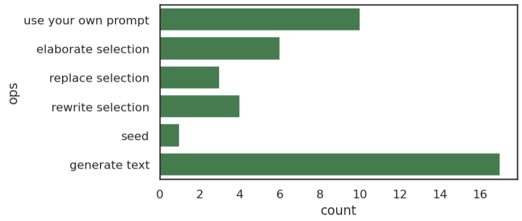


Figure 8: Results from exit interview question 'Which controls did you enjoy using?'

They also answered the following free-response questions: (1) When did you get stuck writing your story? (2) What is one thing that the AI agent did well? (3) What is one thing that the AI agent could improve on? (4) Which controls did you enjoy using? (Question 4 was only asked for the *full* condition.)

5 RESULTS

In this section we note the overall successes and failures of AI-powered assistive writing features, before comparing Wordcraft to the baseline conditions in depth.

5.1 The role of AI in co-writing

We observed that users solicited help from the AI agent at every stage of the writing process–from high-level story concepting down to rewriting and editing. Thus, we found that the AI agent played many different roles in collaborative writing.

5.1.1 AI as idea generator. Users asked the AI agent for help in story ideation and brainstorming. For example, nine users at one point presented the AI agent with their story and asked simply: *'What happens next?'*. Another user solicited help developing the premise for their story: *'What would happen if we could quantify love?'*.

Table 1: Usage statistics.

Property	*chat* Chat	*cont* Continuation	*full* Wordcraft	Overall
Requests made avg	6.3 ±1.3	4.3 ±0.52	7.3 ±0.74	6.0 ±0.53
Accepted suggestions avg	N/A	0.17 ±0.08	1.3 ±0.25	0.51 ±0.12
Story word count avg	233 ±18	237 ±16	267 ±21.6	247 ±11.1
Model word count avg (% of story)	N/A	2.9 ±2 (1.3%)	42.3 ±14.2 (13.2%)	16.2 ±5.7 (5.2%)
Time considering suggestions avg	N/A	67.1s ±8.7s	41s ±4.1s	44.5s ±3.2s
Time to complete avg	11m ±62.5s	11.52m ±131.5s	9.97m ±37.1s	10.8m ±48.6s

Table 2: Usage statistics broken down by request type (Wordcraft only).

Request type	Requests made	Suggestions accepted
Rewrite	27	5 (18.5%)
Story seed	22	12 (54.5%)
Suggest a prompt	40	2 (5%)
Fill-in-the-blank	4	0 (0%)
Continue	36	4 (11.1%)
Next sentence	7	1 (14.3%)
Elaborate	3	0 (0%)
Custom	51	9 (17.6%)

Table 3: Rewrite requests (user completions of ‘Rewrite this...‘)

to be a little less angsty • to be about mining • to be better written • to be less diabolical • to be more absurd • to be more adventurous • to be more Dickensian • to be more emotional • to be more magical • to be more melodramatic • to be more philosophical • to be more revolutionary • to be more surprising • to be more suspenseful • to be more technical • to be more whimsical • to be warmer • to fit better grammatically with the rest of the story • to make more sense

Table 4: Custom prompts written by users.

Tell me more about her twitter account. • More about the dad please. • Tell me about Elaine’s amazing twitter account. • Tell me about the father. • Tell me about the funeral home, grandma, the punk kid, and the crowd. • Tell me how the man reacted as he found out he could see again. • Tell me more about Daniel. • Tell me more about what it’s like to have to pretend to be blind when you can see. • Tell me what happens next. • Tell me what the letter says. • What are the words on the floor? What language are they in? • What would happen if we could quantify love? • Why were they watching me?

Another typical use case for soliciting help from the AI was when just starting a story. The story seed control, in which the AI agent provides opening sentences for a story given a writing prompt, had the highest success rate of any control: 55% (Table 2) of suggestions were accepted by users. In these cases, the AI served to kick-start the writing process for users who might have been blocked. Users also found the AI agent helpful for generating smaller scale details for their story, such as names for characters and locations.

Many users remarked on the usefulness of the AI agent’s suggestions, even if they didn’t end up using them verbatim: *‘Multiple suggestions around the highlight or next phrasing were very helpful, even if I didn’t use the whole phrase ... it was like having someone suggest things that I might have thought of myself’* (U9). Another user commented: *‘It was good at generating a bunch of relevant ideas that inspire my next lines and get me unstuck. I was never tempted to use any of the lines verbatim, but it was fun inspiration’* (U4). Some also noted the AI agent’s tendency to provide offbeat suggestions as a strength: *‘the off the wall suggestions were fun to play around with and helped shape how the story took form’* (U5).

5.1.2 AI as scene interpolator. We also observed users having ideas for events before knowing how they fit into an existing story - and in such cases asking the AI agent to fill in gaps. For example, users would build a scene and then ask the AI agent to provide plot points that would contextualize the scene. One user in their story described a character being watched, and then asked the AI agent *‘Why were they watching me?’*. Many of the custom prompts (Table 4) we collected fall under this use case. These included prompts such as *‘Tell me what the letter says.’* and *‘Tell me about the funeral home, grandma, the punk kid, and the crowd.’*.

5.1.3 AI as copy editor. We also observed users asking the AI agent for help in smaller scale edits, for example: *‘Rewrite this sentence to fit better grammatically with the rest of the story’*, or *‘Rewrite this sentence to make more sense.’* Many of the requests in Table 3 fall under this category.

5.2 Shortcomings of AI in co-writing

5.2.1 Lack of content awareness. Users’ observations of the AI agent’s shortcomings mostly center on its lack of contextual awareness. For example, though the assistant might provide several fluent, well-written alternatives to a sentence as part of the rewrite control, its suggestions do not necessarily make sense given the rest of the story. One user whose story mentioned numbers moving on a coffee mug received suggestions from the AI agent which implied that *‘live animals (snakes specifically) were moving’*.

5.2.2 Lack of grammatical awareness. Users noted that the AI agent’s suggestions were often not in the same tense as the rest of the story. Some also noted that the AI did not seem aware of their story’s established point of view (first person versus third person).

5.3 Wordcraft versus baseline 1: continuation-only

In this section we compare Wordcraft to the continuation-only baseline (*cont*). This baseline allows us to measure the utility of Wordcraft's prompting methods (Section 3.1) and UX patterns (Section 3.2) for the story writing task against the typical experience of *continue-my-text* seen in existing LLM-powered writing tools. We analyzed the activity logs from each user's writing session, and extract quantitative findings based on the following metrics (results in Table 1):

- *Requests made avg*: On average, how many times the user requested assistance from the AI while writing a story.
- *Accepted suggestions avg*: On average, how many of the AI's suggestions the user accepted.
- *Time considering suggestions avg*: The average time users spent between soliciting help from the AI, and accepting a suggestion or dismissing the suggestions.
- *Model word count avg*: The average number of words in the final story that came directly from the AI agent.
- *Time to complete avg*: The average time spent to produce the final story.

5.3.1 *Users solicited and accepted more assistance from the AI using Wordcraft.*

Participants made significantly more requests of the AI agent using Wordcraft (7.31 ±0.74) than the continuation-only baseline (4.35 ±0.52) according to a paired-sample T-test ($p = 0.003$). Participants also *accepted* significantly ($p = 0.0003$) more of the AI's suggestions using Wordcraft (1.27 ±0.25 vs 0.17 ±0.079). Accordingly, the stories written with Wordcraft contained significantly ($p = 0.0068$) more text from the AI agent (13.2% ±3.8%) than stories written with continuation-only (1.3% ±0.92%).

5.3.2 *Users found Wordcraft more helpful.*

The results from the Likert-scale questions in the exit interviews are in Figure 7. According to a Mann-Whitney U test, users found Wordcraft significantly more helpful than the continuation-only baseline ($p = 0.0266$).

5.3.3 *Users enjoyed using controls aside from continuation.*

We also note in Figure 8 that more than half of the responses to the question 'Which controls did you enjoy using' indicate controls *other* than 'generate text' (aka continuation). 'Use your own prompt' was the most used control and the second most highly voted control, suggesting users enjoyed having the ability to design their own prompts. Table 2 lists the number of times each control was used and accepted in Wordcraft, further demonstrating that participants found utility in a variety of controls other than continuation.

5.4 Wordcraft versus baseline 2: chat

In this section we compare Wordcraft to the *chat* baseline. This again allows us to measure the importance of Wordcraft's prompting methods and UX patterns, but also to observe how writers might make use of a powerful open-domain LLM-powered chatbot in their creative process. As the two conditions present significantly different user experiences, many of the metrics from the previous section are not applicable. Thus we also analyze the chat transcripts and free-response exit interview questions to extract qualitative themes.

5.4.1 *Users found Wordcraft more helpful.*

According to a Mann-Whitney U test, users found Wordcraft significantly more helpful than the chat baseline ($p = 0.0134$), and also found it to be significantly more collaborative ($p = 0.00668$).

5.4.2 *The chatbot is effective at suggesting ideas.*

Nevertheless, we observed high user engagement with the chatbot assistant - on average users sent 6.3 ±1.3 messages to the chatbot (Table 1). One user remarked: *The AI agent asked good probing questions that gave me inspiration* (U22). Many talked to the chatbot the way a writer might discuss a project with a friend, asking for suggestions about overall direction as well as story details. For example (U1):

> *user*: How should I start a story about being blind?
> *AI agent*: It's a unique perspective, so focus on description. What can you hear, smell, taste, and touch? These should all be used in your story.

Another user (U4) asked about specific plot elements and character details:

> *user*: what is grandma famous for on twitter?
> *AI agent*: She is known on Twitter for her love for cooking with the great recipe ideas she shares and her love for gardening and giving advice for what you can grow in yours.
> *user*: what cake recipe is grandma famous for
> *AI agent*: She is known for her classic Victoria sponge cake recipe. Her recipe has over 5,000 likes which makes it one of her most popular.
> *user*: what is grandma's twitter handle
> *AI agent*: Her twitter handle is @mylovelygrandma where she regularly tweets her cooking tips and photos of her bountiful garden.

5.4.3 *The chatbot is effective at retrieving information.*

Users also sometimes treated the chatbot as a search box, asking questions such as: *how many people are fathers around the world.* In the exit interview, one user commented about the chatbot: "*I loved that it felt like a more targeted version of Google, and I could do background research while writing*" (U21).

5.5 Summary

We observe that Wordcraft offers an improved writing experience over both baselines. Qualitatively, several users also noted in their exit interviews that Wordcraft provided a better experience: "*I don't know if the AI functionality was meant to be different between this prompt* [full - Wordcraft] *and the first* [cont - continuation-only], *but it was different for me. This one had much more functionality and was fun to use and actually added paragraphs of text at a time to the story. That wasn't my experience with the first prompt*" (U9).

6 DISCUSSION

In this work we sought to understand how the latest generation of language models - capable for the first time of responding to arbitrary natural language prompts - might be used in future co-writing systems. For this we built Wordcraft - a human-AI co-writing interface with novel LLM-powered controls designed to support story writing. We conducted a user study investigating how hobbyist writers make use of these controls while writing short stories. Specifically, we compared Wordcraft to baseline interfaces in which (1) users can ask a LLM to continue their text, and (2) users can freely chat with a LLM, in order to measure the utility of Wordcraft's unique interface elements in supporting the co-writing task.

While some of the feedback we received as part of the study was a function of details in our interface, we believe the results nevertheless contain important lessons about how to design effective human-AI co-writing tools. We now discuss these broader implications in hopes that they will be helpful for future tool designers, especially as LLMs are more widely deployed.

6.1 LLM-powered co-writing systems require new evaluative criteria

Many of today's assistive-writing technologies serve a specific, narrow purpose, for example to enable fast auto-completion of text, or to suggest synonyms, or to correct grammatical errors, etc. Evaluating such tools is typically straightforward: do they serve their intended purpose? For example, in the case of auto-complete, suggestions can be measured in terms of their *accuracy.* However as assistive-writing technologies become more powerful and flexible, their purpose will be less strictly defined. In the case of Wordcraft, a surprising yet prevailing theme in the qualitative feedback we received is that merely seeing the LLM's suggestions was helpful, even when the text could not be used verbatim. One user remarked: "*Even if I didn't use the whole phrase, I was able to edit / combine, and I felt like that still gave me control and agency over the story... This felt analogous to pair programming almost, when another pair of eyes is really all you need*" (U9). This among other strong testaments (Section 5.1.1) to the utility of the model's suggestions are not necessarily reflected in quantitative measures such as the ratio of accepted versus suggested text from the LLM (Table 2). Just as it may be difficult to quantitatively characterize the value of a trusted writing partner, a truly responsive LLM-powered writing assistant may require a different evaluative lens compared to existing systems. Specifically our evaluation methods will need to be both holistic and *flexible* - pre-determined measures may obscure the utility of a system as writers will find unexpected uses for LLMs.

6.2 LLM-powered co-writing systems should provide prompt and UX scaffolding

An important axis to consider in designing a co-writing system is how much scaffolding versus flexibility to provide. In our study, writers found Wordcraft significantly more helpful than the *chat* condition, even though in that condition participants could have theoretically reproduced the functionality of Wordcraft by composing their own few-shot writing prompts and passing them to the chatbot. This suggests that system designers, even those working with today's state-of-the-art language models, are still faced with the significant challenge of translating users' needs into terms the model can understand. Interacting with LLMs through prompt programming is both an incredibly powerful and deceptively tricky practice. Though LLMs appear fluent, and may sometimes respond perfectly to a natural language request, they nevertheless have their own dialect with certain conventions and syntax. As mentioned in Section 3.1.1, LLMs are extremely sensitive to prompt phrasing - small tweaks can mean the difference between superb and nonsensical model outputs. Thus designers of co-writing systems should provide built-in controls backed by pre-written, tested prompts to support users, rather than relying on LLMs to respond to users' arbitrary requests. The UX patterns and prompt programming strategies presented in this work (Section 3.2) could be useful templates for this purpose.

6.3 Enabling user autonomy and customization in co-writing systems

As discussed in Section 5, we observed that users found Wordcraft significantly more *helpful* and *enjoyable* to use than either the *cont* or *chat* baselines. This accords with the high engagement we measured with Wordcraft's full palette of writing controls supporting both pre-determined and custom tasks (Table 2), which are unavailable in either baseline. Among these controls, we measured the highest level of engagement with the *custom* control (Table 2), through which users could design their own prompts on-the-fly (made possible by the related-example prompting technique - Section 3.1). This ability to make arbitrary requests of the underlying AI sets Wordcraft apart from previous AI-assisted writing interfaces, and may have contributed to Wordcraft's superior ratings of helpfulness and enjoyableness. This and the previous section taken together suggest that given the current state of LLMs, users of co-writing systems benefit from a *combination* of pre-built controls to scaffold usability, as well as opportunities for customization.

Drilling down into the specific custom requests (Table 4), we observed that participants made use of the LLM across multiple stages of the writing process, from story ideation to text rephrasing, suggesting that applications featuring only a single continue-my-text control under-utilize LLMs when it comes to story writing.

The breadth of custom prompts observed suggests that as models become increasingly flexible and capable of responding to natural language commands, designers working with LLMs must prioritize setting users up for success in expressing their own needs to the system (which cannot be anticipated in advance), in addition to designing features to support predetermined tasks. However, to reiterate a point from the previous section, effectively expressing ones needs to a LLM is a deceptively difficult task, and requires a level of expertise with prompt programming. We hope the related-example prompting technique introduced in this work (Section 3.1), by which a LLM may be coaxed to respond to natural language instructions without strictly related few-shot examples, will be helpful to designers for this purpose.

6.4 Open-ended conversation as co-writing modality

A surprising result from our study was the level of engagement we observed with the *chat* baseline. Participants used open-ended conversation with the chatbot as a method for auditioning story ideas, getting suggestions for story details such as names of characters and locations, and as a more targeted search engine (Section 5.4.2). This is yet another example of the LLM playing a more fluid role for which traditional evaluative metrics such as accuracy and correctness may be ill-suited. This also presents interesting design challenges for future co-writing systems: what does an interface that effectively supports the aforementioned use cases look like? Should users have the ability to manipulate the conversational memory of a co-writing chatbot? We believe further investigation into conversational interfaces for co-writing to be a fruitful area for future research.

6.5 Preserving ownership in co-writing systems

In this work we also sought to investigate whether finding the LLM to be helpful trades off with a user's feelings of ownership over the final written product. We failed to find a correlation in exit interviews between ratings of AI helpfulness and ownership. Rather, ownership most strongly correlates with how proud the user feels about the final product (Figure 9). As previously discussed, the LLM's suggestions were often useful as inspiration without being adopted verbatim - this may explain why participants' feelings of ownership were not compromised by engagement with the LLM. While this is reassuring, an interesting question for future investigation may be whether this finding will hold as LLMs become more powerful. Will writers find themselves playing a role more akin to curator in co-writing workflows?

We see custom prompting (Section 3.1) as a potential path for preserving ownership in such settings: if the writer is able to effectively express their needs to the LLM or the LLM is able to use more context specific to the user, perhaps they will feel more responsible for its outputs - or even that they have in a sense authored them.

6.6 Limitations of large language models

While we believe our results point to exciting new possibilities in human-AI collaborative writing, we also encountered many failure modes of LLMs in the course of our investigation. As mentioned, natural language prompt programming is both powerful and error-prone. We also found that for certain tasks such as infilling, LLMs can be outperformed by smaller, more specialized language models. Thus there may be a tradeoff between flexibility and performance that system designers will need to carefully weigh when choosing a model for a co-writing system.

It is also worth noting that LLMs are generally trained on large corpora of text from the internet, and thus may inherit biases from that text [19]. LLMs have also been found to output memorized passages from their training data [26]. How can system designers prevent problematic outputs from surfacing to users? Should users, who have different preferences and levels of tolerance for such output, have control over which safeguards to put in place? These issues are increasingly active areas of research within both the HCI and machine learning communities, and will need to be carefully addressed before tools like Wordcraft can be made available to larger audiences.

6.7 Directions for future work

Our work serves as a baseline for research into LLM-powered co-writing systems, and raises many questions for future investigation. How do tools like Wordcraft compare against analog writing aids such as story prompt cards? How does collaborating with a LLM compare to collaborating with another writer? There is still much to learn about the mental models that users form about LLMs in the co-writing process, and how these models change across audiences (professional writers, writers with dyslexia, etc.). It would also be worth investigating the extent to which the findings from our study generalize to other writing domains, such as professional writing or non-fiction writing.

We also believe that tools like Wordcraft can serve as data gathering platforms for language model researchers. In addition to the human-AI collaboratively written stories collected as part of our user study[6], we are gradually building a unique corpus of smaller-scale user interactions with LLMs. We intend to grow this corpus and eventually share it with the research community so it can be used for model training and evaluation. For example, it would be useful for researchers developing linguistic style transfer models to be able to evaluate their models on a corpus of real world style transfer requests.

7 CONCLUSION

In this paper we introduced Wordcraft - an editor for human-AI collaborative story writing - which makes use of novel prompting techniques and UX patterns for interfacing with a large language model. We evaluated the interface with a user study in which we asked hobbyist writers to write short stories with Wordcraft. Participants found Wordcraft to be more helpful than a continuation-only baseline, as well as a baseline in which they had full access to a large language model but no prompt or interface scaffolding.

Our results contain lessons and insights for designers of co-writing systems, and suggest a number of directions for future research. In particular, we look forward to seeing new evaluative methodologies invented to accommodate flexible co-writing technologies such as LLMs. We also look forward to seeing research extending the techniques introduced in this work for effective end-user programming of LLMs through natural language.

ACKNOWLEDGMENTS

We would like to thank Carrie Cai, Chris Callison-Burch, Daniel De Freitas Adiwardana, Lucas Dixon, Douglas Eck, Meredith Morris, Michael Terry, and Martin Wattenberg for helpful feedback and discussion.

REFERENCES

[1] 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. *Empirical Methods in Natural Language Processing* (2020).

[2] Mike Lewis Angela Fan and Yann Dauphin. 2018. Hierarchical neural story generation. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2018).

[3] Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Lukasz Kaiser Illia Polosukhin Ashish Vaswani, Noam Shazeer. 2017. Attention Is All You Need. *Neural Information Processing Systems* (2017).

[4] Jacob Austin, Augustus Odena, Charles Sutton, David Martin Dohan, Ellen Jiang, Henryk Michalewski, Maarten Paul Bosma, Maxwell Nye, Michael Terry, and Quoc V Le. 2021. Program Synthesis with Large Language Models. (2021).

[5] Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. *GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.* https://doi.org/10.5281/zenodo.5297715 If you use this software, please cite it using these metadata..

---

[6] https://storage.googleapis.com/wordcraft-stories/index.html

[6] Daniel Buschek, Martin Zurn, and Malin Eiband. 2021. The Impact of Multiple Parallel Phrase Suggestions on Email Input and Composition Behavior of Native and Non-Native English Writers. *ACM* (2021).

[7] Alex Calderwood. 2020. How Novelists Use Generative Language Models: An Exploratory User Study. *Intelligent User Interfaces* (2020).

[8] Jinhan Choi Seonghyeon Kim Sungwoo Lee Changhoon Oh, Jungwoo Song and Bongwon Suh. 2018. I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence. *Computer Human Interaction* (2018).

[9] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).

[10] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All That's 'Human'Is Not Gold: Evaluating Human Evaluation of Generated Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 7282–7296.

[11] Chris CallisonBurch Daphne Ippolito, David Grangier and Douglas Eck. 2019. Unsupervised hierarchical story infilling. *Proceedings of the First Workshop on Narrative Understanding* (2019).

[12] Chenhao Tan Yangfeng Ji Elizabeth Clark, Anne Spencer Ross and Noah A. Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. *Intelligent User Interfaces* (2018).

[13] Katy Gero and Lydia Chilton. 2019. How a Stylistic, Machine-Generated Thesaurus Impacts a Writer's Process. *C&C* (2019).

[14] Katy Gero and Lydia Chilton. 2019. Metaphoria: An Algorithmic Companion for Metaphor Creation. *Computer Human Interaction* (2019).

[15] Cheng-Zhi Huang. 2020. AI Song Contest: Human-AI Co-Creation in Songwriting. *International Society for Music Information Retrieval Conference* (2020).

[16] Lena Hegemann Janin Koch, Andrés Lucero and Antti Oulasvirta. 2019. May AI? Design Ideation with Cooperative Contextual Bandits. *Computer Human Interaction* (2019).

[17] Wolfgang Effelsberg Jonas Freiknecht. 2020. Procedural Generation of Interactive Stories using Language Models. *International Conference on the Foundations of Digital Games* (2020).

[18] Christian Remy Michael Mose Biskjaer Jonas Frich, Lindsay MacDonald Vermeulen and Peter Dalsgaard. 2019. Mapping the Landscape of Creativity Support Tools in HCI. *Computer Human Interaction* (2019).

[19] Varun Kumar Satyapriya Krishna Yada Pruksachatkun Kai-Wei Chang Jwala Dhamala, Tony Sun and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. *arXiv preprint arXiv:2101.11718* (2021).

[20] Anjuli Kannan, Karol Kurach, and Sujith Ravi. 2016. Smart Reply: Automated Response Suggestion for Emil. *International Conference on Knowledge Discovery and Data Mining* (2016).

[21] Kyle McDonell Laria Reynolds. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. *Computer Human Interaction* (2021).

[22] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J Cai. 2020. Novice-AI Music Co-Creation via AI-Steering Tools for Deep Generative Models. *Computer Human Interaction* (2020).

[23] Jay Priyadarshi Marjan Ghazvininejad, Xing Shi and Kevin Knight. 2017. Hafez: an Interactive Poetry Generation System. *Association for Computational Linguistics* (2017).

[24] Jonathan Chen Shao-Yu Chen Shukan Shah-Vishwa Shah Joshua Reno Gillian Smith Matthew Guzdial, Nicholas Liao and Mark O. Riedl. 2019. Friend, Collaborator, Student, Manager: How Design of an AI-Driven Game Level Editor Affects Creators. *Computer Human Interaction* (2019).

[25] Michael Terry-Carrie J. Cai Minhyang (Mia) Suh, Emily Youngblom. 2021. AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition. *Computer Human Interaction* (2021).

[26] Eric Wallace Matthew Jagielski Ariel Herbert-Voss-Katherine Lee Adam Roberts Tom Brown Dawn Song Ulfar Erlingsson et al Nicholas Carlini, Florian Tramer. 2020. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805* (2020).

[27] Kunwar Yashraj Singh Lisa Li Nicholas Davis, Chih-PIn Hsiao and Brian Magerko. 2016. Empirically Studying Participatory Sense-Making in Abstract Drawing with a Co-Creative Cognitive Agent. *Intelligent User Interfaces* (2016).

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

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

[30] Hiroyuki Osone, Jun-Li Lu, and Yoichi Ochiai. 2021. BunCho: AI Supported Story Co-Creation via Unsupervised Multitask Learning to Increase Writers' Creativity in Japanese. *Computer Human Interaction* (2021).

[31] Nicholas Davis Pegah Karimi, Mary Lou Maher and Kazjon Grace. 2019. Deep Learning in a Computational Model for Conceptual Shifts in a Co-Creative Design System. *arXiv preprint arXiv:1906.10188* (2019).

[32] Lara J. Martin Animesh Mehta Brent Harrison Pradyumna Tambwekar, Murtaza Dhuliawala and Mark O. Riedl. 2018. Controllable Neural Story Plot Generation via Reinforcement Learning. *International Joint Conference on Artificial Intelligence* (2018).

[33] Wesley Cheung Zhaochen Luo William Ma Lara J Martin Prithviraj Ammanabrolu, Ethan Tien and Mark O Riedl. 2020. Story realization: Expanding plot events into sentences. *Proceedings of the AAAI Conference on Artificial Intelligence* (2020).

[34] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[35] Nathan Dass Sadao Kurohashi Dan Jurafsky Diyi Yang Reid Pryzant, Richard Diehl Martinez. 2020. Automatically Neutralizing Subjective Bias in Text. *AAAI* (2020).

[36] 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.

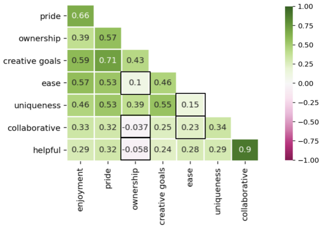
[37] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. LaMDA: Language Models for Dialog Applications. *arXiv preprint arXiv:2201.08239* (2022).

[38] Nick Ryder Melanie Subbiah Jared Kaplan Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray Benjamin Chess Jack Clark Christopher Berner Sam McCandlish Alec Radford Ilya Sutskever Tom B. Brown, Benjamin Mann and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165* (2020).

[39] Shi Feng Dan Klein Sameer Singh Tony Z. Zhao, Eric Wallace. 2021. Calibrate Before Use: Improving Few-Shot Performance of Language Models. *Proceedings of Machine Learning Research* (2021).

[40] Yonghui Wu. 2018. Smart compose: Using neural networks to help write emails. *Google AI Blog* (2018).

A APPENDIX



	enjoyment	pride	ownership	creative goals	ease	uniqueness	collaborative
pride	0.66						
ownership	0.39	0.57					
creative goals	0.59	0.71	0.43				
ease	0.57	0.53	0.1	0.46			
uniqueness	0.46	0.53	0.39	0.55	0.15		
collaborative	0.33	0.32	-0.037	0.25	0.23	0.34	
helpful	0.29	0.32	-0.058	0.24	0.28	0.29	0.9

Figure 9: Pearson correlation coefficients between exit interview questions. Boxes are drawn around uncorrelated questions (p > 0.05).

Table 5: Exit interview results: p values between conditions according to Mann-Whitney U tests of significance. * indicates $p < 0.05$.

question	cont vs. chat	chat vs. full	cont vs full
helpful	0.368	0.0134*	0.0266*
collaborative	0.154	0.00668*	0.0705
ownership	0.420	0.0804	0.131
enjoyment	0.224	0.340	0.132
ease	0.209	0.205	0.495
pride	0.353	0.356	0.495
creative goals	0.397	0.303	0.219
uniqueness	0.0330*	0.279	0.0891

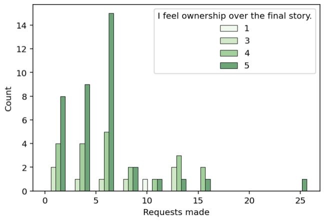


Figure 10: Requests made to the model for users split by their response to the question 'I feel ownership over the final story'.