CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities

Mina Lee
minalee@cs.stanford.edu
Stanford University
United States

Percy Liang
pliang@cs.stanford.edu
Stanford University
United States

Qian Yang
qianyang@cornell.edu
Cornell University
United States

ABSTRACT

Large language models (LMs) offer unprecedented language generation capabilities and exciting opportunities for interaction design. However, their highly context-dependent capabilities are difficult to grasp and are often subjectively interpreted. In this paper, we argue that by *curating and analyzing large interaction datasets*, the HCI community can foster more incisive examinations of LMs' generative capabilities. Exemplifying this approach, we present CoAuthor, a dataset designed for revealing GPT-3's capabilities in assisting creative and argumentative writing. CoAuthor captures rich interactions between 63 writers and four instances of GPT-3 across 1445 writing sessions. We demonstrate that CoAuthor can address questions about GPT-3's language, ideation, and collaboration capabilities, and reveal its contribution as a writing "collaborator" under various definitions of good collaboration. Finally, we discuss how this work may facilitate a more principled discussion around LMs' promises and pitfalls in relation to interaction design. The dataset and an interface for replaying the writing sessions are publicly available at https://coauthor.stanford.edu.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Natural language generation**.

KEYWORDS

Human-AI collaborative writing, GPT-3, language models, dataset, crowdsourcing, natural language generation, writing assistants.

ACM Reference Format:
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 (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3491102.3502030

1 INTRODUCTION

Large language models (LMs) offer exciting opportunities for novel interaction design. Recent LMs (such as GPT-2 [45], GPT-3 [7], GPT-J [57], Jurassic-1 [36], Megatron-Turing-NLG [31], and Gopher [46]) can generate a wide variety of prose and dialogues with an unprecedented level of fluency out of the box. Through fine-tuning, these models can further become specialized at particular tasks, such as composing emails [8] or providing health consultation [58]. As a result, the HCI community has become interested in the opportunities surrounding LMs' generative capabilities. Some have started leveraging off-the-shelf LMs for rapid prototyping of novel natural language interactions [64]; others have started crafting end-user-facing applications with fine-tuned LMs directly[1] [69], even though how soon such applications can become production-ready remain highly disputable [1, 24].

Harnessing LMs' generative capabilities to power interaction designs begins with a *holistic* understanding of these capabilities [5, 68]; this includes understanding what LMs can and cannot do under diverse interaction contexts. For example, when designing the mode of interaction between writers and GPT-3 for writing assistants, designers may ask: Can GPT-3 contribute new ideas to one's writing, or does it merely expand on existing ideas? Does this ideation capability differ in the context of writing fictional stories versus persuasive arguments? To what extent does this capability fluctuate when its decoding parameters change? Answers to such questions guide early interaction design process. Without them, envisioning how an LM may serve writers' needs—or when and how it may fall short—becomes a shot in the dark.

In this paper, we investigate how HCI researchers can examine LMs' generative capabilities to inform interaction design. Any LM's performance fluctuates significantly depending on the preceding text [38], decoding parameters [27], among other factors. Examining such variable capabilities requires more than interviewing its users [66] or tinkering with the model. The challenge becomes even more salient in interactive settings: After a writer and a model takes turns in writing a story and iteratively edits it, how can one tease out and characterize the model's contribution to the writing, or how well it served the writer's needs?

This paper proposes *curating and analyzing large interaction datasets* as one way to address these challenges. Datasets have long been useful in evaluating LMs in Natural Language Processing (NLP) research [25, 32, 47, 56]. We argue that, when thoughtfully designed for HCI, datasets can also reveal what LMs can do for interaction design (Section 3). Exemplifying this approach, we present CoAuthor, a dataset designed for revealing GPT-3's generative

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3502030

[1] More examples of end-user-facing LM applications exist outside academic research, such as GPT-3-powered copywriting tools (e.g. Copy.ai, Copysmith, Omneky, Jarvis, Writesonic), creative writing tools (AI Dungeon, AI Writer, ShortlyAI, Rytr, QuillBot), and programming tools (e.g.TabNine, Copilot).

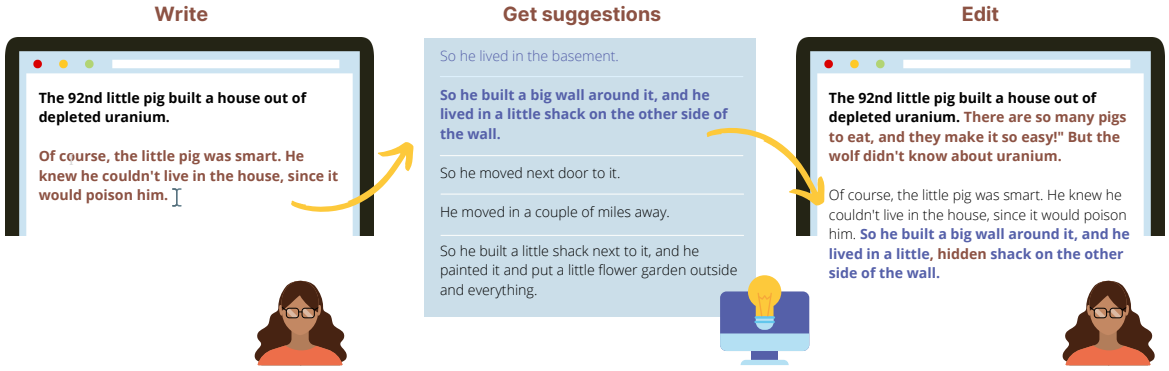


Figure 1: We present CoAuthor, a dataset designed for revealing GPT-3's generative capabilities for interactive writing. It contains rich interactions between 63 writers and 4 instances of GPT-3 across 1445 writing sessions. Each session starts with a prompt (black text). Writers then freely write (brown), request suggestions from GPT-3 (blue), accept or dismiss suggestions, and edit accepted suggestions or previous texts in any order they choose.

capabilities in assisting creative and argumentative writing. It captures rich interactions between 63 writers and four instances of GPT-3 across 1445 writing sessions (Section 4).

We demonstrate that CoAuthor can help to answer high-level questions about GPT-3's generative capabilities. Specifically, we reason about its language capabilities (ability to generate fluent text), ideation capabilities (ability to generate new ideas), and collaboration capabilities (ability to work jointly with writers) using CoAuthor (Section 5.1). The dataset can also help researchers investigate GPT-3's contribution as a writing "collaborator" under various definitions of good collaboration (Section 5.2). We provide a tool for replaying all writing sessions in CoAuthor, giving designers a *felt* understanding of the interactions. The dataset and a replay interface are publicly available at https://coauthor.stanford.edu.

This paper makes three contributions. First, it identifies a need for holistic understanding of LMs' generative capabilities for interaction design. Second, it proposes curating and analyzing large interaction datasets as a viable approach to making LMs' generative capabilities more accessible to the HCI community; this opens up new research opportunities in designing and mining large interaction datasets as a research contribution. Finally, CoAuthor offers a vivid depiction of GPT-3's capabilities in assisting creative and argumentative writing, facilitating a more principled discussion around GPT-3's promises and pitfalls in interaction design.

2 RELATED WORK

2.1 Understanding Technological Capabilities

2.1.1 Types of Understanding. Appropriate interaction design for a new technology requires a deep understanding of its capabilities and limitations. Concretely, this understanding serves two purposes [9]:

- A *specific, felt* capability understanding concerns how the technology can help users in particular contexts, how that interaction may unfold, and what user experiences it may entail. It guides designers in making detailed interaction and user experience (UX) design choices [66].
- A *holistic* capability understanding concerns what the technology is capable and incapable of doing broadly, across various contexts. It gives structure to designers' considerations around how the technology may provide value to different users and what guardrails are necessary to ensure its appropriate use [23, 43].

2.1.2 Ways to Develop Understandings. Researchers have created systems to help designers grasp the capabilities of new or partially-understood technologies. For example, Arduino made accessible the interaction design possibilities of sensors and motors [39]; Wekinator [22] and Teachable Machine [11] made accessible the otherwise abstract capabilities of supervised machine learning classifiers. This approach enables designers to tinker with the technology easily and repeatedly to gain a specific, felt understanding. In addition, designers can use replay enactment [29] to further materialize the dynamics of interactions between users and systems and make complex system behavior more tangible. By "tinkering with a scale" (observing how systems react to different user inputs) and potentially replaying many interactions, designers can develop a more holistic understanding of what the technology can and cannot do broadly [37, 50]. This paper adds to this line of research by curating a large interaction dataset that can be replayed to provide both holistic and felt understandings.

2.2 Understanding Language Models' Generative Capabilities

2.2.1 Language Models' Generative Capabilities. The goal of Natural Language Generation (NLG) is to produce fluent text in many domains, such as machine translation [21], summarization [34], dialogue [30], style transfer [18], and programming code [13]. In recent years, building large LMs has become a common approach to NLG. Unlike traditional models designed to perform a single task (e.g. do translation *or* summarization), recent LMs [7, 31, 36, 46, 57]—the ones that this paper focuses on—learn task-agnostic language representations through pre-training. These LMs can power vastly different tasks (e.g. do translation *and* summarization) [34]. Through additional fine-tuning, pre-trained LMs can further be specialized to given tasks and contexts (e.g. composing emails [8] or providing

health consultation [58]). It would be naïve, however, to think that these LMs have achieved language mastery, or can be trusted with all tasks and contexts. Pre-trained on massive amounts of text on the Internet, LMs are known to produce linguistically flawed, factually incorrect, or even ethically problematic text at times [1, 24]. Guardrails are necessary when using these models.

2.2.2 Challenges in Understanding Generative Capabilities. Understanding LMs' generative capabilities for interactive writing is challenging for at least two reasons:

- LMs' generative capabilities are *highly context-dependent*; therefore, it is difficult to cover all contexts. It may be tempting to describe, for example, GPT-3 as capable of assisting users with writing, coding, and carrying a dialogue. But considering all contexts and characterizing when and how GPT-3 succeed or fail to perform the tasks is effectively intractable. Overestimation of what LMs can thwart interaction design [66].
- LMs' generative capabilities can be *subjectively interpreted*; therefore, they are susceptible to varying evaluations even within a given interaction context. As authors, we can all resonate with how difficult it is to assess how a co-author has helped us with writing a paper. Formally, this assessment requires analysis based on various definitions of good collaboration, at multiple levels of abstraction.[2] Assessing the functional and experiential value of machine-generated text shares similar complexities.

2.2.3 Limitations of Traditional Methods. HCI research has taken two approaches to investigating LMs' generative capabilities for interactive writing. The most traditional approach is contextual inquiry, inviting writers to write with an LM and interviewing them afterward [10, 14, 26, 63, 64, 66]. For example, Calderwood et al. [10] interviewed four professional novelists after they wrote fictional stories with GPT-2. This approach reveals rich insights about how novelists interpreted the capabilities of LMs in specific contexts. However, it is unclear to what extent the findings about GPT-2 would generalize to other writing contexts, to other non-professional writers, to future versions of GPT, or even to other configurations of the same model. Similarly, tinkering with GPT-3 in the Playground (a text box where one can submit a prompt to generate a completion) [42] is unlikely to cover diverse contexts. In other words, contextual inquiry is more effective in capturing the subjective interpretation of LMs' generative capabilities than covering diverse contexts.

An emerging approach to investigating LMs' capabilities is to log interactions and analyze them afterward [8, 48]. For example, Roemmele and Gordon [48] varied the degree of randomness of suggestions generated by a recurrent neural network and tracked writers' edits to suggestions as a strategy for evaluation. Likewise, Buschek et al. [8] logged interactions between native or non-native English speakers and GPT-2 for email writing and analyzed their behavior patterns. Although this approach may provide less rich insights than interviews, it can cover relatively diverse contexts across tasks and writers, while allowing for a fine-grained analysis of interactions. However, most previous work considered restricted interaction settings (e.g. strict turn-taking [15]) and adapted LMs and interfaces (e.g. find-tuned GPT-2 [58]) on specific tasks (e.g. email writing [8]), thereby making it hard to generalize to other tasks and configurations of the same model.

[2]For example, a human writing collaborator can enhance the fluency and the sense of audience in the writing (contribution at a text production level) [6, 53], can expand the pool of knowledge and ideas (ideation and thinking) [19], can better harness the socialization opportunity with the discourse communities (socialization) [65, 70], and more. On an interaction level, good writing collaboration can exhibit different interaction patterns, such as different levels of mutuality and equality [2].

2.3 Datasets in HCI

The challenges of understanding highly context-dependent and subjectively interpreted capabilities are not unique to LMs [68]. In understanding advances in technologies, a different approach has emerged: *developing interaction datasets and providing tools for data analyses* [17, 51, 54, 61, 67]. We distinguish datasets from logs, as logs are usually byproducts of user studies and are not meant to be reused. Nevertheless, this dataset approach shares strengths with the log analysis approach in that it can cover diverse contexts, while supporting the subjective interpretation of LMs' capabilities in a different way.

- Datasets can cover diverse contexts. For example, Theodorou et al. [54] published a dataset of photos taken by blind and low vision users, for assessing how well different object recognition models can serve this user population.
- Datasets can account for the subjective interpretation of capabilities and allow various interpretations. For example, Cuadra et al. [17] provided a video dataset capturing user interactions with a voice assistant without defining which interactions are good. Instead, it opened up discussion around how "good" interactions should be defined, and relatedly, what data-driven interactions are desirable (e.g. What non-verbal cues should a voice assistant detect? How should it respond accordingly?).

The HCI potential of the dataset approach informed this work, with the focus on LMs' generative capabilities for interactive writing.

2.4 Datasets in NLP

Datasets are central to evaluating LMs' generative capabilities in NLP [12, 25, 28, 32, 49, 55]. Typically, a dataset is designed based on a task and collected at scale, containing text across diverse topics from multiple sources and annotators. These datasets are often assembled and combined into a benchmark in order to assess LMs' capabilities comprehensively. For example, Hendrycks et al. [28] proposed a benchmark consisting of 57 datasets including elementary mathematics, US history, computer science, and law, to assess LMs' world knowledge and problem-solving ability. This benchmarking practice has been particularly helpful in assessing recent LMs that can perform many tasks, since datasets are *reusable* when evaluating many LMs, and are easily *expandable* to new tasks and contexts. However, we argue that the underlying assumption of most datasets is the full-automation of tasks rather than augmentation. In other words, they do not consider interactive settings where users can guide and correct systems' generated outputs, but rather expect LMs to generate correct answers alone. As a result, these datasets tend to not capture the *process* of writing, but rather focus on *result*. In this work, we aim to design reusable and expandable datasets that capture the writing process.

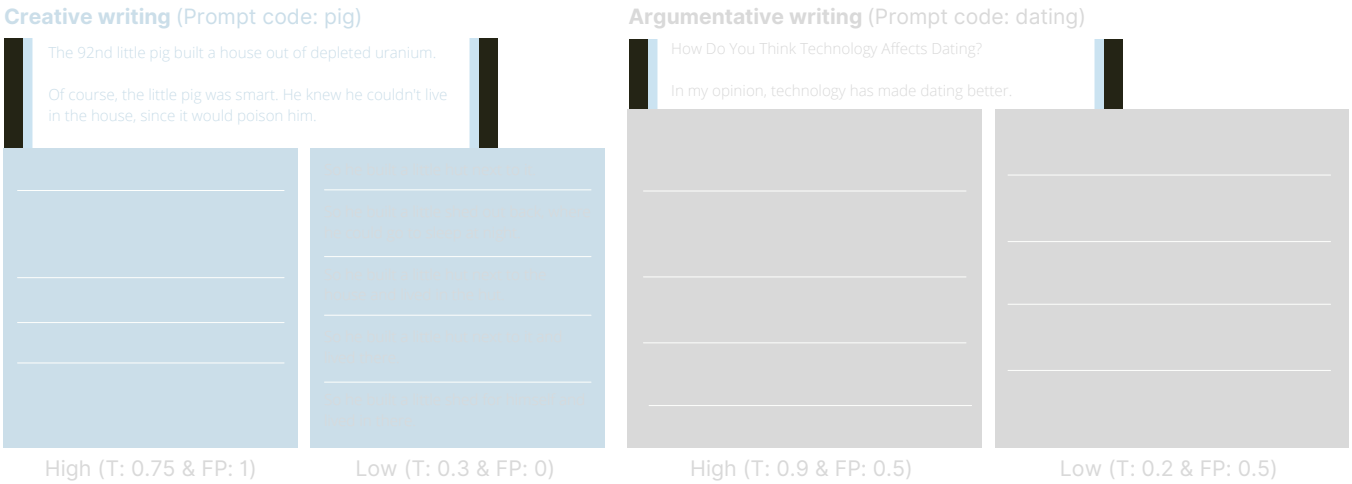


Figure 2: We contrast the capabilities of GPT-3 with high randomness and low randomness in creative and argumentative writing. To control randomness, we varied two decoding parameters: temperature (T) and frequency penalty (FP).

3 DESIGNING DATASETS FOR HCI

We set out to investigate LMs' generative capabilities for interactive writing, making them more accessible for interaction design. Informed by the preceding review of literature, we propose four desiderata for *large interaction datasets* that can capture LMs' generative capabilities:

- **Cover diverse contexts**: Datasets should cover a wide range of contexts such as writing tasks, writing prompts, and writers to account for the highly context-dependent capabilities of LMs.
- **Support subjective interpretations**: Datasets should refrain from imposing a single label or metric. Instead, they should allow designers to extract meaning from interactions and analyze them according to their own design goals.
- **Capture processes, not just results**: Datasets must capture the process of writing that can provide designers a felt understanding of interaction.
- **Allow for possible reuse and expansion**: Datasets need to be reusable and expandable, given the fast advances of LMs and the resource-intensive process of data collection.

4 DESIGNING COAUTHOR

Applying these design principles, we created CoAuthor, a dataset for understanding LMs' generative capabilities for interactive writing. In what follows, we first explain how we designed the dataset based on the four desiderata, and then provide an overview of the dataset. In Section 5, we provide example analyses of the dataset, demonstrating its utility to the HCI community.

4.1 Dataset Design

Cover diverse contexts. CoAuthor contains interactions between writers and LMs across multiple writing tasks, writing prompts, and writers.

- Writing tasks: CoAuthor covers creative writing and argumentative writing, which have distinct goals and require different sets of writing skills. Creative writing, especially story writing, involves structural elements such as character development, narrative, and plot, while infusing the structure with imagination. Argumentative writing requires a writer to investigate a topic by collecting, generating, evaluating evidence, and then establish a position on the topic concisely [33].
- Writing prompts: CoAuthor contains 20 writing *prompts*, brief passages of text that provides a potential topic idea or starting point. We retrieved 10 creative writing prompts from the WritingPrompts subreddit [62], as these prompts have been successful in attracting writers and providing writing inspiration. For argumentative writing, we used prompts from The New York Times [41] in order to provide an accessible, well-balanced set of topics. Appendix B lists the prompts used.
- Writers: We recruited 63 crowd workers (writers) from Amazon Mechanical Turk to account for individuals with potentially different backgrounds and writing styles.

Support subjective interpretations. CoAuthor supports multiple analytical perspectives and goals by providing various measurements in three categories that can be used to define good collaboration: writing outcome, writer perception, and writer behavior.

- Writing outcome: We consider a *writing outcome* to be the artifacts collected at the end of a writing session (e.g. a full list of events and final texts), with its associated measurements, such as time, length, total number of queries, acceptance rate, and written-by-writers rate (i.e. the proportion of the final text written by writers as opposed to the system).
- Writer perception: We asked writers to fill out a survey to understand their *perception* of LM's generative capabilities as well as overall experience (e.g. ownership and satisfaction) after each writing session.
- Writer interaction: We measured how much *interaction* writers had with GPT-3 in the writing process using the notions of equality and mutuality from Storch [52].

Capture processes, not just results. CoAuthor preserves details of rich interactions as *events* and *event blocks*.

Category	Event	Event source	Key binding	Description
System	system-initialize	API	-	Initialize editor
Text (delta)	text-insert	{User, API}	(any key)	Insert text
	text-delete	User	delete	Delete text
Cursor (range)	cursor-forward	{User, API}	{↓, →}	Move cursor forward
	cursor-backward	User	{↑, ←}	Move cursor backward
	cursor-select	User	shift + {↓, →, ↑, ←}	Select range of text
Suggestion	suggestion-get	User	tab	Request new suggestions
	suggestion-open	API	-	Show suggestions
	suggestion-reopen	User	shift + tab	Reopen previous suggestions
	While suggestions are shown			
	suggestion-up	User	↑	Navigate to suggestion above
	suggestion-down	User	↓	Navigate to suggestion below
	suggestion-select	User	enter	Select suggestion
	suggestion-close	{User, API}	esc or (any key)	Hide suggestions

Table 1: List of events. Text events have associated metadata "delta," containing information on inserted or deleted text. Likewise, cursor events have associated metadata "range," containing information on start and end indices of cursor selection.

Category	Event block	Event sequence	Event source
System	init	system-initialize	API
Text	insert	(text-insert)+	User
	delete	(text-delete)+	User
Cursor	cursor	(cursor-forward | cursor-backward | cursor-select)+	User
Suggestion	query	suggestion-get (suggestion-close | cursor-forward)? suggestion-open	User (API)? API
	reopen	suggestion-reopen	User
	navigate	(suggestion-up | suggestion-down)+	User
	choose	suggestion-select suggestion-close text-insert	User API API
	dismiss	suggestion-close	User

Table 2: List of event blocks. Event blocks are deterministic, non-overlapping abstraction of a sequence of events. Event sequence and source are represented using regular expression syntax.

- Events: An *event* can be inserting or deleting text, moving a cursor forward or backward, getting suggestions from the system, or accepting or dismissing suggestions. A list of all events are shown in Table 1 (e.g. text-insert(a)). Formally, an event is a tuple of event name, timestamp, and snapshot of the current editor, which is designed to preserve every detail about interactions.
- Event blocks: Once recorded, events are abstracted into *event blocks*. An event block is a deterministic, non-overlapping abstraction of a sequence of events (e.g. text-insert(a) text-insert(b) → insert(ab)), which is designed to be conducive for further processing and analysis. Table 2 lists all event blocks.

Allow for possible reuse and expansion. We designed CoAuthor to be reusable and easily expandable in the future.

- Writing sessions: CoAuthor consists of a set of *writing sessions*. In each writing session, a writer was presented with a prompt and given an instance of a model with associated decoding parameters. Designers can potentially use a subset of the writing sessions or collect more sessions based on their design goals.
- Language model: To generate suggestions, we used GPT-3 [7] without any adaptation (e.g. fine-tuning). We only varied the randomness of suggestions by changing its decoding parameters (similar to Roemmele and Gordon [48]) as shown in Figure 2. See Appendix A.2 for details.
- System: We used a model- and task-agnostic user interface and interaction design that resembles a text editor (Section 4.2). Future research can easily expand this study and dataset to different writing tasks, prompts, and writers using the same user interface and interaction design.

4.2 Data Collection Interface

Our user interface and interaction design was informed by related designs (e.g. Write With Transformer and Buschek et al. [8]). For alternative interface and interaction designs, we refer readers to Clark et al. [14] and Coenen et al. [16].

4.2.1 Interface. Our interface was a text editor initialized with a writing prompt and keyboard shortcuts for functionality (Figure 3). The text editor was implemented using Quill. The editor supported all typical interactions, such as typing, selecting, editing, and deleting text, and cursor movement via keys and mouse. At the bottom of the screen, a timer for minimum required writing time was shown. After the time was over, the "Get verification code" button was enabled, which writers clicked to finish writing sessions.

4.2.2 Interactions. When writers press the tab key, the system provided five suggestions in a popup box below the cursor (Figure 3). While fetching the suggestions, the icon next to the title (Writing with AI) spun to indicate that the system was getting suggestions. Selecting suggestions in the list was possible via mouse (point and click) or keyboard (arrow up and down to change selection, enter to accept selected suggestion). These main commands were explained as part of the study and shown in the keyboard shortcuts in the interface.

4.2.3 Suggestions. We consider a suggestion "accepted" if a writer selected it from the list. An accepted suggestion is automatically appended at the end of the current text. A group of suggestions is "dismissed" if a writer continues to type, clicks outside of the suggestion popup box, or presses the escape key. The suggestions are automatically hidden as they are dismissed.

4.3 Dataset Curation Procedure

4.3.1 Writers and prompts. We recruited crowd workers (writers) on Amazon Mechanical Turk. From 201 writers who participated in our qualification round, we qualified 100 writers. Among these writers, 63 writers participated in the main round, where 62 writers were native English speakers and one was not. Table 4 and 5 in Appendix B list all prompts in the order they were shown to writers. Writers could write about each prompt up to five times, or choose to skip prompts if they wished not to write. We paid them $2.50 for each writing session. See Appendix D for details.

4.3.2 Instructions. We instructed writers to spend at least ten minutes writing in response to a given prompt. In the instructions, we specified that this was an open-ended task, and the only two requirements were (1) to ensure the story or essay has a clear ending or a clear stance and conclusion, and (2) to collaborate with the system to write a story or an essay for ten minutes. We included self-evaluation of two requirements after each writing session in order to remind writers of the requirements and control the quality of final texts. See Figure 11 in Appendix E for full instructions.

4.3.3 Survey. We asked writers to fill out a survey after each writing session to contrast the capabilities of LMs across different prompts and qualities associated with the system. The survey questions consisted of five sections: writer information (native vs. non-native English speaker), assessment of the known benefits of collaborative writing (fluency, pooling of ideas, and enhanced quality), perceived capabilities of LMs, perceived limitations of LMs, and overall experiences (ownership, satisfaction, and willingness to reuse). See Appendix C for the full list of questions.

Figure 3: Interface used for data collection of CoAuthor. Our interface was a text editor in which writers press the tab key to get suggestions from the system whenever desired.

4.4 Dataset Overview

Table 3 shows overall statistics about CoAuthor. The dataset contains 830 stories written by 58 writers for creative writing and 615 essays written by 49 writers for argumentative writing. On average, each writing session results in 418 words of text, contains 11.8 queries to the system, has acceptance rate of 72.3%, and 72.6% of text is written by writers.

5 DEMONSTRATING USES OF COAUTHOR

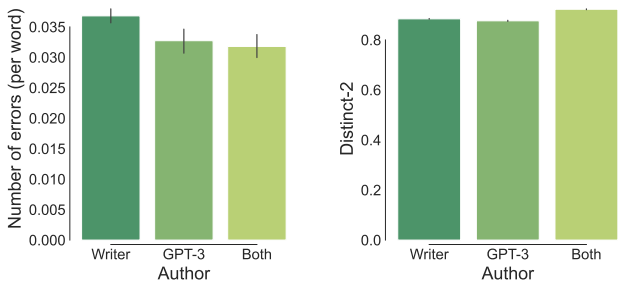
This section demonstrates the usefulness of CoAuthor for revealing GPT-3's generative capabilities in assisting creative and argumentative writing. Specifically, it allows designers to explore the generative capabilities of LMs holistically and to reason about its contribution as a writing "collaborator" under various definitions of good collaboration.

5.1 GPT-3's Generative Capabilities

To gain a holistic understanding of GPT-3's generative capabilities for interactive writing, we looked at three aspects of capabilities using CoAuthor and compared our findings to existing hypotheses. First, we study *language* capabilities (ability to generate fluent text). In Section 5.1.1, we show that the sentences generated by GPT-3 had less spelling and grammar errors than the sentences written by writers in CoAuthor. Second, we focus on *ideation* capabilities (ability to generate new ideas). In Section 5.1.2, we present evidence that GPT-3 is capable of providing new ideas to writers, influencing their subsequent writing. Lastly, we investigate *collaboration* capabilities (ability to work jointly with writers). In Section 5.1.3, we demonstrate that the amount of collaboration between writers

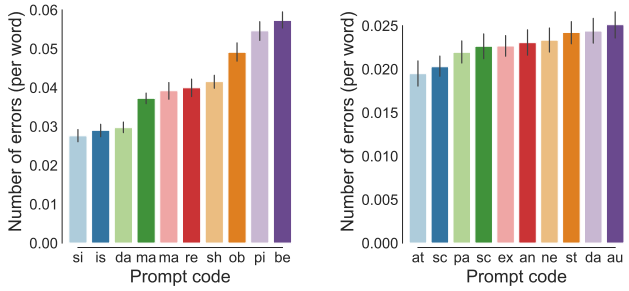
	Overall			Writing sessions				
	Prompts	Writers	Sessions	Time (minutes)	Length (words)	Queries	Acceptance rate (%)	Written by writers (%)
Creative	10	58	830	11.6	446	12.8	75.7	72.7
Argumentative	10	49	615	10.6	380	10.3	67.6	72.5
Combined	20	63	1445	11.2	366	11.8	72.3	72.6

Table 3: Overall statistics of CoAuthor. The dataset consists of a set of writing sessions in the two types of writing: Creative and argumentative writing. For each type, ten writing prompts were provided, from which writers could write up to five times per prompt. The dataset contains 1445 writing sessions written by 63 writers. On average, each writing session is 418 words long, contains 11.8 queries to the system, has acceptance rate of 72.3% (how often writers accepted suggestions from GPT-3), and results in 72.6% of final texts written by writers (as opposed to GPT-3).



(a) Grammaticality (b) Vocabulary diversity

Figure 4: Sentences written by both writers and GPT-3 had fewer spelling and grammatical errors (a) and contained more diverse vocabulary (b) compared to sentences written by writers alone and GPT-3 alone.



(a) Creative writing (b) Argumentative writing

Figure 5: For both creative (a) and argumentative (b) writing, the number of spelling and grammar errors per word (y-axis) in GPT-3-generated sentences vary across writing prompts (x-axis).

and GPT-3 varied significantly across writers, but less depended on writing prompts and the randomness of suggestions.

5.1.1 Language Capabilities: Ability to Generate Fluent Text. To study language capabilities of GPT-3, we compared the grammaticality and vocabulary diversity of sentences written by writers alone, GPT-3 alone, and writers and GPT-3 together in CoAuthor. For simplicity, we considered the sentences from final texts, as opposed to sentences during the writing process that might have been edited or deleted later on and do not appear in final texts.

How grammatical is GPT-3's writing? Figure 4 (a) shows that sentences written by writers had more spelling and grammar errors compared to GPT-3-generated sentences in final texts. We measured *grammaticality* by averaging over the number of spelling and grammar errors across sentences using LanguageTool. Overall, the number of errors per word (averaged across all sentences in creative and argumentative writing) was 0.037 ± 0.001 for writers, 0.033 ± 0.001 for GPT-3, and 0.032 ± 0.001 for both (the number next to the average indicates the standard error of measurement). This matches with Dou et al. [20]'s finding that sentences written by writers tend to have more grammar and usage errors compared to those written by GPT-3.

How diverse is GPT-3's vocabulary? Figure 4 (b) shows that sentences written by both writers and GPT-3 contained more diverse vocabulary compared to the sentences written by writers alone and GPT-3 alone. The vocabulary diversity was measured by counting the number of unique bigrams scaled by total number of generated words (distinct-2) [35]. Overall, the distinct-2 score (averaged across all sentences in creative and argumentative writing) was 0.884 ± 0.001 for writers, 0.876 ± 0.001 for GPT-3, and 0.923 ± 0.001 for both. The result may imply that the use of suggestions from GPT-3 encouraged writers to use more diverse vocabulary.

This matches with the previous findings about machine-generated text being less diverse than human-authored counterparts [27]. Note that the sentences in final texts only include suggestions from GPT-3 which were accepted by writers. In other words, dismissed suggestions generated by GPT-3 (which may contain much less diverse vocabulary) were not included in the previous analysis. When we computed the distinct-2 score on *all* suggestions generated by GPT-3, the score was even lower (0.868 ± 0.001), further confirming the previous findings.

Do prompts influence GPT-3's grammaticality? The grammaticality of GPT-3-generated text varies across prompts (Figure 5). This observation matches with recent findings that LMs have high

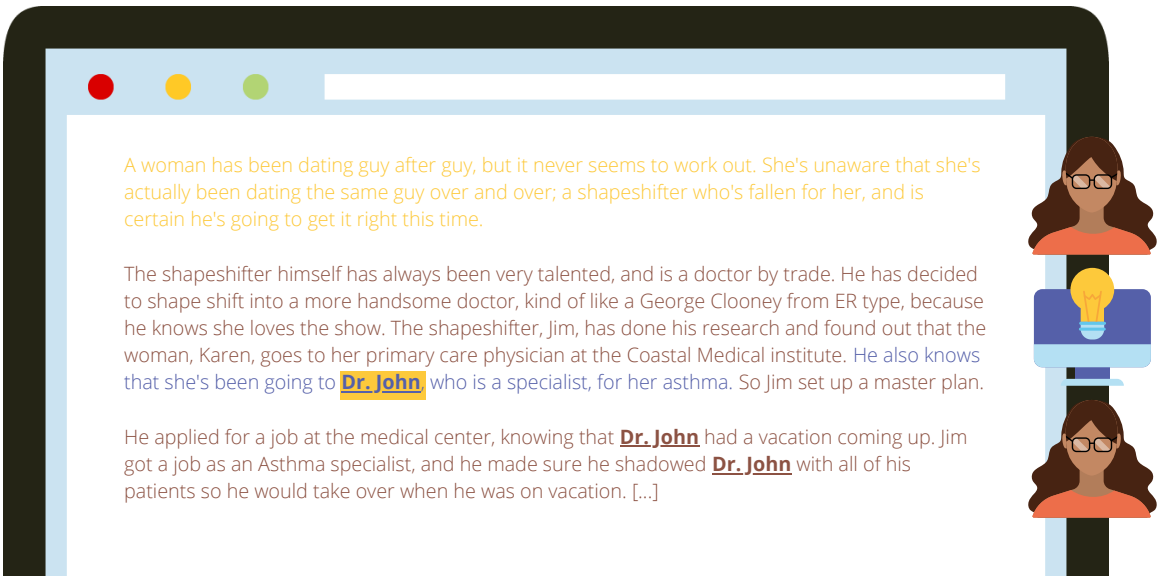


Figure 6: Example of a story in which a writer accepted a suggestion from GPT-3 with a new named entity "Dr. John" and used the entity in the subsequent writing. The prompt is shown in black, sentences written by the writer in brown, and sentences written by GPT-3 in blue.

variance performance based on the choice of prompts [38]. In survey responses, we find varying degree of writers' reactions to the suggestions they received. Positive responses include "*Didn't notice any grammar errors or anything*" and "*I am not a woman, or who reads Romance novels, but the AI apparently is! It did a grand job of writing this time; beside grammar and formatting, there is nothing I saw that could be improved.*" Negative ones include "*Some of the AI suggestions had some typos, so they were a little grammatically incorrect. For example, the AI suggested a sentence with the word "imminent" being spelled "emminent," which is not correct. I used the sentence, but had to correct the misspelling. So the grammar could be improved slightly.*"

5.1.2 Ideation Capabilities: Ability to Generate New Ideas. Some existing work uses the amount of text written by the system or the number of edits made by writers to estimate the system's "usefulness" or degree of contribution [15, 48]. However, this notion of usefulness is limited, as writers may find inspiration in suggestions (or portions thereof) or even dismissed suggestions. To find evidence that LMs can provide new ideas, we identified GPT-3-generated sentences with new *named entities* (a real-world object, such as a person and location, that can be denoted with a proper name) that did not appear in their previous contexts. Then, we checked whether they were reused by writers in subsequent writing.

Figure 6 shows an example story from CoAuthor where a new named entity was introduced by GPT-3 and subsequently reused by a writer. We used Stanza [44] to identify named entities and exact match to check whether they were reused in subsequent writing. Note that this is likely to underestimate GPT-3's contributions, since new ideas are not always expressed as named entities and names may appear in a different form in the later text (e.g. pronouns).

How often do suggestions contain new named entities? 13% and 7% of accepted suggestions contained new named entities in creative and argumentative writing, respectively (Figure 7). We suspect that this difference could be due to the different writing types (e.g. stories are more likely to need new names and locations) and the different sets of decoding parameters (temperature and frequency penalty as shown in Figure 2).

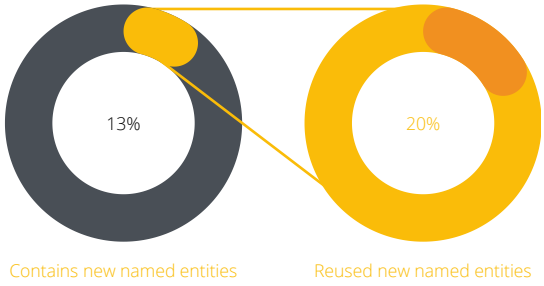


Figure 7: In creative writing, 13% of suggestions (generated by GPT-3) contained new named entities, among which 20% were reused by writers in subsequent writing.

How often are the new named entities reused by writers? Among the new named entities proposed by GPT-3, 20% and 14% of them were reused by writers in subsequent writing in creative and argumentative writing, respectively (Figure 7). Survey responses show that some writers liked and found the suggestions with new named entities helpful (e.g. "*I especially found the names helpful. I was actually trying to think of a stereotypical rich jock name, and the AI provided me with Chadwick. Perfect!*", "*I found the suggestions that introduced new characters to be most helpful. They really helped push the story along nicely.*").

Does the randomness of GPT-3 influence its likelihood of generating new named entities? The GPT-3 instance with high randomness was more likely to generate sentences with new named entities compared to the one with low randomness, while the likelihoods of the entities being reused by writers were similar. In creative writing, accepted suggestions from GPT-3 with high randomness contained new named entities more often (15 ± 1%), compared to the ones from GPT-3 with low randomness (10 ± 0%). However, the likelihoods of new named entities being reused by writers were similar for both high randomness (19 ± 2%) and low randomness (22 ± 2%). Similarly, in argumentative writing, accepted suggestions from GPT-3 with high randomness contained new named entities more often (9 ± 1%), compared to the ones from GPT-3 with low randomness (6 ± 0%). Yet, the likelihoods of new named entities being reused by writers were similar for both high randomness (13 ± 2%) and low randomness (15 ± 3%).

5.1.3 Collaboration Capabilities: Ability to Work Jointly. To investigate the extent that writers interact with GPT-3 in the writing process, we adapted the notions of equality and mutuality from Storch [52]. We redefined *equality* as how evenly a writer and GPT-3 distributed turns (e.g. the degree of deviation from even division of work) and *mutuality* as the level of interaction a writer has with GPT-3 (e.g. querying multiple times, choosing suggestions, reopening suggestions, and navigating through suggestions). Concretely, we computed equality and mutuality scores for a writing session by counting the number of event blocks as follows. Let

$$\mathcal{H} = \{\text{insert}\},\ \mathcal{M} = \{\text{choose}\}$$

be the sets representing human-generated event blocks and machine-generated event blocks, respectively. Also, let

$$\mathcal{I} = \{\text{insert, choose, reopen, navigate}\},$$
$$\mathcal{A} = \{\text{dismiss, insert, delete}\}$$

denote the sets of event blocks corresponding to human-machine interactions and event blocks corresponding to writing alone, respectively. Given a set of events $\{e_i\}$, we define:

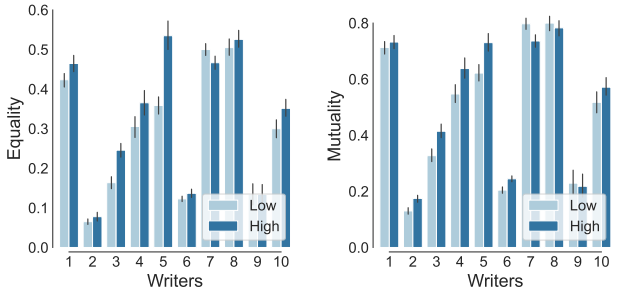
$$\text{equality} = 1 - \frac{\sum_i [e_i \in \mathcal{H}] - \sum_i [e_i \in \mathcal{M}]}{\sum_i [e_i \in \mathcal{H}] + \sum_i [e_i \in \mathcal{M}]},$$

where $[P] = 1$ if P is true and 0 if not. Moreover, define:

$$\text{mutuality} = \frac{\sum_i [e_i \in \mathcal{I}]}{\sum_i [e_i \in \mathcal{I}] + \sum_i [e_i \in \mathcal{A}]}.$$

An equality score of 1 means perfect division of turns for writing and 0 means that there was no turn change (i.e. it was written entirely by a writer or GPT-3). For mutuality, score 1 means writers interacted with GPT-3 the entire time (without writing on their own as a result) and 0 means writers never interacted with GPT-3. Note that these scores are not based on final texts, but event blocks. This accounts for text written (by either writers or GPT-3) during the writing session, that may not be present in final texts.

Do writers and prompts influence collaboration? Both collaboration equality and mutuality varied greatly depending on writers, but less on prompts. Figure 8 shows varying degrees of equality and mutuality scores across writers: for instance, Writer #1 scored around 0.45 and 0.75 for equality and mutuality, whereas Writer #2 scored below 0.1 and 0.2 (the scores are averaged over all writing sessions). On the other hand, both equality and mutuality scores did not fluctuate as much across writing prompts, as shown in Figure 9. This result may indicate that differences between writers are more likely to influence the degree of collaboration rather than those of writing prompts (see Table 4 and 5 in Appendix B for the full list of prompts).

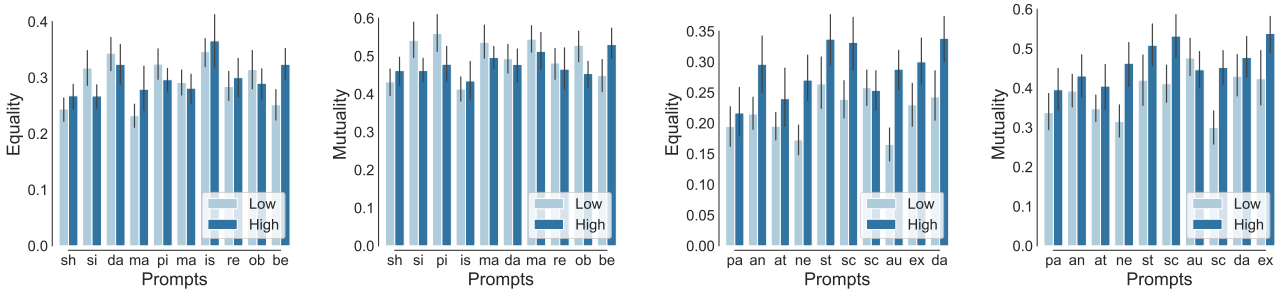


(a) Equality **(b) Mutuality**

Figure 8: Equality (a) and mutuality (b) (y-axis) vary across writers (x-axis) greatly. Also, some writers had more equal and mutual collaboration with GPT-3 with high randomness (dark blue), whereas others had such collaboration with GPT-3 with low randomness (light blue).

Does the randomness of GPT-3 influence collaboration? Writing sessions with GPT-3 with *high* randomness received slightly higher equality and mutuality scores in argumentative writing, while the difference was not statistically significant in creative writing. In argumentative writing, the sessions with GPT-3 with high randomness resulted in equality of 0.29 ± 0.01 and mutuality of 0.47 ± 0.01, whereas GPT-3 with low randomness resulted in 0.22 ± 0.01 and 0.38 ± 0.01 on average (Figure 9 (b)). When aggregated by prompts, 9 out of 10 prompts (90%) had higher equality and mutuality scores with GPT-3 with *high* randomness. On the other hand, in creative writing, writing sessions with GPT-3 with high randomness resulted in 0.30 ± 0.01 for equality and 0.48 ± 0.01 for mutuality, whereas the sessions with GPT-3 with low randomness resulted in 0.29 ± 0.01 and 0.49 ± 0.01, respectively (Figure 9 (a)). When aggregated by prompts, 5 out of 10 prompts (50%) and 8 out of 10 prompts (80%) had higher equality and mutuality scores with GPT-3 with *low* randomness.

On a granular level, writers have different preferences over suggestions generated by GPT-3 with high randomness and low randomness. Figure 8 shows that some writers (e.g. Writer #3 and #5) had more equal and mutual collaboration when they interacted with GPT-3 with high randomness, whereas others (e.g. Writer #7) had this type of collaboration with GPT-3 with low randomness. This could be due to writers having different purposes of collaborating with the system that made one more preferable than the other in two writing types. For example, in creative writing, some writers used the system to advance plots (62.9%), whereas some used it to add details to stories (51.4%) according to the survey's multiple choice question. In this example, the former group of writers might have preferred suggestions from GPT-3 with high randomness, as they tend to be more diverse. On the other hand, the latter might



(a) Creative writing **(b) Argumentative writing**

Figure 9: Equality and mutuality (y-axis) and vary across prompts (x-axis) in creative (a) and argumentative writing (b), but to the less extent than across writers.

have preferred suggestions from GPT-3 with low randomness, as they tend to be more grammatical and coherent [48].

5.2 Various Definitions of Good Collaboration

In addition to addressing broad questions about GPT-3's generative capabilities, CoAuthor can also help researchers reason about GPT-3's contribution as a writing "collaborator" under various definitions of good collaboration. Here, we consider two examples definitions (productivity and ownership) and show how CoAuthor may provide preliminary evidence for formulating hypotheses regarding interaction design. Note that we use simplified definitions and generate hypotheses based on correlations. To validate the hypotheses, further experiments with interventions are necessary.

5.2.1 Increasing writers' productivity. Consider a design scenario where a designer considers how to design a GPT-3-powered autocomplete system to increase writers' productivity. CoAuthor reveals a number of factors that are positively correlated with the amount of text the writers and GPT-3 end up producing. These factors include the time writers spent writing, the number of queries, and the number of accepted suggestions. The correlation was much stronger for the number of queries and accepted suggestions (0.42 and 0.48) compared to the time writers spent writing (0.29). This observation suggests that having suggestions from GPT-3 has the potential to increase writers' productivity. The designer can further investigate individual instances where this correlation was strong or weak and make more informed decisions, before deploying GPT-3.

5.2.2 Increasing writers' feeling of ownership. Consider another scenario where a designer wants to increase writers' feeling of ownership over their GPT-3-assisted writing. In this case, the designer could consider devising ways to keep the fraction of text written by writers to text written by GPT-3 relatively high. We observe correlation between the ownership writers have over final texts and the fraction of text written by writers. For ownership scores (rated as a 5-point Likert scale) and the fraction of text written by writers, the Pearson correlation coefficient was 0.3 in both creative and argumentative writing, whereas it was 0.1 and 0.0 for satisfaction scores. This result may imply that the more writers get suggestions from GPT-3, the less they write and the less they feel ownership over final texts.

On the other hand, encouraging writers to make more edits (on their own writing and GPT-3-generated sentences) may not be as effective in increasing writers' feeling of ownership. In human-human collaborative writing, Birnholtz et al. [3] suggested that the quantity of collaboration (e.g. number of comments and edits) may affect writers' perceived ownership of final texts and attractiveness of the group task. However, we did not observe meaningful correlation between the amount of edits by writers and their ownership or satisfaction scores in CoAuthor. We approximated the number of edits by counting the number of delete and cursor event blocks in each writing session. For ownership score, the Pearson correlation coefficient was 0.1 in both creative and argumentative writing. For satisfaction score, the coefficient was 0.0 and −0.1 in creative and argumentative writing.

6 DISCUSSION

Excitement about LMs' promises over their perils are often rooted in observations of their particular behaviors in very restricted interaction settings. These observations are rarely cross-referenced with the literature in NLP that attempts to examine LMs' generative capabilities holistically. In this section, we aim to bring together these discussions, which often occur in isolation. We first discuss the role of datasets as boundary objects between the HCI and NLP communities. Then, we describe potential use cases of CoAuthor in HCI and NLP research, exemplifying the potential of CoAuthor serving as a boundary object.

6.1 Datasets as Boundary Objects

We argue that datasets have the potential to serve as boundary objects between the HCI and NLP communities. For example, datasets in NLP can provide resources to help HCI researchers reason about LMs' generative capabilities in a holistic manner. HCI researchers can provide analytical tools for NLP researchers to better investigate LMs's capabilities interactive settings. Moreover, large interaction datasets in HCI can embed human-centered values and practices, and further incorporate them into technical advances, when used to train or evaluate LMs. By having datasets as boundary objects,

the HCI and NLP communities can easily communicate results and have shared understanding of LMs' generative capabilities.

6.2 Potential Use Cases of CoAuthor

6.2.1 Formulate Hypotheses. HCI researchers can use CoAuthor to formulate hypotheses via replay enactment [29]. Concretely, researchers can replay writing sessions in CoAuthor, which materializes the dynamics of interactions between writers and GPT-3 and makes complex system behavior more tangible. Through replay, researchers may discover, for example, that some writers prefer to get suggestions from GPT-3 in the beginning, whereas others prefer to do so throughout the writing process. Researchers with an eye towards building personalized writing assistants may notice similarities and differences across writers, or specific needs for certain writers.

6.2.2 Assess the Plausibility of Hypotheses. CoAuthor can serve as a starting point to assess the plausibility of hypotheses. For instance, to investigate how the style, voice, or tone of a writer or GPT-3 influences that of the other over time (i.e. linguistic accommodation), researchers can first check the existence such phenomena in CoAuthor as supporting evidence to warrant further investigation (e.g. recruiting a specific set of participants for initial interview). Similarly, if the above hypothesis is true, researchers can examine whether the influence is uni-directional (e.g. from writer to GPT-3) or bi-directional (e.g. both from writer to GPT-3 and from GPT-3 to writer) by analyzing events and associated timestamps within each writing session. How users adapt to systems over time is of interest for not only HCI researchers but also NLP researchers [59, 60].

6.2.3 Train and Evaluate Language Models. NLP researchers can train and evaluate LMs on CoAuthor to better support interactive writing. One can expect that a LM fine-tuned on *accepted* suggestions may generate suggestions that are more desirable by writers, compared to the same LM without any fine-tuning. Most current LMs are trained and evaluated on datasets that do not consider interactive settings. On the other hand, our dataset considers an interactive setting, which seamlessly embodies in-situ generation and evaluation. Additional information in the dataset (e.g. when writers asked for suggestions, which suggestions they accepted, dismissed, and modified, and how they perceived the capabilities of GPT-3) can provide supervision signals for LMs to learn desirable behaviors.

7 CONCLUSION

In this work, we identified a critical need for understanding LMs' generative capabilities for interaction design. We argued that *curating and analyzing large interaction datasets* is one viable approach, as it can cover diverse interaction contexts and support various interpretations of good collaboration. Exemplifying this approach, we created CoAuthor, a dataset containing rich interactions between 63 writers and four instances of GPT-3 across 1445 writing sessions. We demonstrated the feasibility of this approach and discussed insights designers can draw from the dataset. We encourage fellow researchers to use, analyze, and extend CoAuthor, based on their respective design goals and research perspectives.

REFERENCES

[1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. arXiv:2101.05783 [cs.CL]

[2] Zsuzsanna I Abrams. 2019. Collaborative Writing and Text Quality in Google Docs. *Language Learning & Technology* 23, 2 (2019), 22–42.

[3] Jeremy Birnholtz, Stephanie Steinhardt, and Antonella Pavese. 2013. Write Here, Write Now! An Experimental Study of Group Maintenance in Collaborative Writing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 961–970.

[4] Joel Bloch. 2007. Abdullah's blogging: A generation 1.5 student enters the blogosphere. *Language Learning & Technology* 11, 2 (2007), 128–141.

[5] Sara Bly and Elizabeth F Churchill. 1999. Design through matchmaking: technology in search of users. *interactions* 6, 2 (1999), 23–31.

[6] Joel Boch. 2007. Abdullah's blogging: A generation 1.5 student enters the blogosphere. *Language Learning & Technology* 11, 2 (2007), 128–141.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901.

[8] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. *The Impact of Multiple Parallel Phrase Suggestions on Email Input and Composition Behaviour of Native and Non-Native English Writers.* Association for Computing Machinery, New York, NY, USA.

[9] Bill Buxton. 2010. *Sketching user experiences: getting the design right and the right design.* Morgan Kaufmann.

[10] 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 the Workshops on Human-AI Co-Creation with Generative Models and User-Aware Conversational Agents co-located with 25th International Conference on Intelligent User Interfaces.* Association for Computing Machinery.

[11] Michelle Carney, Barron Webster, Irene Alvarado, Kyle Phillips, Noura Howell, Jordan Griffith, Jonas Jongejan, Amit Pitaru, and Alexander Chen. 2020. Teachable Machine: Approachable Web-Based Tool for Exploring Machine Learning Classification. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8.

[12] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of Text Generation: A Survey.

[13] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating Large Language Models Trained on Code. arXiv:2107.03374 [cs.LG]

[14] 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* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 329–340.

[15] Elizabeth Clark and Noah A. Smith. 2021. Choose Your Own Adventure: Paired Suggestions in Collaborative Writing for Evaluating Story Generation Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics, Online, 3566–3575.

[16] Andy Coenen, Luke Davis, Daphne Ippolito, Emily Reif, and Ann Yuan. 2021. Wordcraft: a Human-AI Collaborative Editor for Story Writing. arXiv:2107.07430 [cs.CL]

[17] Andrea Cuadra, Hansol Lee, Jason Cho, and Wendy Ju. 2021. Look at Me When I Talk to You: A Video Dataset to Enable Voice Assistants to Recognize Errors. *arXiv preprint arXiv:2104.07153* (2021).

[18] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.*

OpenReview.net.
[19] Richard Donato. 1994. Collective scaffolding in a second language. *Vygotskian approaches to second language research* (1994), 33–56.
[20] Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A Smith, and Yejin Choi. 2021. Scarecrow: A Framework for Scrutinizing Machine Text. *arXiv preprint arXiv:2107.01294* (2021).
[21] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond English-Centric Multilingual Machine Translation. arXiv:2010.11125 [cs.CL]
[22] Rebecca Fiebrink and Perry R Cook. 2010. The Wekinator: a system for real-time, interactive machine learning in music. In *Proceedings of The Eleventh International Society for Music Information Retrieval Conference (ISMIR 2010)(Utrecht)*.
[23] William W. Gaver. 1991. Technology Affordances. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New Orleans, Louisiana, USA) *(CHI '91)*. ACM, New York, NY, USA, 79–84. https://doi.org/10.1145/108844.108856
[24] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 3356–3369.
[25] Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*. Association for Computational Linguistics, Online, 96–120.
[26] Katy Ilonka Gero and Lydia B Chilton. 2019. Metaphoria: An algorithmic companion for metaphor creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
[27] Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying Human and Statistical Evaluation for Natural Language Generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 1689–1701.
[28] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
[29] Kenneth Holstein, Erik Harpstead, Rebecca Gulotta, and Jodi Forlizzi. 2020. Replay Enactments: Exploring Possible Futures through Historical Data. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS '20)*. Association for Computing Machinery, New York, NY, USA, 1607–1618. https://doi.org/10.1145/3357236.3395427
[30] Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A Simple Language Model for Task-Oriented Dialogue. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 20179–20191.
[31] Paresh Kharya and Ali Alvi. 2021. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model. https://developer.nvidia.com/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/
[32] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking Benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 4110–4124.
[33] Purdue Online Writing Lab. 2021. Argumentative Essays. https://owl.purdue.edu/owl/general_writing/academic_writing/essay_writing/argumentative_essays.html
[34] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7871–7880.
[35] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, 110–119.
[36] Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. 2021. Jurassic-1: Technical Details and Evaluation.
[37] Youn-Kyung Lim, Erik Stolterman, and Josh Tenenberg. 2008. The anatomy of prototypes: Prototypes as filters, prototypes as manifestations of design ideas. *ACM Transactions on Computer-Human Interaction (TOCHI)* 15, 2 (2008), 7.
[38] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. arXiv:2107.13586 [cs.CL]
[39] Nikolas Martelaro, J.D. Zamfirescu-Pereria, David Goedicke, David Sirkin, and Wendy Ju. 2020. Make This! Introduction to Electronics Prototyping Using Arduino. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–4.
[40] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.
[41] The Learning Network. 2021. 300 Questions and Images to Inspire Argument Writing. https://www.nytimes.com/2021/02/01/learning/300-questions-and-images-to-inspire-argument-writing.html
[42] OpenAI. 2020. OpenAI GPT-3 playground: GPT-3 demo. https://gpt3demo.com/apps/openai-gpt-3-playground
[43] Fatih Kursat Ozenc, Miso Kim, John Zimmerman, Stephen Oney, and Brad Myers. 2010. How to support designers in getting hold of the immaterial material of software. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2513–2522.
[44] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Online, 101–108.
[45] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* 1, 8 (2019).
[46] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, 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 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 Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed 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. arXiv:2112.11446 [cs.CL]
[47] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, 784–789.
[48] Melissa Roemmele and Andrew S. Gordon. 2018. Automated Assistance for Creative Writing with an RNN Language Model. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion* (Tokyo, Japan) *(IUI '18 Companion)*. Association for Computing Machinery, New York, NY, USA, Article 21, 2 pages.
[49] Sebastian Ruder. 2021. Challenges and Opportunities in NLP Benchmarking. http://ruder.io/nlp-benchmarking
[50] Donald A Schön. 1984. *The reflective practitioner: How professionals think in action*. Vol. 5126. Basic books.
[51] Rob Semmens, Nikolas Martelaro, Pushyami Kaveti, Simon Stent, and Wendy Ju. 2019. *Is Now A Good Time? An Empirical Study of Vehicle-Driver Communication*

Timing. Association for Computing Machinery, New York, NY, USA, 1–12.
[52] Neomy Storch. 2002. Patterns of interaction in ESL pair work. *Language learning* 52, 1 (2002), 119–158.
[53] Neomy Storch. 2005. Collaborative writing: Product, process, and students' reflections. *Journal of second language writing* 14, 3 (2005), 153–173.
[54] Lida Theodorou, Daniela Massiceti, Luisa Zintgraf, Simone Stumpf, Cecily Morrison, Edward Cutrell, Matthew Tobias Harris, and Katja Hofmann. 2021. *Disability-First Dataset Creation: Lessons from Constructing a Dataset for Teachable Object Recognition with Blind and Low Vision Data Collectors*. Association for Computing Machinery, New York, NY, USA.
[55] Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*. Association for Computational Linguistics, Tokyo, Japan, 355–368.
[56] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Brussels, Belgium, 353–355.
[57] Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.
[58] Lu Wang, Munif Ishad Mujib, Jake Williams, George Demiris, and Jina Huh-Yoo. 2021. An Evaluation of Generative Pre-Training Model-based Therapy Chatbot for Caregivers. arXiv:2107.13115 [cs.HC]
[59] Sida I. Wang, Samuel Ginn, Percy Liang, and Christopher D. Manning. 2017. Naturalizing a Programming Language via Interactive Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 929–938.
[60] Sida I. Wang, Percy Liang, and Christopher D. Manning. 2016. Learning Language Games through Interaction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 2368–2378.
[61] Sarah Wiegreffe and Ana Marasović. 2021. Teach Me to Explain: A Review of Datasets for Explainable NLP. arXiv:2102.12060 [cs.CL]
[62] WritingPrompts. 2021. https://www.reddit.com/r/WritingPrompts
[63] Qingyang Wu, Lei Li, Hao Zhou, Ying Zeng, and Zhou Yu. 2020. Importance-aware learning for neural headline editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9282–9289.
[64] Tongshuang Wu, Michael Terry, and Carrie J. Cai. 2021. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. *CoRR* abs/2110.01691 (2021). arXiv:2110.01691
[65] Luxin Yang. 2014. Examining the mediational means in collaborative writing: Case studies of undergraduate ESL students in business courses. *Journal of Second Language Writing* 23 (2014), 74–89.
[66] Qian Yang, Justin Cranshaw, Saleema Amershi, Shamsi T. Iqbal, and Jaime Teevan. 2019. Sketching NLP: A Case Study of Exploring the Right Things To Design with Language Intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 185, 12 pages. https://doi.org/10.1145/3290605.3300415
[67] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating How Experienced UX Designers Effectively Work with Machine Learning. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. ACM, New York, NY, USA, 585–596. https://doi.org/10.1145/3196709.3196730
[68] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376301
[69] Wenmian Yang, Guangtao Zeng, Bowen Tan, Zeqian Ju, Subrato Chakravorty, Xuehai He, Shu Chen, Xingyi Yang, Qingyang Wu, Zhou Yu, Eric P. Xing, and Pengtao Xie. 2020. On the Generation of Medical Dialogues for COVID-19. *CoRR* abs/2005.05442 (2020). arXiv:2005.05442
[70] Soobin Yim and Mark Warschauer. 2017. Web-based collaborative writing in L2 contexts: Methodological insights from text mining. *Language Learning & Technology* 21, 1 (2017), 146–165.

A SYSTEM SETTINGS

A.1 Server

Our server is written in Python using Flask, and is used by the frontend both to request suggestions and to save events in a writing session. We deployed the system on our institution's infrastructure.

A.2 Decoding parameters

We used the following decoding parameters for GPT-3 [7] to generate suggestions.

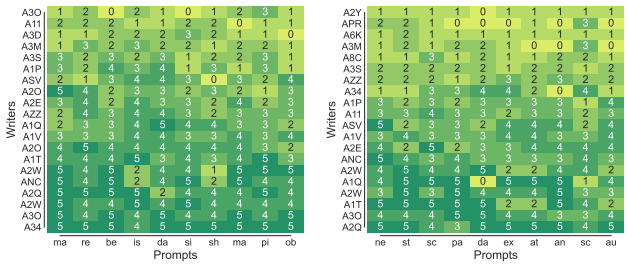
- Engine: davinci
- Response length (word piece): 30
- Temperature: 0.2, 0.3, 0.75, 0.9
- Top P: 1
- Frequency penalty: 0, 0.5, 1
- Presence penalty: 0
- Best of: 1

A.3 Suggestions

Once GPT-3 generated an output given a prompt, it was parsed as a list of sentences using Stanza [44], and only the first sentence was used as a suggestion for readability. In addition, despite querying GPT-3 five times, we oftentimes ended up with a fewer suggestions due to suggestions containing swear words, duplicate strings, and empty strings, which were filtered out and not shown to writers. For more rigorous filtering, we recommend using toxicity detection to exclude inappropriate suggestions.

B WRITING PROMPTS

Table 4 and 5 show ten prompts used in creative writing and argumentative writing, respectively. The prompts were retrieved from the WritingPrompts subreddit [62] and The New York Times [41] with minor modifications. Prompt codes were assigned by authors to easily refer to specific prompts in the paper.



(a) Creative writing **(b) Argumentative writing**

Figure 10: For both (a) creative and (b) argumentative writing, the number of times a writer (y-axis) chose a writing prompt (x-axis). Each writer could continue working on the same prompt up to five times.

Recall that writers could write about each prompt up to five times or choose to skip prompts if they wished not to write. Figure 10 shows the number of writing sessions per prompt and the top 20 writers with most participation. We observe that some writers chose

Prompt code	Prompt text (Source URL)
shapeshifter	A woman has been dating guy after guy, but it never seems to work out. She's unaware that she's actually been dating the same guy over and over; a shapeshifter who's fallen for her, and is certain he's going to get it right this time. (https://www.reddit.com/r/WritingPrompts/comments/7xihva/wp_a_woman_has_been_dating_guy_after_guy_but_it/)
reincarnation	When you die, you appear in a cinema with a number of other people who look like you. You find out that they are your previous reincarnations, and soon you all begin watching your next life on the big screen. (https://www.reddit.com/r/WritingPrompts/comments/7ezd5t/wp_when_you_die_you_appear_in_a_cinema_with_a/)
mana	Humans once wielded formidable magical power. But with over 7 billion of us on the planet now, Mana has spread far too thinly to have any effect. When hostile aliens reduce humanity to a mere fraction, the survivors discover an old power has begun to reawaken once again. (https://www.reddit.com/r/WritingPrompts/comments/7i3bs6/wp_humans_once_wielded_formidable_magical_power/)
obama	You're Barack Obama. 4 years into your retirement, you awake to find a letter with no return address on your bedside table. It reads "I hope you've had a chance to relax Barack... but pack your bags and call the number below. It's time to start the real job." Signed simply, "JFK." (https://www.reddit.com/r/WritingPrompts/comments/6b3rmg/wp_youre_barack_obama_4_months_into_your/)
pig	Once upon a time there was an old mother pig who had one hundred little pigs and not enough food to feed them. So when they were old enough, she sent them out into the world to seek their fortunes. You know the story about the first three little pigs. This is a story about the 92nd little pig. The 92nd little pig built a house out of depleted uranium. And the wolf was like, "dude." (https://www.reddit.com/r/WritingPrompts/comments/hytfcd/wp_then_the_92nd_little_pig_built_a_house_out_of/)
mattdamon	An alien has kidnapped Matt Damon, not knowing what lengths humanity goes through to retrieve him whenever he goes missing. (https://www.reddit.com/r/WritingPrompts/comments/8p3ora/wp_an_alien_has_kidnapped_matt_damon_not_knowing/)
sideeffect	When you're 28, science discovers a drug that stops all effects of aging, creating immortality. Your government decides to give the drug to all citizens under 26, but you and the rest of the "Lost Generations" are deemed too high-risk. When you're 85, the side effects are finally discovered. (https://www.reddit.com/r/WritingPrompts/comments/8on59a/wp_when_youre_28_science_discovers_a_drug_that/)
bee	Your entire life, you've been told you're deathly allergic to bees. You've always had people protecting you from them, be it your mother or a hired hand. Today, one slips through and lands on your shoulder. You hear a tiny voice say "Your Majesty, what are your orders?" (https://www.reddit.com/r/WritingPrompts/comments/88p6rp/wp_your_entire_life_youve_been_told_youre_deathly/)
dad	All of the "#1 Dad" mugs in the world change to show the actual ranking of Dads suddenly. (https://www.reddit.com/r/WritingPrompts/comments/6gl289/wp_all_of_the_1_dad_mugs_in_the_world_change_to/)

Table 4: For creative writing, we retrieved prompts from the WritingPrompts subreddit [62] and used them with minor modifications.

to write about each prompt repeatedly (e.g. A2W, A3O, and A34 in Figure 10 (a)), whereas others wrote about each prompt one or twice and moved onto next prompt (e.g. A3O and A11 in Figure 10 (a)). Also, writers sometimes skipped certain prompts completely (e.g. A1Q chose not to write about dating (da) in Figure 10 (b)).

C SURVEY QUESTIONS

Our survey consisted of five sections: writer information, benefits of collaborative writing, perceived capabilities of LMs, perceived limitations of LMs, and overall experiences. For writer information, we asked writers whether English is their first language, accounting for the different capabilities of LMs on native and non-native English speakers [8] (Is English your first language?). For benefits of collaborative writing, we wanted to understand whether human-LM collaborative writing has the known benefits of human-human collaborative writing, such as increase in fluency [4], pooling of knowledge and ideas [19], and enhanced writing quality [53]. To this end, we asked writers whether the suggestions they received contributed to the fluency of the resultant text, whether the suggestions helped them come up with new ideas, and whether they felt like that they would have written a better essay if they wrote the essay alone in 7-point Likert scale. For perceived capabilities of LMs, we wanted to understand capabilities of LMs perceived by writers. Specifically, we considered the notion of *competence* (having expert knowledge and ability to perform a task successfully) [40] and asked writers whether they think the system was competent in writing, whether the system was capable of writing creative stories or persuasive essays, and whether the system understood what they were trying to write. For perceived limitations of LMs, we asked writers which aspects of the suggestions (that they received during each writing session) can be improved and ask them to provide specific examples they have. For overall experience, we included common questions asked in NLP papers. Then, we checked whether some of our hypotheses are meaningfully reflected and observable through these questions. We asked writers about ease of writing (It was easy to write with the system), satisfaction (I am satisfied with the story/essay I wrote.), confidence (I am confident in my ability to write a story/essay with the help of the system.), ownership (I feel like the story/essay is mine.), and willingness to reuse (If the system is available for free, I would reuse the system.).

D QUALIFICATION ROUND

In the qualification round, the following conditions were used to allow experienced crowd workers (writers) to participate in our qualification round:

Prompt code	Prompt text (Source URL)
screen	How Worried Should We Be About Screen Time During the Pandemic? The coronavirus pandemic ended the screen time debate: Screens won. We all now find ourselves on our screens for school, for work and for connecting with family and friends during this time of social distancing and increased isolation. But should we be worried about this excessive screen use right now? Or should we finally get over it and embrace the benefits of our digital devices? (https://www.nytimes.com/2021/01/22/learning/how-worried-should-we-be-about-screen-time-during-the-pandemic.html)
dating	How Do You Think Technology Affects Dating? Have you had any experience with dating? Have you ever used dating apps? If so, what has it been like for you? If not, why not? How do you think technology — like apps, Netflix, social media and texting — affects dating and relationships? In your opinion, does it improve or worsen romantic interactions? How so? (https://www.nytimes.com/2018/02/21/learning/how-do-you-think-technology-affects-dating.html)
pads	Should Schools Provide Free Pads and Tampons? Have you ever experienced period shaming, or "period poverty"? Should schools step in to help? Should schools be required to provide free pads and tampons to students? How are pads and tampons similar to toilet paper, soap, Band-Aids and other products that are already provided in schools? How are they different? (https://www.nytimes.com/2020/11/18/learning/should-schools-provide-free-pads-and-tampons.html)
school	What Are the Most Important Things Students Should Learn in School? In your opinion, what are the most important things students should learn in school? What is the most important thing you have learned in school? How has this knowledge affected your life? How do you think it will help your success in the future? (https://www.nytimes.com/2019/02/21/learning/what-are-the-most-important-things-students-should-learn-in-school.html)
stereotype	What Stereotypical Characters Make You Cringe? What stereotypical characters in books, movies or television shows make you cringe and why? Would you ever not watch or read something because of its offensive portrayal of someone? (https://www.nytimes.com/2017/11/16/learning/what-stereotypical-characters-make-you-cringe.html)
audiobook	Is Listening to a Book Just as Good as Reading It? Do you listen to audiobooks? What are the benefits, in your opinion, of listening instead of reading? Are there advantages to reading that cannot be gained by listening? Which method do you prefer? Why? (https://www.nytimes.com/2018/12/12/learning/is-listening-to-a-book-just-as-good-as-reading-it.html)
athletes	Should College Athletes Be Paid? Do you think college athletes should be paid? Or is a college scholarship and other non-monetary perks like the opportunity to play in front of cheering fans enough? [...] What possible difficulties or downsides might there be in providing monetary compensation to players? (https://www.nytimes.com/2019/02/26/learning/should-college-athletes-be-paid.html)
extremesports	Is It Selfish to Pursue Risky Sports Like Extreme Mountain Climbing? Some sports, like extreme mountain climbing, are dangerous. Since there are varying degrees of risk in most, if not all, sports (such as the possibility of concussions, broken bones and even death), how does one decide where the line might be drawn between what is reasonable and what is not? Are some sports simply too dangerous to be called a sport? (https://www.nytimes.com/2019/04/29/learning/is-it-selfish-to-pursue-risky-sports-like-extreme-mountain-climbing.html)
animal	Is It Wrong to Focus on Animal Welfare When Humans Are Suffering? Would you be surprised to hear that a study found that research subjects were more upset by stories of a dog beaten by a baseball bat than of an adult similarly beaten? Or that other researchers found that if forced to choose, 40 percent of people would save their pet dog over a foreign tourist. Why do you think many people are more empathetic toward the suffering of animals than that of people? In your opinion, is it wrong to focus on animal welfare when humans are suffering? Why do you think so? (https://www.nytimes.com/2018/04/11/learning/is-it-wrong-to-focus-on-animal-welfare-when-humans-are-suffering.html)
news	Are We Being Bad Citizens If We Don't Keep Up With the News? In your opinion, are we being bad citizens if we don't keep up with the news? Do you think all people have some responsibility to know what is going on in the world? Does engaging with current events actually do anything at all? Why do you think the way you do? (https://www.nytimes.com/2018/03/20/learning/are-we-being-bad-citizens-if-we-dont-keep-up-with-the-news.html)

Table 5: For argumentative writing, we retrieved prompts from The New York Times [41] and used them with minor modifications. At the end of each prompt, we added "In my opinion," to give GPT-3 a clear signal to start responding to the prompt (as opposed to generating the continuation of the prompt).

- HIT Approval Rate (%) for all Requesters' HITs is greater than 97
- Location is the United States
- Number of HITs Approved is greater than 10000.

We specified two requirements for passing the qualification in the instructions (Figure 11 in Appendix E): (1) to ensure a story has a clear ending or an essay has a clear stance and conclusion, and (2) to collaborate with the system to write a story or an essay for at least ten minutes.

We used the following rubrics below to rate submissions in the qualification round. The authors of this paper manually rated 201 submissions with the goal of qualifying writers who demonstrated that they could meet two requirements. Writers whose submissions were rated as 4 or 5 were qualified to participate in the main round. Note that writers who did not interact with the system were

automatically disqualified, since we wanted to confirm that the writers were mindful of the requirement and demonstrated that they understood the functionality and could interact with the system.

- 5: A great story with a clear ending, or a great essay with a clear stance and conclusion
- 4: A reasonable story with a clear ending, or a reasonable essay with a clear stance and conclusion
- 3: A story *without* a clear ending, or an essay *without* a clear stance and conclusion
- 2: A below average story *without* an ending, or a below average essay *without* a clear stance and conclusion
- 1: Spam

E INSTRUCTIONS

Figure 11 shows the instructions used for Amazon Mechanical Turk, specifically the ones used for creative writing in the qualification round. The instructions for the main round and argumentative writing were nearly identical except for the first paragraph (which specified whether this is the qualification round or main round) and specific wordings for stories and essays.

Note: This is a **qualification round** for writing with AI. You need to participate **only once** for the qualification round! Please excuse us if takes several days to get qualified as we are going through everyone's story one by one.

We are looking for participants who are interested in **collaborating with AI** to write short, creative, and interesting stories. If you prefer to write alone or prefer to let AI write everything for you ***should not*** participate in this HIT.

The goal of the qualification round is to make sure that participants (1) know how to collaboratively write with AI (e.g. taking suggestions and editing them) and (2) can write a short, interesting story that has a clear ending. The main round will have an identical task of story writing but with potentially different prompts.

Please complete this HIT only if you are interested in writing short stories in the next round of story collection. We appreciate your interest! :)

Step 1: Consent Form

DESCRIPTION: You are invited to participate in a research study on human-AI interaction, which aims to deepen our understanding of how humans and an AI agent collaborate in the writing process and identify strengths and weaknesses of AI in assisting humans as a writing assistant. You will be asked to write a story in the provided text editor. The editor will show suggestions for next sentences from which you can choose a suggestion to incorporate into your story. Your textual interaction (e.g. insert, delete, select) can be recorded and released for research purposes.

TIME INVOLVEMENT: Your participation will take approximately 15 minutes.

RISKS AND BENEFITS: The risks associated with this study is that suggestions generated by AI may contain offensive, triggering, or factually unreliable contents due to the prevalence of such contents on the Internet. If you feel considerable distress from these contents, please discontinue participation at any time and reach out to trained counselors from free and confidential treatment resources such as Suicide Prevention Lifeline (1-800-273-8255) or Crisis Text Line (text SIGNS to 741741) (more resources can be found at https://www.cdc.gov/mentalhealth/tools-resources/individuals). Also, because textual interaction can be released for research purposes, we advise you not to write any identifiable private information in the text editor. The benefits which may reasonably be expected to result from this study are to write one or more stories, gain confidence in writing, and potentially reduce the fear of writing. We cannot and do not guarantee or promise that you will receive any benefits from this study.

PAYMENTS: You will receive $2.5 as payment for your participation.

PARTICIPANT'S RIGHTS: If you have read this form and have decided to participate in this project, please understand your participation is voluntary and you have the right to withdraw your consent or discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled. The alternative is not to participate. Note that you are only entitled to receive the payment upon the completion of your participation. The results of this research study may be presented at scientific or professional meetings or published in scientific conferences. Your individual privacy will be maintained in all published and written data resulting from the study.

CONTACT INFORMATION: Questions: If you have any questions, concerns or complaints about this research, its procedures, risks and benefits, contact the Protocol Director, [ANONYMIZED] at [ANONYMIZED] or email at [ANONYMIZED].

Independent Contact: If you are not satisfied with how this study is being conducted, or if you have any concerns, complaints, or general questions about the research or your rights as a participant, please contact [ANONYMIZED] to speak to someone independent of the research team at [ANONYMIZED] or toll free at [ANONYMIZED], or email at [ANONYMIZED]. You can also write to [ANONYMIZED].

Please print a copy of this page for your records.

If you agree to participate in this research, please click the checkbox below to indicate your consent.

☐ **I agree to participate in this research. (required)**

Step 2: Writing with Artificial Intelligence

Goal: Write an interesting story with an AI-powered writing assistant!

Whenever you feel stuck or just want some inspiration, **press the tab key** -- our AI-powered system will show suggestions at the end of your story. If you need more ideas, just press the tab key multiple times to get a new set of suggestions each time.

If you like any of the suggestions, **click** them to add to your story! You can also use **up/down arrow** keys to navigate the suggestions and press the **enter** key to add to your story. Revise your text or suggestions to make overall story more interesting and creative.

Requirements: This is an open-ended task. The only two requirements are (1) to ensure your story has a **clear ending** and (2) to **collaborate with AI** to write a story for at least **10 minutes**. If you're idle for more than 1 minute, we reserve the right to reject your HIT.

Instructions:

1. When you click the link below, you will be given a text editor with a prompt.
2. Write a short, interesting story that starts with the prompt and has a clear ending.
3. While writing your story, you can write on your own or use the suggestions at any point.
4. Once you're done, click the finish button which will give you a verification code.
5. Copy and paste your story and verification code below.

Warning: Please ***do not write any individually identifiable information*** in your story!

Click here to start writing! ([ANONYMIZED])

Self-evaluation to check whether the two requirements are met

Title (optional)

Copy and paste your story - do not write here directly! (required)

☐ **My story has a clear ending. (required)**

☐ **I collaboratively wrote my story with AI. (required)**

Enter verification code (required)

Step 3: Survey

Instructions:

1. Complete the survey by clicking the link below.
2. Once you're done, click the submit button which will give you a survey code.
3. Copy and paste your survey code below.

Click here to start survey! ([ANONYMIZED])

Enter survey code (required)

Submit

Figure 11: Instructions used for Amazon Mechanical Turk.