The HaLLMark Effect: Supporting Provenance and Transparent Use of Large Language Models in Writing with Interactive Visualization

Md Naimul Hoque
nhoque@umd.edu
University of Maryland
College Park, MD, USA

Tasfia Mashiat
tmashiat@gmu.edu
George Mason University
Fairfax, VA, USA

Bhavya Ghai
bhavyaghai@gmail.com
Amazon
New York, NY, USA

Cecilia Shelton
sheltonc@umd.edu
University of Maryland
College Park, MD, USA

Fanny Chevalier
fanny@dgp.toronto.edu
University of Toronto
Toronto, ON, Canada

Kari Kraus
kkraus@umd.edu
University of Maryland
College Park, MD, USA

Niklas Elmqvist
elm@cs.au.dk
Aarhus University
Aarhus, Denmark

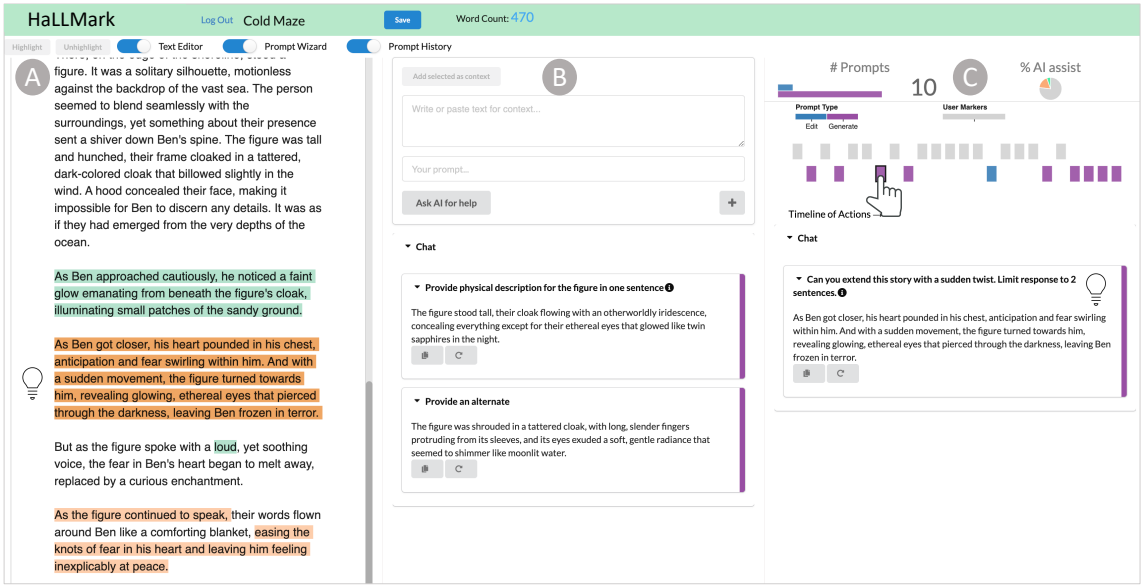

Figure 1: The HaLLMark System. (A) Text editor for viewing and editing text. The system highlights text written (orange) and influenced (green) by the AI. There are three toggle buttons on top of the editor to turn on and off the three views (columns) of the interface. (B) Prompting interface for large language models such as GPT-4. The user can see the prompts and AI responses for the current session. (C) Summary statistics show the number of prompts and percentage of user-written text and AI assistance. Below is a timeline of a user's writing actions (grey rectangles) and interaction with the AI (purple and blue rectangles). The user can hover over any glyphs in the timeline to see the relevant prompt and linked text in the text editor.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05.
https://doi.org/10.1145/3613904.3641895

ABSTRACT

The use of Large Language Models (LLMs) for writing has sparked controversy both among readers and writers. On one hand, writers are concerned that LLMs will deprive them of agency and ownership, and readers are concerned about spending their time on text generated by soulless machines. On the other hand, AI assistance can improve writing as long as writers can conform to publisher

policies, and as long as readers can be assured that a text has been verified by a human. We argue that a system that captures the provenance of interaction with an LLM can help writers retain their agency, conform to policies, and communicate their use of AI to publishers and readers transparently. Thus we propose HaLLMark, a tool for visualizing the writer's interaction with the LLM. We evaluated HaLLMark with 13 creative writers, and found that it helped them retain a sense of control and ownership of the text.

CCS CONCEPTS

• **Human-centered computing → Visualization systems and tools**.

KEYWORDS

Creative writing, co-writing, LLMs, agency, visualization.

ACM Reference Format:

Md Naimul Hoque, Tasfia Mashiat, Bhavya Ghai, Cecilia Shelton, Fanny Chevalier, Kari Kraus, and Niklas Elmqvist. 2024. The HaLLMark Effect: Supporting Provenance and Transparent Use of Large Language Models in Writing with Interactive Visualization. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3613904.3641895

1 INTRODUCTION

Large Language Models (LLMs) are here to stay. Their arrival has been particularly widely panned in the area of creative writing, where doomsayers are hinting at a future where it will become impossible to distinguish between a writer's original work and text generated by an LLM. Worse yet, LLMs are feared to potentially inundate us with a cornucopia of poor text. What happens to original thought if we are subjected to writing forever recycled and regurgitated from original work that has already been produced by our forebears? It has even been argued that some essential aspect of what it means to be human is lost if creative writing is performed by machines. While these are valid concerns, it remains that LLMs are just another tool in a long line of tools, and we should find ways to harness the technology in a just, responsible, and ethical way rather than attempt to suppress it. But in trying to embrace this technology, one critical concern for writers who want to use LLMs is properly attributing contributions from the AI. This gives rise to an "ownership tension" among writers and hampers their agency and control over the process [25, 51]. Identifying contributions from AI is also important for writers who need to conform to AI-assisted writing policies [28, 53]. Authors currently have few mechanisms to track their accountability with regard to these rules and policies.

In this work, we argue that capturing interactions between AI and writers as the document evolves (i.e., provenance information) and supporting interactive exploration of such provenance information will **improve** the writer's agency, control, and ownership of the final artifact (e.g., short stories, novels, poems). Prior research suggests that externalizing provenance can help balance automation and agency in human-AI collaboration [60]. Provenance can also help writers conform to AI-assisted writing policies and provide transparency to publishers and readers. To explore this design space, we first reviewed existing guidelines and policy documents on the use of LLMs from several professional, educational, and academic organizations (Section 3). This review informed us about the types of information that writers should be aware of in AI-assisted writing. We then developed HaLLMark, a web-based technology probe [35] that integrates an authoring interface with LLM support that stores and visualizes a writer's interaction with an LLM (Section 4). The system facilitates writers in self-reflecting on their use of AI by clearly highlighting AI contributions.

To validate HaLLMark, we engaged a group of creative writers to use it to write a short story during remote evaluation sessions and collected their feedback and resulting stories (Section 5). Our findings suggest HaLLMark encouraged writers to actively evaluate AI-assistance from the onset of the writing process. As a result, it instilled a sense of control in the writer's mind and improved ownership of the final artifact. Our findings also suggest HaLLMark will help writers conform to AI-assisted writing policies without the need for manually generating disclosures. Writers were therefore confident that HaLLMark will help them become more transparent and that it is an effective medium to communicate the use of AI in writing to external parties (i.e., publishers). We close the paper by outlining how other stakeholders (e.g., readers, peer-reviewers, publishers) could potentially use and benefit from our approach (Section 6). We also discuss the broader impact of our work on the ethical use of LLMs in the writing and publication industry.

2 RELATED WORK

Our work intersects with prior art on intelligent writing support tools, concerns on the use of LLMs, and the use of visualization in literature and writing. Here we describe each of these topics.

2.1 Writing Support Tools

Computers have long been used as writing support tools, harking back to the spelling check feature on Microsoft Word in 1997 and all the way to modern-day Large Language Models (LLMs). In fact, one could argue that the ubiquitous typewriter and printing press are examples of enabling technology for writers. One popular category of writing support tools is paid software such as Scrivener [49] and Granthika [27] that try to enhance the organizational capabilities of a writer. An additional category of writing support tools targets academic writers who need support of various kinds. More relevant to our work are tools that enhance creativity of writers through *interactive* and *intelligent* features. Examples of such works include support for metaphor creation [24, 40], automatic text summarization [14], interaction with literary styles [61], and support for the iterative revising process [17, 18].

More recently, the introduction of LLMs has fueled a new generation of writing support and *co-writing* tools. These tools can generate human-like text and inspire new narrative angles and ideas. For example, CoAuthor [43] and Wordcraft [76] can generate new sentences and passages to help writers develop short stories. Dramatron [51] is a similar kind of system but generalizes to long-form writing through hierarchical chaining of prompts. Sparks [25] focuses on science writing, while TaleBrush [12] can generate texts to match a character arc sketched by the author. HaLLMark, the tool proposed in this paper, is built on similar mechanisms but has a different focus: to leverage interactive provenance to help

writers reflect on their use of the LLM, to conform to new policies on AI-assisted writing, and to retain their ownership as well as transparently communicate the influence of the AI on the text.

2.2 Concerns around LLMs for Co-Writing

While the capabilities of LLMs—similar to other generative AI tools—have awed writers from different domains, their use for creative purposes is controversial [20]. The Writers Guild of America (WGA) and the Screen Actors Guild – American Federation of Television and Radio Artists (SAG-AFTRA) were recently on strike, the former from May to September 2023, and the latter from July to November 2023. Along with typical demands such as better pay structure, especially for streaming services, the main demands from protesters were to add contract language that protects them from being replaced by machines (writers from AI-generated text and actors from studios using their AI-generated likenesses). Similar sentiments have been reported in several recent studies. Writer concerns for LLMs include agency and ownership [51, 72], ethics and plagiarism [72], and lackluster, stereotyped text [25, 51].

Creative writers who are pushing the envelope of technology view generative AI as just another tool that can help them support their work more efficiently. Even among these supporters, a majority prefer to limit the use of generative AI to supporting their editing, brainstorming, or organizing rather than asking it to creatively generate the text of their work in order to retain their own agency, ownership, and artistic expression [4]. This echoes principles of meaningful human control in generative AI articulated by Epstein et al. [20]. Given the paired, if sometimes conflicting, interests of writers who want to both embrace the affordances of new generative AI technologies and also carefully and thoughtfully limit the ratio of AI-generated text output in their final work, it is clear that we need guidelines and policies for ensuring responsible use of LLMs in writing that account for the strengths and weaknesses of the tools as well as ethical concerns regarding their use.

In response, organizations and publishers such as the U.S. Copyright Office [53], Author's Guild [28], and ACM [22] have released guidelines for AI-assisted writing. These policies ask writers to track their interactions with LLMs (i.e., provenance) and report interactions to show that writers had creative control over the generation of the text. However, it is not clear how writers can operationalize these guidelines in their writing and report use of AI transparently. This paper first formalizes the guidelines from existing policies into actionable items and then presents a tool that writers can use to conform to the policies. The result is a tool that supports provenance for authors, helping them regain agency and authorship, while at the same time, allowing them to conform to the policies and transparently communicate the process to others (e.g., readership, publishers).

2.3 Visualization for Text and Writing

Data visualization can be particularly helpful for summarizing large volumes of text [8]. Text is largely an unstructured data format, making it difficult to see hidden patterns in text-based artifacts, such as a novel, document collection, or newspaper article. Text visualization is the area of visualization research that invents new representations to summarize and comprehend text data [1, 8]. Examples of text visualizations include the ubiquitous word cloud [67], wordle [69], and the Word Tree [71]. More complex representations also convey structure in a text corpus, such as Phrase Nets [65], TextFlow [13], Elastic Documents [5], and ThemeDelta [23]. Jänicke et al. [36] provide a survey on the use of text visualization and analytics in support of close and distant reading in the digital humanities.

Despite the prevalence of text visualization in the academic community, application of these representations in writing support tools is limited. One example is DramatVis Personae (DVP) [32], a visualization system for mitigating nuanced social biases in creative writing. In follow-up work, the DVP authors developed a tool to visualize different character traits [33]. Finally, Poemage [50] helps literary scholars understand the sonic properties of a poem.

Visualization has long been used to provide explainable machine learning [73] and NLP models [11, 45], and these ideas have also recently begun to be applied to LLMs. Recently, Jiang et al. [39] proposed Graphologue, that converts responses from LLMs to interactive graphs for fast and non-linear sensemaking. Sensescape [63] supports multilevel organization of information gathered from LLM responses. However, none of these tools focus on supporting writing, provenance, or transparency, a gap we aims to address.

2.4 Visualizing and Tracking Collaboration

A field of research relevant to our work is visualization methods and systems proposed to track contributions from multiple parties in a collaborative setting [46, 66]. For instance, researchers have shown that visualization can track provenance in collaborative writing [7, 10, 37, 64, 75]. History Flow [68] and DocuViz [70] are examples of visualization techniques for studying co-authorship patterns (cooperation and conflicts) in collaborative writing.

Several visualization systems have been proposed to track provenance in human-AI collaboration [57]. For example, tracking model and data performance is a key need for interactively developing machine learning models. Amershi et al. proposed ModelTracker [2] to support this need. ModelTracker is an interactive visualization that summarizes traditional summary statistics and graphs while displaying example-level performance to enable direct error examination and debugging. Chameleon [29] uses a collection of visualizations to allow users to compare data features, splits, and performance across data versions. Other works in this area have focused on visualizing contributions from AI and humans for a specific task. For example, Rogers and Crisan recently proposed AutoML Trace [58], a system for visualizing contributions from humans and AI in AutoML. Wu et al. [74] showed that decomposing an LLM task into multiple sub-tasks, chaining them, and then allowing users to investigate how a previous task influences the subsequent task improved users' task quality, sense of control, and system transparency.

While these works inspired us to utilize visualization in externalizing provenance information, AI-assisted writing—especially with LLMs—is a fairly new research area with emerging challenges for tracking provenance. These challenges include understanding the types of information that writers should be aware of in AI-assisted writing, coupling requirements from writers and policies on AI-assisted writing, and supporting writers' needs with interactive visualization. This paper aims to address these challenges.

Dimension	Category	Examples
D1. Prompt category	Asking for editing an existing text	No need to report common editorial assistance (grammar check and paraphrasing) [21, 22, 53]
	Asking for generating new texts	Report prompts if they are used to generate an extensive amount of text [22]
D2. Using AI response	Explore: AI response was not inserted in the text	Report use if AI generated new ideas [21]
	AI response was inserted in the text	Highlight text written by the AI [53, 55]
D3. Summary statistics	Number of prompts used; Percentage of text written by the AI	AI-written text should not be more than 5% [28]

Table 1: AI-assistance writing policies. Summary of types of information required by AI-assisted writing policies.

3 FORMATIVE ANALYSIS OF AI-ASSISTED WRITING POLICIES

To better understand the dimensions involved in concerns around intellectual ownership and ethical use of LLMs, we perform an analysis of publishing outlet policies for AI co-writing. This included aspects such as when, how, and to what extent LLMs can be used in the creative authoring process. Table 1 contains our consolidated typology of current types of information that are important to be aware of during AI-assisted writing, per our findings.

A tool supporting awareness and deeper understanding of these dimensions will help authors effectively and responsibly leverage powerful aids. As a starting point, we reviewed existing guidelines for using generative AI models from the U.S. Copyright Office [53], The Author's Guild [28], educational organizations such as the Modern Language Association [3], several creative writing publishers (e.g., [55]), and academic venues and publishers (e.g., ACL [21] and ACM [22]). We recognize that the list is not exhaustive and as we collectively learn about usage, policies will evolve.

We conducted thematic analysis [9] to identify key dimensions from the policies. Two authors of this paper individually reviewed the policies and open-coded the recommendations from the policies. The authors then met to discuss key observations and developed a codebook. After that, the authors open-coded the policies again by following the codebook. The inter-rater reliability between the coders was 0.91 (Jaccard's similarity). Finally, the coders met to resolve disagreements and finalize the themes. The full research team participated in the discussion with the coders regularly. The full list of the coded policies is available in our OSF repository.

3.1 Patterns and Differences

There was a consensus among policies and guidelines, irrespective of their domain (government, creative, academic, and education), on the need to **report** the extent of contribution from AI in the creation of content. While specific instructions vary between policies, all encourage authors to be transparent and disclose the use of AI. Another shared sentiment across policies and domains is that AI cannot be granted authorship; rather, authors should be responsible for content generated by the AI and should acknowledge that they have themselves verified all AI-generated content.

We also noticed some differences in the policies from different domains. For instance, policies from creative writing are more concerned about copyright issues and preventing LLMs from training on books without writers' permission than providing guidelines on how writers should use LLMs [28]. In comparison, policies from academic venues provide explicit guides for using and reporting LLMs [21, 22]. Academic policies also put more emphasis on fact-checking and proper referencing [21] than policies from creative domains. Policies from educational venues are concerned about plagiarism and academic integrity [3].

Finally, there is ambiguity in the description of what needs to be reported and how writers should do that. For example, the U.S. Copyright Office mentions several ways to disclose the use of AI: a brief note, acknowledgments, or providing exact contents provided by AI. On the other hand, ACL discourages writers from using AI-generated text directly. *Nature* directs authors to report the use of AI in the method section or alternative section of the manuscript [52], but does not explicitly specify what writers should report.

3.2 Information Typology

Table 1 presents the three major themes that emerged from the analysis—each of which corresponds to a dimension capturing the information required by AI-assisted writing policies. The first dimension, recurrent in the policies, separates prompts that seek new content from prompts that merely request edits to an existing text (**D1**). Most policies recognize that we have been using computers for editorial tasks such as spelling and grammar check for a long time and writers do not need to disclose the use of LLMs for such tasks [21]. However, writers should report prompts that seek new content. For example, ACM directs authors to provide a list of generative prompts used for such purposes [22].

All policies ubiquitously ask writers to disclose text generated by AI (**D2**). For example, The Author's Guild specifies that *"Authors shall disclose to Publisher if any AI-generated text is included in the submitted manuscript."* Other policies have similar clauses. Some publishers reserve the right to reject papers that were mostly generated by AI [53]. An interesting case is when an author does not directly use AI-generated text in their article, but still draws inspiration or ideas from the AI, or derives narrative angles from it. ACL requires authors to also acknowledge such use [21].

While it is rare, some publishers recommend specific thresholds for using AI-written texts (**D3**). For example, The Author's Guild restricts the authors to limit the use of AI-generated text to only 5% [28]. However, it does not provide any guidelines on how this 5% should be measured in practice nor where to report these statistics.

4 THE HALLMARK SYSTEM

We used our formative analysis of the AI-writing policies (Table 1) to inform the design of HaLLMark, a technology probe [35] that couples a text editor, an LLM, and an interaction history. "Technology probes are simple, flexible, and adaptable technologies with the goal of understanding user needs and desires in a real-world setting, field-testing the technology, and motivating the design of new technologies" [35]. We decided on such a probe as a suitable method as it would allow us to understand how tracking provenance information can help writers, what features in HaLLMark writers find most useful, and what would be the design of future systems aiming to ensure human agency and transparency in AI-assisted writing. Two domain experts, who are professors of English and writing studies at our university and co-authors of this paper, provided feedback during the development cycle of HaLLMark. In this section, we first present the system's design rationale and then describe its details.

4.1 Design Rationale

We designed HaLLMark based on the following design rationale:

DR1 **Capture and externalize AI vs. human provenance.** The three prominent dimensions of information from AI-assisted writing policies are *D1. prompt category*, *D2. how writers used AI responses*, and *D3. summary statistics*. Thus, our primary goal is to store this information when a writer interacts with an LLM. To support provenance, the system should externalize this information to the writer in an easily understandable format. The writer should then be able to freely go back and forth in the interaction history.

DR2 **Integrate provenance in artifact.** Externalizing interaction history may not be enough. To make sense of the history, the system should connect the artifact (i.e., the text document) with the history [32, 33]. Moreover, some policies require authors to highlight text written by AI in the artifact. Thus, content generated by the AI should be clearly visible in the text and linked to the information stored for DR1.

DR3 **Integrate writer's judgment in provenance.** While we aim to store user interaction automatically, we may not be able to store all information automatically. For example, one category in Table 1 references when a writer takes inspiration from AI outputs, but does not use the output directly. Since the influence here is implicit and difficult to quantify, we decided to facilitate mechanisms for writers to integrate this information in the interaction history. To improve agency for writers, we also decided that writers should be able to edit or modify the interaction history if needed. One implication of this decision is that our tool is not a tool to enforce AI-writing policies; rather it is a tool for writers to be able to measure their own compliance, while being able to design disclosures and be transparent.

DR4 **Extensible/flexible.** While existing policies provided a baseline for our work, they are still evolving. The technology around LLMs is also particularly fluid. Thus, our design should be extensible to new requirements in the future.

4.2 Visual Interface

Figure 1 shows the full interface for HaLLMark. Because of the rich dependencies between text, prompts, and interaction history, we opted for a visual approach with data visualization components. The interface is divided into three modules: a) a rich text editor; b) an interface to interact with GPT-4; and c) a module to visualize interaction history with LLMs. By default, we allocate 1/3 of the screen width to each module. However, we provide toggle buttons to hide or show each module. Whenever a writer toggles the visibility of a module, we redistribute the available width to the visible modules equally. For example, a writer can hide the GPT-4 and visualization module to write "distraction-free". We describe individual components of HaLLMark below, and include a video of the tool as supplementary material.

4.2.1 Prompting LLMs. The prompting interface in HaLLMark bears similarities with the current ChatGPT interface. It has a text box for writing the prompt and an optional text box for specifying the context of the prompt (Figure 2B). A user can highlight a portion of the text in the editor to be automatically selected as an additional context for prompting the AI (Figure 2). The response from an LLM such as GPT-4 gets appended below the text boxes. The prompt wizard suggests several standardized creative composition interactions, such as "summarize," "elaborate," "enumerate," "introduce," and "conclude." A writer can write a free-form prompt or choose one from the standardized recommendation.

4.2.2 Prompt Card. We encapsulate each prompt and the relevant AI response in a card, the popular UI component for designing modular objects (Figure 3). As per **DR1**, we categorize each prompt as either seeking generation of new contents or editorial help on an existing text. We use the following soft prompt with the actual prompt to identify the category: `"For the input text, reply 'Edit' or 'Generate' if the text intends to edit existing text or generate new text. Consider paraphrasing an existing text, or grammatical and spelling check as an Edit. Input sentence - " + input prompt.` To validate the performance of this method, we created a dataset of 150 prompts. Two authors of this paper collaboratively created the prompts and then labeled them as either targeted at generating new content or focused on editorial support of existing content. We ensured that the prompts spanned a wide range of writing compositions and were challenging to decode. We then measured the accuracy of the soft prompt in classifying the prompts correctly. The accuracy was 96%. The list of prompts and their labels are available in the our OSF repository. We encode the category of the prompt at the right border of the card body with either `purple` color for indicating generation or `blue` color for indicating edit.

4.2.3 Visualizing AI vs. human provenance. We externalize the provenance information using several interactive visualizations (**DR1**). We visualize two summary statistics in HaLLMark (Figure 4A): the counts for the prompts in a bar chart; and in a pie chart, the percentages for AI-written, AI-influenced, and text written by the writer. We consider text written by the AI and then used verbatim to be as AI-written and highlight them with an `orange` color in the text editor and pie chart. When a user copies full or parts of the response from the prompt card and pastes it into the text editor,

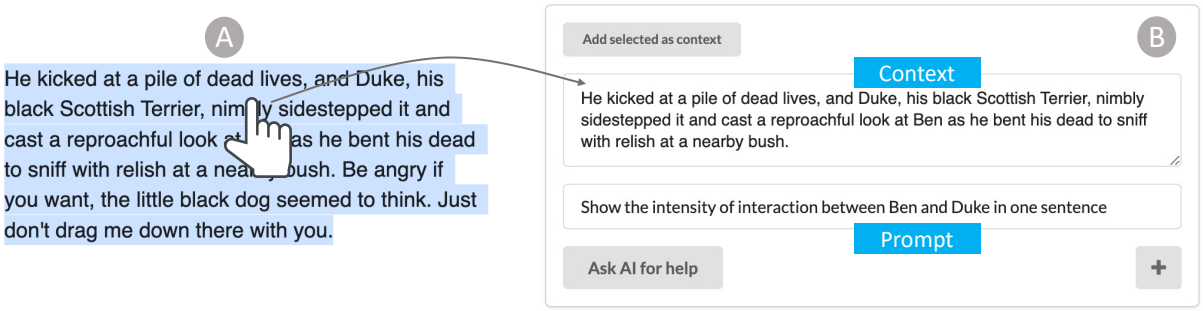

Figure 2: Prompting GPT-4 in HaLLMark. A) By highlighting any portion of the text in the text editor, the user can select that text as context for prompting GPT-4. B) The selected text is automatically pasted into the context box. The user can specify the task to perform in the prompt box.

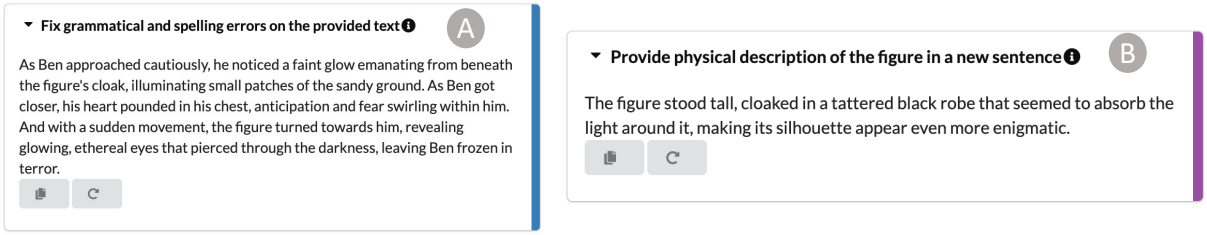

Figure 3: Design of the prompt card. We encapsulate each prompt and AI response in a card. The title shows the prompt. Users can hover over the information icon to see the context. Each card contains a copy button and redo button for regenerating the AI response. We categorize each prompt as either seeking new contents (blue) or seeking editorial help (purple) on an existing text. For instance, A) shows a prompt seeking new content, and B) is a prompt seeking editorial help.

we automatically highlight that text as AI-written, update the pie chart, and link the prompt with the text (**DR2**). If a user re-writes or edits a portion of the AI-written text, we remove the highlight from the specific portion of the text and mark it as written by the user. The other parts of the text remain as AI-written.

Writers can manually mark any text as AI-influenced (**DR3**). For instance, after re-writing an AI-written text, writers who feel that the text is influenced by the AI can mark the text as such. Following **DR3**, we do not automatically detect AI-influenced text, and rather leave it to the writers' discretion and judgment. This said, flagging potential AI-influenced text for writers to critically review could still be useful. To this end, we explored several automatic methods aimed at identifying AI-influenced text. For instance, following Kim et al. [41], we experimented with the BLEU score—a measure used to evaluate the similarity between a piece of text and references. We calculated the BLEU score between text re-written by a person with the responses from LLMs to see if we can reliably detect AI-influenced text. We also experimented with OpenAI's classifier for detecting AI-written text [54]. We decided against using these methods as our domain experts did not find the methods to be consistent and sufficiently reliable. OpenAI also lists the shortcomings of such methods [54]. We determined that wrong predictions and interpretations could negatively impact the user's experience, trust, and agency in the user study, and decided to omit the feature.

The timeline shows interaction history in a linear fashion (Figure 4B), using a colored rectangle (either blue or purple) for each prompt. We insert a new grey rectangle in the timeline each time the user writes a new line to show writing activity in comparison to prompting AI. The timeline is scalable to more variables and information as we can encode the new information in a new row in the timeline (**DR4**). The timeline also extends horizontally when the rectangle width becomes less than a threshold (default 5px) and provides a scroll bar to see the extended content.

A user can hover over the colored rectangles to see the linked prompts (Figure 4) and the linked text, if any (**DR2**). Upon clicking any rectangle, the respective prompt stays visible to the user.

4.2.4 Linking Visualization and Artifact. We used QuillJS [56] as a rich text editor in our interface, where a user can read or write textual content, and apply traditional formatting. Beyond these traditional operations, a user can also perform the following actions related to AI-assisted writing in the text editor:

- **Manually label text.** Following **DR3**, a user can select a portion of the text and then label the text as either AI-written or AI-influenced using a button named `Highlight`. Additionally, the user can link a prompt from prompt history with the highlighted text (Figure 5). This helps writers to manually annotate any text in the case where our system cannot automatically annotate them.
- **Manually remove label.** In a similar manner, writers can remove annotations (AI-written or AI-influenced) and links with prompts by first highlighting a portion of the text and then clicking a button named `Unhighlight`.

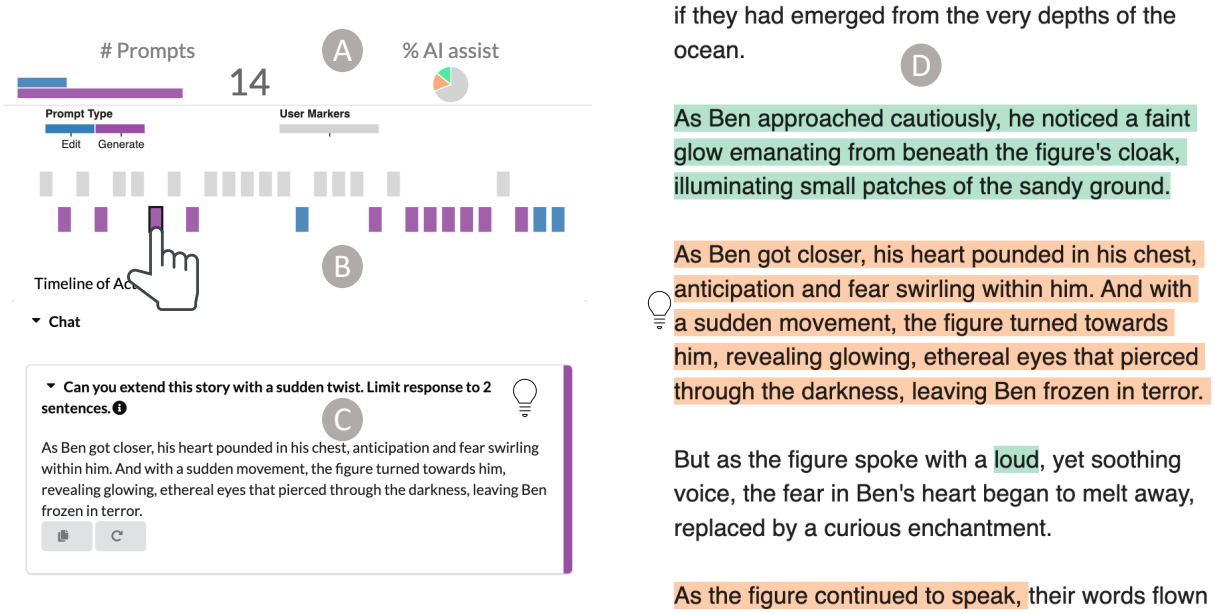

Figure 4: Visualization and interaction in HaLLMark. (A) Summary statistics: number of prompts and percentage of assistance from AI. (B) The timeline shows the prompts (blue or purple tiles) in the context of the user's writing behavior (e.g., writing a new sentence). Hovering over a colored tile will show the respective (C) prompt and text highlighted in the editor (D).

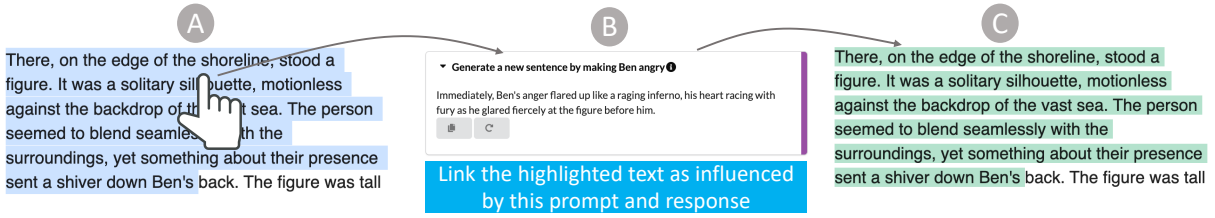

Figure 5: Manually linking a portion of the text with a prompt in HaLLMark. A) The user highlights a portion of the text. B) The user can link the text with a prompt from the prompt history. They can either label it as AI-written or AI-influenced. In this case, the writer labels it as AI-influenced. C) The text color changes to green to indicate the change in the label.

- **Click on annotated text.** A user can see the linked prompt to an annotated text by clicking on it (**DR2**).

5 EVALUATION

We conducted a user study with 13 creative writers. The overarching goal of the study was to answer our original high-level research question: *How can externalizing provenance information help AI-assisted co-writing?* We used HaLLMark as a technology probe to explore this question. Since provenance can impact many facets of AI-assisted writing (Sections 1 and 2), we seek to answer the following specific research questions (RQs):

RQ1: How does HaLLMark affect a writer's interaction with an LLM?

RQ2: How does HaLLMark affect a writer's ownership concerns while receiving AI-writing support from an LLM?

RQ3: How does HaLLMark help a writer to communicate the extent of their use of LLMs?

RQ4: How does HaLLMark help a writer to conform to policies on AI-assisted writing?

5.1 Different Forms of Writing

While HaLLMark is generalizable to different genres of writing (e.g., creative, argumentative, academic), each genre has different goals, tasks, and styles. Instead of recruiting writers from diverse domains, we conducted our case study with creative writers. Thus, while our evaluation might inform the adoption of HaLLMark in other domains, our results may well be specific to fiction.

Creative writers commonly use a wide range of techniques that are regarded as integral to creative, expressive writing: vivid, concrete language; metaphors and similes; syntactical variety; alliteration; and other literary devices. Furthermore, the impact of LLM *hallucination* [38] is less problematic for creative writing than non-fiction and academic writing, thus removing a potential confound. In addition, there is currently a significant backlash from creative

writers about the use of AI in creative writing, instrumented by the recent strike from screenplay writers and artists. LLMs directly threaten their bread and butter. Previous studies also reported that creative writers have ownership and agency issues when using LLMs [51, 76]. Given the premise of this work, creative writers are perfectly suited to help us answer our research questions.

5.2 Study Conditions

We conducted a repeated-measures within-subject experiment with the following two conditions (counterbalanced):

- **C1.** Baseline: A ChatGPT-like interface with a text editor. Participants are able to write, use GPT-4, and see a list of prompts and responses from GPT-4 in a sidebar. We include a screenshot of the baseline in the supplement.
- **C2.** HaLLMark: Our tool with all interactive support.

5.3 Participants

We recruited participants by advertising in our university's Writing Center as well as English, Literature departments. Our participants varied in terms of self-reported gender (male = 5, female = 7, prefer not to say = 1, other = 0), age (min = 19 years, max = 56 years, mean= 26 years, SD = 4.2 years), experience in writing different creative materials (fiction, non-fiction, short stories, and poems), and years of experience as creative writers (min = 5 years, max = 26 years, mean = 9.2 years, SD = 4.3 years). All participants had published works in their portfolio. Participants received a $40 USD gift card for their time.

All participants reported prior experience in using LLMs (e.g., ChatGPT) or were aware of their use in creative writing, but none mentioned LLMs to be a critical part of their writing process. Five participants reported using ChatGPT infrequently for various editorial writing tasks such as rephrasing a text or changing the mood of the text. Two participants had used ChatGPT to explore different narrative angles. Other participants had tested ChatGPT out of intellectual curiosity. Two participants had actively participated in the 2023 WGA/SAG-AFTRA strike.

5.4 Tasks

It is difficult to design tasks with objective goals for creative writers [32, 33]. Their work typically does not adhere to predefined structures and depends on their artistic styles and idiosyncrasies [33]. Thus, we opted to ask writers to write short stories using our interfaces for a fixed amount of time (20 mins) while prompting GPT-4. We aimed to study their interaction patterns and collect feedback through semi-structured interviews to answer the RQs.

5.5 Measures

Since the study tasks did not involve any objective goals, we opted for a qualitative methodology. Another reason for this choice is that concepts relevant to our study (e.g., agency, transparency, ownership) are mostly abstract concepts and are difficult to operationalize quantitatively [76]. Instead, we designed a semi-structured interview for capturing writers' feedback. We asked writers about how the study conditions impacted their interaction with LLMs, agency, control, and ownership. We also asked them about the usefulness of each interface to support communication and transparency around AI-assisted writing. The interview script is available on OSF.

We also asked participants to rate the study conditions on a 7-point Likert scale across three subjective dimensions:

- *Ownership*: On a scale of 1 (not at all comfortable) to 7 (very comfortable), how comfortable would you be in publishing the short story under your name?
- *Communication and Transparency*: On a scale of 1 (not helpful at all) to 7 (very helpful), how helpful would the tool be to you in communicating your use of AI to others (e.g., publishers, readers) for transparency?
- *Conformity*: On a scale of 1 (not helpful at all) to 7 (very helpful), how helpful would the tool be to you in conforming to the given AI-assisted writing guideline?

Finally, following prior literature [41, 44], we asked participants to rate each condition on a 7-point Likert scale (1: strongly disagree, 7: strongly agree) across the following six dimensions for capturing the usability of LLM support:

- *Helpful*: "I found the AI helpful."
- *Ease*: "I found it easy to write the advertisement."
- *Experiment*: "I felt that I experimented with various ideas and generated alternatives."
- *Iteration*: "I felt that I iterated various times on ideas and the generation process."
- *Pride*: "I am proud of the final output."
- *Unique*: "The story I wrote feels unique."

5.6 Procedure

Before each session, we asked participants to familiarize themselves with the policy on AI-assisted writing from the U.S. Copyright Office [53]. We also asked participants to think about the plots and settings for two short stories, but asked them not to start writing in advance of the research study session. Each session started with participants signing the consent form and a brief introduction about the goal of the study from the study administrator. After that, we introduced the first study condition (tool) with a brief demo. We encouraged participants to ask questions at this stage and then to explore different features of the tool using a training story.

Participants then started the first writing session (20 minutes) in which they were asked to write a short story. We clarified to the participants that they did not need to finish the full story; rather this is a timed experience.

At the end of the first writing session, participants filled out a survey to provide their subjective experience. Participants then started the second writing session where they were asked to write their second story using the other interface, following the same procedure as for the first condition. We followed this with a second survey and then concluded the study with a semi-structured interview. During the interviews, participants shared their experience with both the baseline and the tool conditions and discussed their use of the LLM in the writing process.

5.7 Analysis Plan

Similar to our formative analysis (section 3), two co-authors independently open-coded the anonymized post-study interview transcripts and then conducted a thematic analysis. The coders met

regularly to discuss and refine the codes and themes. The coders also discussed the codes and themes with the entire research team. The initial inter-rater agreement was 0.86 (Jaccard's similarity).

For quantitative measures and subjective ratings, we used numerical estimation methods to derive 95% confidence intervals (CIs) for all measures [16]; non-parametric bootstrapping with $R = 1,000$ iterations. We also report the standardized effect size (Cohen's d).

5.8 Results

5.8.1 RQ1: Interaction with LLM. We found that HaLLMark significantly changed the writers' interaction with the LLM compared to the baseline. In the post-study interviews, participants mentioned that HaLLMark instilled a sense of awareness and encouraged them to actively evaluate AI-assistance from the beginning of the process (P1-4, P7, P10-13). For example, P2 and P8 said,

> *"I liked [HaLLMark] better because I was trying to use the AI without overusing it, and there were times when I felt like I was [doing that]. But then it said, 'Oh, you know, 90 or 95% of this writing is yours,' so, you know, more than I thought. So that was nice to have, and I liked having that information all the time."* (P2)

> *"I was keeping an eye on the text highlighted by yellow color and the percentage of that in the pie chart. It certainly made me conscious and encouraged me to modify text generated by the machine."* (P8)

On the flip side, some participants mentioned that it is possible that the tool may make some writers nervous and overly conscious about overusing the LLM, particularly in the eyes of readers and publishers, or even other writers (P5-6). This stigma can hamper their creative process.

We found evidence of the impact of HaLLMark in the percentage of AI-written text in the final stories. As defined in Section 4.2.3, we consider text directly generated by GPT-4 as AI-written and exclude text that was generated by GPT-4 but later re-written by the writers. We also measured text marked as AI-influenced by the writers. Figure 6A shows the percentage of text written and influenced by the AI in the final document for the two conditions. On average, the stories contained 13.66% (CI = [6.30, 19.76]) text written by the AI when participants used the baseline. In comparison, the stories contained only 3.48% (CI = [1.23, 5.26]) text written by AI when participants used HaLLMark. Additionally, participants labeled 2.99% (CI = [0.00, 4.77]) of the total text as AI-influenced in the stories written using HaLLMark.

However, we did not observe any difference in the number of prompts used in the two conditions. Regardless of the condition, participants preferred asking the AI to generate new content. Figure 6B shows the number of prompts used by participants in the two conditions. On average, participants used 2.65 (CI = [1.50, 3.80]) prompts seeking editorial help with the baseline. Participants used a similar amount of editorial prompts (2.86 with CI = [1.80, 3.89]) while using HaLLMark. The small effect size of 0.04 (Cohen's d) indicates no practical difference between the conditions.

Participants used prompts seeking generation more frequently than editorial prompts. On average, participants asked 5.66 (CI = [4.00, 6.76]) prompts seeking new content while using the baseline. Participants used a similar amount of prompts seeking new contents (6.29 with CI = [5.10, 7.37]) while using HaLLMark. The small effect size of 0.1 (Cohen's d) indicates a very small practical difference between the two conditions.

5.8.2 RQ2: Agency and Ownership. We found evidence that the situational awareness provided by HaLLMark improved writers' control over the process. As a result, writers were able to measure their contribution better when using HaLLMark. P7's comment below summarizes their experience,

> *"[HaLLMark] made me feel less confused in a way, even to myself, like what did I generate? What did the AI generate? What was influenced by the AI? It felt easy to apply the green highlighting for what was influenced, and the fact that it just automatically applying the orange highlighting for what the AI had generated felt pretty seamless. And it gave me sort of this feeling of reassurance and control that I did not find in the [baseline] interface."* (P7)

Using HaLLMark, some participants were able to perceive AI as a collaborator, rather than as an external agent (P3, P9, P13). For example, P13 said,

> *"With all the information showing my work and AI's work, it felt less robotic and more like I was collaborating with someone."* (P3)

Of course, there is danger inherent with anthropomorphizing AI [15, 60]; AI models are not persons and thus cannot be authors in the true sense, and there are legal, safety, security, trust, and reliability concerns in such relationships [20, 60].

The overall positive experience was reflected in the subjective *ownership* ratings provided by the participants (Figure 7A). On average, the rating for the baseline was 3.39 (CI = [2.00, 4.76]), and for HaLLMark, the rating was 4.92 (CI = [4.15, 5.69]). An effect size of 0.46 (Cohen's d) indicates a medium effect of the study condition.

5.8.3 RQ3: Communication and Transparency. Most participants preferred HaLLMark to communicate the extent of AI contributions in the final artifact. According to P1,

> *"I do not even see how I can use the first one [baseline] for communicating. [With HaLLMark], you can literally copy the text with colors and send it to someone in seconds. You can send the pie chart and the rectangles for more breakdowns."* (P1)

However, two participants had reservations against using HaLLMark for communication. P8 was worried that people might "nitpick" their writing if it was completely transparent and that readers would harshly criticize the use of AI. P4 preferred the timeline and text highlighting for communicating AI contributions, but did not want to share the summary statistics. They felt that readers might reduce their work to a single number (e.g., only 80% of the text was written by the author). The timeline would presumably show their contribution more clearly.

Figure 7B shows the participants ratings for how useful HaLLMark they felt it is for communication and transparency. On average, the rating for the baseline was 3.14 (CI = [1.85, 4.69]). The average rating for HaLLMark was 6.31 with CI = [5.61, 6.92]. The

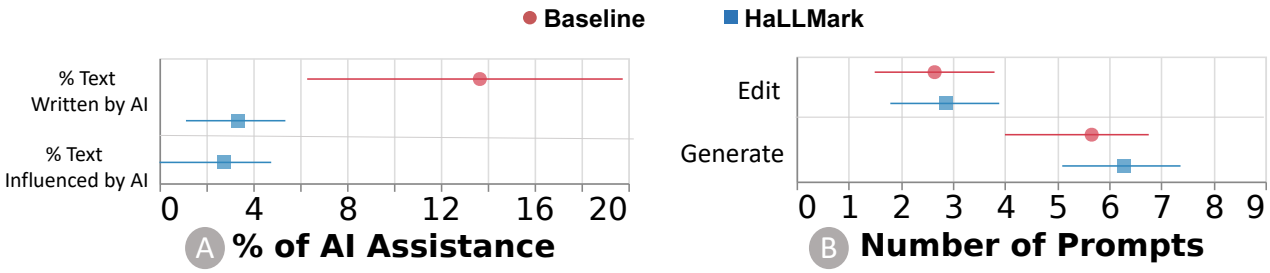

Figure 6: Percentage of AI assistance and number of prompts and while using the baseline tool and HaLLMark. Error bars show 95% confidence intervals (CIs). The baseline condition did not have the option to label text as AI-influenced. Thus, we see only one mark for that category in Figure A.

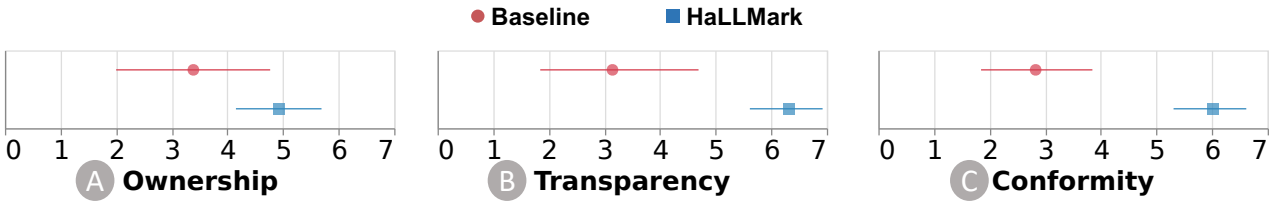

Figure 7: Self-reported subjective ratings. Collected for ownership, transparency, and conformity of AI policies.

standardized effect size (d = 2.27) shows a very large effect of the study condition.

5.8.4 RQ4: Conformity to AI-assisted Writing. All participants preferred HaLLMark for evaluating conformity to AI-assisted writing policies. For example, P5 said,

> *"When I read the policy before the study, I was like 'Ooh! this will be such a pain in the [posterior[1]].' But, then when I used the tool, I was like, 'Okay, this is easy!' I would totally use the second tool [HaLLMark] if I needed to follow a policy like this."* (P5)

We noticed a large difference in the subjective rating for this dimension for the two conditions (Figure 7C). On average, the subjective rating for the baseline was 2.83 (CI = [1.85, 3.85]). The average rating for HaLLMark was 6.00 with CI = [5.31, 6.61]. The standardized effect size (d = 2.10) indicates a very large effect of the study condition.

5.8.5 Subjective Perception of LLM Support. Figure 8 shows participants' subjective ratings on the usability of LLM support [41, 44]. Similar to previous research [41], we did not observe any significant difference in these dimensions.

5.8.6 Cognitive Load and Usability. Overall, participants found HaLLMark to be easy to use. Prompting LLMs is a relatively new activity for writers. HaLLMark added an extra task on top of prompting: tracking and verifying provenance information. However, participants did not report any excessive cognitive load due to this task in the post-study interviews. We believe there are three reasons for that. First, the general feedback from all participants indicates that verifying provenance information is a real need for writers who want to use LLMs and writers do not see this as an extra task. Second, several participants appreciated the use of visualization to seamlessly integrate the tracking task in HaLLMark. Participants used words such as "easy-to-understand", "simple", and "cool" to describe the visualizations and interaction. Finally, participants found the "distraction-free" mode useful to focus on specific tasks. We noticed several participants turned on and off the three modules of HaLLMark in different combinations to switch between writing, prompting, and validating provenance information. For example, P6 turned off the GPT-4 and visualization modules whenever they were writing, turned on the GPT-4 module for prompting the LLM, and then turned on the visualization module for seeing the summary statistics and prompt history.

Participants also suggested several improvements to HaLLMark. P3 and P6 suggested adding an option to turn on and off the text highlighting, as they can become distracting to writers for long-term use. In the current implementation, the text highlighting is always on. P9 wondered if they could add notes to the prompt or the text editor directly. This might be useful for providing an explanation if needed. P1 asked for more control over prompt generation, as for instance controlling the randomness of the text generation.

6 DISCUSSION, LIMITATIONS, AND FUTURE WORK

Our findings show that capturing and externalizing provenance information have a significant impact on how writers interact with LLMs as well as on their agency and control over the process. Our writer participants used HaLLMark to easily track the AI's contribution with respect to their own contribution. The tool helped writers maintain a level of AI contribution that they were comfortable with. This provided a sense of control in the writers' minds

[1] *Equus asinus.*

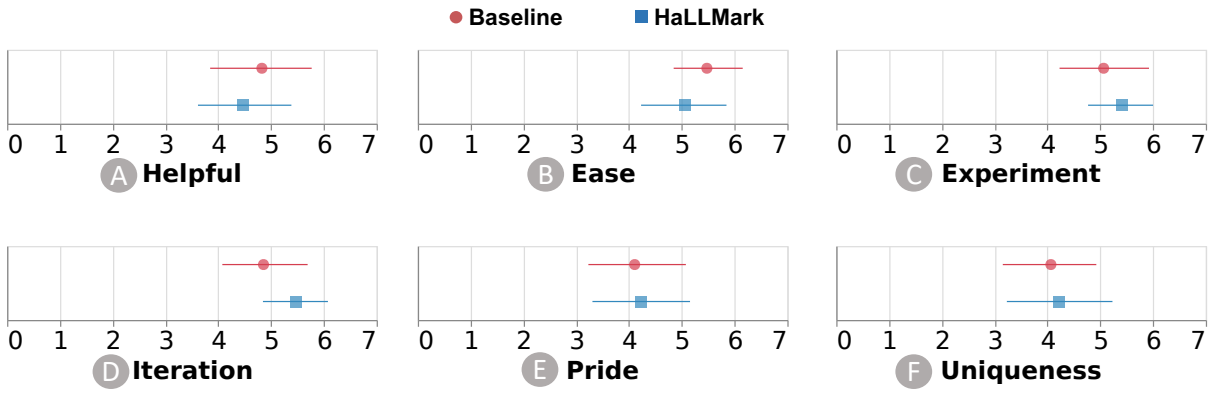

Figure 8: Subjective perception of LLM support. LLM support across six dimensions [41, 44].

and improved their ownership of the final artifact. Feedback from our participants indicate that HaLLMark can help them become more transparent about the co-writing process and conform to AI-assisted writing policy without manually preparing disclosures. Below we discuss the broader impacts of our work on interactive and intelligent writing support tools.

6.1 Normalize Transparency and Accountability in AI-assisted Writing

Interactive provenance information helped writers make informed decisions while producing their short stories. While writers were generally enthusiastic about sharing the provenance information with readers, some writers were not comfortable reporting the AI-generated parts publicly. They thought that disclosing the use of LLMs might make them susceptible to criticism. The fear of criticism is understandable, as it might diminish the perceived contribution of the human writer and presumably lead to disapproval of their creativity [62]. We note that such concerns are not unique to the use of intelligent tools—similar taboos exist for the use of reference materials for inspiration [30].

We believe the right way to remove the stigma around AI-assisted writing is by encouraging writers to be more transparent and accountable, and democratize tools to support this goal [34, 48]. However, this also requires that readers, publishers, and other writers become charitable and open-minded about LLMs going forward.

6.2 Design Implications

6.2.1 Writers Want to Use LLMs for Content Generation, not Editing. We found that participants in the study were mostly interested in generating new content or ideas (not the whole story) using GPT-4. Although existing policies indirectly encourage writers to use the LLM for editorial purposes [21], writers did not find that useful during the study. Rather, they were intrigued by its generation power and wanted to use it for overcoming challenges such as writer's block, difficulties in expressing nuanced and expressive details about a new scene, or taking narrative inspiration. While this result is in line with several previous studies [25, 51], it contrasts with a survey that found that 60% of the surveyed writers want to use LLMs for editorial purposes [4].

One explanation behind this contrasting result could be that we collect writers' feedback based on the experience with tangible interfaces whereas the survey depended on writers' preconceived perceptions of LLMs. It is also possible that writers' perceptions have changed since the time of the survey (May 2023) due to the introduction of policies on AI-assisted writing. Another caveat here is that the writing sessions in our study were short (20 mins) and likely do not fully capture how writers might use tools such as HaLLMark for long-form writing (e.g., fiction). Nevertheless, we believe that, in the future, organizations and authorities will likely benefit by focusing on devising policies for ethical and responsible content generation, rather than limiting writers to the use of LLMs for editorial purposes.

6.2.2 When to Write, When to Prompt, and When to Verify? Our findings indicate that when and how writers want to switch between writing, prompting, and tracking provenance depends on writers' personal style, needs, and idiosyncrasies. During the study, we noticed several participants turned on and off the three modules in different combinations. This observation indicates that future AI-assisted writing tools will benefit from allowing writers to freely switch between writing and AI-related tasks. In the future, we aim to conduct a longitudinal study to comprehensively understand how the tracking task impacts writing and when and how writers want to verify provenance information. Recent studies on understanding writers' needs for AI support are inspiring in this scenario [26, 51].

6.2.3 Adopting HaLLMark in Writing Tools. HaLLMark is currently a standalone writing tool. However, we aim to make our tool open-sourced (currently available in the OSF repository), allowing others to build upon it or modify it to their needs. For example, with appropriate modification to our codebase, HaLLMark can be turned into a plugin. Writers can then install HaLLMark in their favorite writing tool and track provenance information. Further, if implemented as a desktop application, HaLLMark can track provenance information across multiple tools (e.g., Microsoft Word, Overleaf, and Google Docs). Alternatively, researchers and organizations can take inspiration from the design of HaLLMark and decide to build their own tracking interface. For example, Microsoft Word already integrates LLMs into its writing interface. Designing a tracking interface should be relatively straightforward. Finally,

many visualization systems now store data using structured languages (e.g., Vega-Lite [59]). One way to make the task of tracking provenance information platform-independent is to focus on storing provenance information in a JSON format and then allowing copy and paste actions for the data across tools and devices.

6.2.4 HaLLMark as a Reading Tool. HaLLMark is primarily focused on helping writers to ensure transparency and accountability in their creative content. However, we believe that the provenance information collected by our approach could be shared in several ways with the recipients of written artifacts. For example, since HaLLMark is a web-based tool, writers could share a URL to the interactive document produced from HaLLMark as a supplement to the peer reviewers or publishers, who can then verify whether the artifact conforms to their writing policies using HaLLMark. Alternatively, HaLLMark could have a new feature that would create a static report of AI assistance (text highlighting on the artifact, summary statistics, and a list of prompts with the static timeline as an overview). Writers could then share the report as a supplement to the actual artifact with the publishers or peer-reviewers. Finally, if our approach is able to store the data in a structured format (as discussed above), writers can store the provenance information and then share it with readers, who can then use their favorite reading tool to verify the contribution from the AI. In all cases, publishers can decide how they want to share the provenance information with the larger audience.

6.3 Limitations

HaLLMark is so far only a technology probe and is not designed for production use. This means that it is limited in terms of the scale and scope of the documents and writing tasks it can produce. While we did not perform any specific stress or performance testing in our evaluation, we believe that the current HaLLMark prototype implementation easily scales to a few thousand words. While not evaluated, in theory, the visualizations in HaLLMark are also scalable to any number of prompts. As mentioned in Section 4.2.3, there is a minimum threshold for the rectangle width (5px). When the rectangle width reaches the minimum, we dynamically increase the width of the SVG instead of decreasing the rectangle width and provide a horizontal scrollbar to see the extended contents. Prior visualization systems have adopted similar methods to make representations scalable [31]. Similarly, as policies evolve, we will likely see more constraints and dimensions, which could be encoded in the timeline by adding new rows in the timeline. Nevertheless, we acknowledge that as stories become longer and more variables are added to the timeline, it might become a daunting task to make sense of the timeline. One way to scale the representation is to aggregate the rectangles in the timeline, a popular approach in visualization design [19, 31]. Our future work will explore these solutions. We have also discussed some practical manifestations of the tool in Section 6.2; for example, a practical use case would be to implement the tool as a plugin that can be integrated with existing writing software, such as Grammarly.

Our study in this paper was limited to creative writers, and their experiences may differ from the general population of all writers. For example, compared to a fiction author, an academic or a journalist must rely on verifiable facts and evidence. This may alter the dynamic for such writers when using an LLM, as LLMs are still notorious for generating hallucinations [38]. Further study for these settings are beyond the scope of this paper.

6.4 Ethical Concerns of LLM-based Co-Writing

It could well be argued that the central argument of this paper—that LLMs are here to stay, and that we should just learn how to best leverage them—is a technopositive, naïve, and perhaps even actively harmful approach to the use of AI in human creativity, and that generative AI should be seen as dangerous technology that should be regulated or even banned. However, we would argue that this is true of virtually any technology. For example, photography was widely hailed as the end of painting but instead freed painters from the curse of realism [20]. Instead, by harnessing these technologies as supertools [60] in support of and subservient to—and not partners or collaborators with—human writers is precisely the approach that we should be taking.

In the end, LLMs are just tools, even if they are highly sophisticated ones. By focusing on conveying the provenance between LLMs and humans, essentially making human verification and influence the gold standard, we can reinforce this notion [42, 47, 60]. After all, while many of us would view the idea of spending hours reading stories that were generated by a soulless machine somewhat insulting, most would likely accept this when assured that the overarching control of the story belonged to an actual human writer. People already accept computer-generated imagery (CGI) in today's movies as a matter of course—why would they not accept similar computer-generated prose, as long as it has been verified (and potentially edited) by the author? The provenance mechanisms presented in this paper, where these prompts, edits, and influences are made explicit in the text itself, is one approach to conveying this interaction history between the writer and the LLM. Cryptology concepts such as NFTs [6]—or placing the entire edit history on a blockchain—may be used to protect the integrity of this history.

Findings from our evaluation clearly indicate that our writer participants are mainly interested in using AI to improve their own writing rather than producing more copy faster. This may not strictly be true of students in educational settings, where LLMs could be argued to do more for the aspiring writer than act as the equivalent of a mere calculator for mathematics education. For example, one sentiment that was expressed by our participants, in line with previous studies [25, 51], is that writer's block may now be a phenomenon of the past, as the AI can always be relied upon to generate many new and fresh ideas of how to continue a story. While we should always be wary of bad (or overworked and stressed) actors that are indeed primarily seeking the ability to generate acceptable copy with a minimum of effort, professional writers harbor pride in the craft of writing [26], as is true among virtually all professionals.

Naturally, there are other ethical considerations that we must consider when putting this technology into the hands of writers. For one thing, it is possible that in spite of the tool's design to support author agency (as evidenced by the ability to writer's ability to edit or modify transaction history), other actors in the publication industry might be compelled to use the tool to surveil AI use and

enforce AI writing policies. Such has been the case for some academic writing support tools, such as Turnitin, which is designed to empower student learners, but has drawn criticism for its potential to police rather than support students. Additionally, given the 2023 strike between the WGA and SAG-AFTRA on the one side, and the Alliance of Motion Picture and Television Producers (AMPTP) on the other, we should ensure not crossing any picket lines by actively making these tools freely available on the internet. In the case of HaLLMark, while we anticipate releasing the tool as open source on Github upon acceptance of this paper, we will add an explicit statement of support for WGA and SAG-AFTRA on the tool website as well as include licensing terms prohibiting the use of the tool to cross the picket line.

7 CONCLUSION

We have presented a technology probe on AI co-writing called HaLLMark that enables an effective form of Large Language Model prompting while storing the provenance of interaction between human writer and AI. Designed based on our review of generative AI guidelines by professional and research organizations, HaLLMark transparently stores the prompting and influences between the LLM and the writer using text highlighting and a visual timeline. We have presented our findings from a qualitative study involving a group of writers using HaLLMark to write a short story or non-fiction article. We found that our writers valued the explicit representation of the AI's influence on their work, but also that the prompting interface yields a smoother and more integrated workflow than the default ChatGPT interface.

Human-AI co-writing is a nascent area of research that is also fraught with controversy. Our work addresses both transparency and prompting for LLMs supporting this modality, but is by no means the final nor optimal approach to either of these open research problems. We hope that future work can build on our findings to derive better supertools that retain human agency and control of the output while leveraging the formidable power of modern foundation AI models. In particular, we think that future research should focus on scaffolding prompts, improving provenance tracking, and adding non-repudiation of textual outputs generated by both human writers and AI models.

ACKNOWLEDGMENTS

While this work deals with AI co-writing, none of it was written using an AI model such as ChatGPT or GPT-4. In other words, HaLLMark was actually not used in producing the copy in this paper. This work was partly supported by grant IIS-2211628 from the U.S. National Science Foundation and Villum Investigator grant VL-54492 by Villum Fonden. Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the funding agency.

REFERENCES

[1] Aretha B. Alencar, Maria Cristina F. de Oliveira, and Fernando V. Paulovich. 2012. Seeing beyond reading: a survey on visual text analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2, 6 (2012), 476–492.

[2] Saleema Amershi, Max Chickering, Steven M. Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. ModelTracker: Redesigning Performance Analysis Tools for Machine Learning. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 337–346. https://doi.org/10.1145/2702123.2702509

[3] Modern Language Association, Conference on College Composition, and Communication. 2023. MLA-CCCC Joint Task Force on Writing and AI Working Paper: Overview of the Issues, Statement of Principles, and Recommendations. https://hcommons.org/app/uploads/sites/1003160/2023/07/MLA-CCCC-Joint-Task-Force-on-Writing-and-AI-Working-Paper-1.pdf Accessed: 2023-08-04.

[4] Author's Guild. 2023. Survey Reveals 90 Percent of Writers Believe Authors Should Be Compensated for the Use of Their Books in Training Generative AI. https://authorsguild.org/news/ai-survey-90-percent-of-writers-believe-authors-should-be-compensated-for-ai-training-use/.

[5] Sriram Karthik Badam, Zhicheng Liu, and Niklas Elmqvist. 2019. Elastic Documents: Coupling Text and Tables through Contextual Visualizations for Enhanced Document Reading. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 661–671. https://doi.org/10.1109/TVCG.2018.2865119

[6] Seyed Mojtaba Hosseini Bamakan, Nasim Nezhadsistani, Omid Bodaghi, and Qiang Qu. 2022. Patents and intellectual property assets as non-fungible tokens; key technologies and challenges. *Scientific Reports* 12, 1, Article 2178 (2022), 13 pages. https://doi.org/10.1038/s41598-022-05920-6

[7] Jeremy Birnholtz and Steven Ibara. 2012. Tracking Changes in Collaborative Writing: Edits, Visibility and Group Maintenance. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. ACM, New York, NY, USA, 809–818. https://doi.org/10.1145/2145204.2145325

[8] Richard Brath. 2021. *Visualizing with Text.* CRC Press, Boca Raton, FL, USA.

[9] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[10] Fanny Chevalier, Pierre Dragicevic, Anastasia Bezerianos, and Jean-Daniel Fekete. 2010. Using text animated transitions to support navigation in document histories. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 683–692. https://doi.org/10.1145/1753326.1753427

[11] Jason Chuang, Christopher D. Manning, and Jeffrey Heer. 2012. Termite: visualization techniques for assessing textual topic models. In *Proceedings of the ACM Conference on Advanced Visual Interfaces*. ACM, New York, NY, USA, 74–77. https://doi.org/10.1145/2254556.2254572

[12] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching Stories with Generative Pretrained Language Models. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 209:1–209:19. https://doi.org/10.1145/3491102.3501819

[13] Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai Gao, Huamin Qu, and Xin Tong. 2011. TextFlow: Towards Better Understanding of Evolving Topics in Text. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2412–2421. https://doi.org/10.1109/TVCG.2011.239

[14] Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. 2022. Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries. In *Proceedings of the ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, USA, 98:1–98:13. https://doi.org/10.1145/3526113.3545672

[15] Ameet Deshpande, Tanmay Rajpurohit, Karthik Narasimhan, and Ashwin Kalyan. 2023. Anthropomorphization of AI: Opportunities and Risks. *CoRR* abs/2305.14784 (2023), 7 pages. https://doi.org/10.48550/arXiv.2305.14784 arXiv:2305.14784

[16] Pierre Dragicevic. 2016. Fair Statistical Communication in HCI. In *Modern Statistical Methods for HCI*. Springer Publishing, New York, NY, USA, 291–330.

[17] Wanyu Du, Zae Myung Kim, Vipul Raheja, Dhruv Kumar, and Dongyeop Kang. 2022. Read, Revise, Repeat: A System Demonstration for Human-in-the-loop Iterative Text Revision. In *Proceedings of the Workshop on Intelligent and Interactive Writing Assistants*. Association for Computational Linguistics, Stroudsburg, PA, USA, 96–108. https://doi.org/10.18653/v1/2022.in2writing-1.14

[18] Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. 2022. Understanding Iterative Revision from Human-Written Text. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, 3573–3590. https://doi.org/10.18653/v1/2022.acl-long.250

[19] Niklas Elmqvist and Jean-Daniel Fekete. 2010. Hierarchical Aggregation for Information Visualization: Overview, Techniques, and Design Guidelines. *IEEE Transactions on Visualization and Computer Graphics* 16, 3 (2010), 439–454. https://doi.org/10.1109/TVCG.2009.84

[20] Ziv Epstein, Aaron Hertzmann, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R. Frank, Matthew Groh, Laura Herman, Neil Leach, Robert Mahari, Alex "Sandy" Pentland, Olga Russakovsky, Hope Schroeder, and Amy Smith and. 2023. Art and the science of generative AI. *Science* 380, 6650 (June 2023), 1110–1111. https://doi.org/10.1126/science.adh4451

[21] Association for Computational Linguistics. 2023. ACL 2023 Policy on AI Writing Assistance. https://2023.aclweb.org/blog/ACL-2023-policy/ Accessed: 2023-08-04.

[22] Association for Computing Machinery. 2023. ACM Policy on Authorship. https://www.acm.org/publications/policies/new-acm-policy-on-authorship Accessed:

2023-08-04.

[23] Samah Gad, Waqas Javed, Sohaib Ghani, Niklas Elmqvist, E. Thomas Ewing, Keith N. Hampton, and Naren Ramakrishnan. 2015. ThemeDelta: Dynamic Segmentations over Temporal Topic Models. *IEEE Transactions on Visualization and Computer Graphics* 21, 5 (2015), 672–685. https://doi.org/10.1109/TVCG.2014.2388208

[24] Katy Ilonka Gero and Lydia B. Chilton. 2019. Metaphoria: An Algorithmic Companion for Metaphor Creation. In *Proceedings of the ACM Conference on Designing Interactive Systems*. ACM, New York, NY, USA, 296. https://doi.org/10.1145/3290605.3300526

[25] Katy Ilonka Gero, Vivian Liu, and Lydia B. Chilton. 2022. Sparks: Inspiration for Science Writing using Language Models. In *Proceedings of the ACM Conference on Designing Interactive Systems*. ACM, New York, NY, USA, 1002–1019. https://doi.org/10.1145/3532106.3533533

[26] Katy Ilonka Gero, Tao Long, and Lydia B. Chilton. 2023. Social Dynamics of AI Support in Creative Writing. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 245:1–245:15. https://doi.org/10.1145/3544548.3580782

[27] Granthika. 2021. Granthika Writing Tool. https://granthika.co Accessed: 2022-02-04.

[28] The Author's Guild. 2023. AG Introduces New Publishing Agreement Clauses Concerning AI. https://authorsguild.org/news/ag-introduces-new-publishing-agreement-clauses-concerning-ai/ Accessed: 2023-08-04.

[29] Fred Hohman, Kanit Wongsuphasawat, Mary Beth Kery, and Kayur Patel. 2020. Understanding and Visualizing Data Iteration in Machine Learning. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376177

[30] Josh Holinaty, Alec Jacobson, and Fanny Chevalier. 2021. Supporting reference imagery for digital drawing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE Computer Society, Los Alamitos, CA, USA, 2434–2442.

[31] Md Naimul Hoque and Niklas Elmqvist. 2024. Dataopsy: Scalable and Fluid Visual Exploration using Aggregate Query Sculpting. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (2024), 1–11. https://doi.org/10.1109/TVCG.2023.3326594

[32] Md Naimul Hoque, Bhavya Ghai, and Niklas Elmqvist. 2022. DramatVis Personae: Visual Text Analytics for Identifying Social Biases in Creative Writing. In *Proceedings of the ACM Conference on Designing Interactive Systems*. ACM, New York, NY, USA, 1260–1276. https://doi.org/10.1145/3532106.3533526

[33] Md Naimul Hoque, Bhavya Ghai, Kari Kraus, and Niklas Elmqvist. 2023. Portrayal: Leveraging NLP and Visualization for Analyzing Fictional Characters. In *Proceedings of the ACM Conference on Designing Interactive Systems*. ACM, New York, NY, USA, 74–94. https://doi.org/10.1145/3563657.3596000

[34] Mohammad Hosseini, David B Resnik, and Kristi Holmes. 2023. The ethics of disclosing the use of artificial intelligence tools in writing scholarly manuscripts. *Research Ethics* 19, 4 (2023), 449–465. https://doi.org/10.1177/17470161231180449

[35] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology Probes: Inspiring Design for and with Families. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 17–24. https://doi.org/10.1145/642611.642616

[36] Stefan Jänicke, Greta Franzini, Muhammad Faisal Cheema, and Gerik Scheuermann. 2015. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. In *State of the Art Reports of the Eurographics Conference on Visualization*. Eurographics Association, Geneva, Switzerland, 83–103. https://doi.org/10.2312/eurovisstar.20151113

[37] Waqas Javed and Niklas Elmqvist. 2013. ExPlates: Spatializing Interactive Analysis to Scaffold Visual Exploration. *Computer Graphics Forum* 32, 3 (2013), 441–450. https://doi.org/10.1111/CGF.12131

[38] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38. https://doi.org/10.1145/3571730

[39] Peiling Jiang, Jude Rayan, Steven P Dow, and Haijun Xia. 2023. Graphologue: Exploring Large Language Model Responses with Interactive Diagrams. In *Proceedings of the ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, USA, 3:1–3:20. https://doi.org/10.1145/3586183.3606737

[40] Jeongyeon Kim, Sangho Suh, Lydia B. Chilton, and Haijun Xia. 2023. Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing. In *Proceedings of the ACM Conference on Designing Interactive Systems*. ACM, New York, NY, USA, 115–135. https://doi.org/10.1145/3563657.3595996

[41] Tae Soo Kim, Yoonjoo Lee, Minsuk Chang, and Juho Kim. 2023. Cells, Generators, and Lenses: Design Framework for Object-Oriented Interaction with Large Language Models. In *Proceedings of the ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, USA, 4:1–4:18.

[42] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1369–1385. https://doi.org/10.1145/3593013.3594087

[43] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI collaborative writing dataset for exploring language model capabilities. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 388:1–388:19. https://doi.org/10.1145/3491102.3502030

[44] Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E. Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael S. Bernstein, and Percy Liang. 2022. Evaluating Human-Language Model Interaction. *CoRR* abs/2212.09746 (2022), 64 pages. https://doi.org/10.48550/arXiv.2212.09746 arXiv:2212.09746

[45] Tak Yeon Lee, Alison Smith, Kevin D. Seppi, Niklas Elmqvist, Jordan L. Boyd-Graber, and Leah Findlater. 2017. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies* 105 (2017), 28–42. https://doi.org/10.1016/j.ijhcs.2017.03.007

[46] Tianyi Li, Yasmine Belghith, Chris North, and Kurt Luther. 2020. CrowdTrace: Visualizing Provenance in Distributed Sensemaking. In *Proceedngs of the IEEE Visualization Conference*. IEEE, Piscataway, NJ, USA, 191–195. https://doi.org/10.1109/VIS47514.2020.00045

[47] Q. Vera Liao, Hariharan Subramonyam, Jennifer Wang, and Jennifer Wortman Vaughan. 2023. Designerly Understanding: Information Needs for Model Transparency to Support Design Ideation for AI-Powered User Experience. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, Article 9, 21 pages. https://doi.org/10.1145/3544548.3580652

[48] Q. Vera Liao and Jennifer Wortman Vaughan. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *CoRR* abs/2306.01941 (2023), 33 pages. https://doi.org/10.48550/ARXIV.2306.01941 arXiv:2306.01941

[49] Literature and Latte. 2021. Scrivener. https://www.literatureandlatte.com/scrivener/overview/ Accessed: 2022-02-04.

[50] Nina McCurdy, Julie Lein, Katherine Coles, and Miriah D. Meyer. 2016. Poemage: Visualizing the Sonic Topology of a Poem. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 439–448. https://doi.org/10.1109/TVCG.2015.2467811

[51] Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 355:1–355:34. https://doi.org/10.1145/3544548.3581225

[52] Springer Nature. 2023. Nature Portfolio. https://www.nature.com/nature-portfolio/editorial-policies/ai

[53] U.S. Copyright Office. 2023. Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence. https://copyright.gov/ai/ai_policy_guidance.pdf Accessed: 2023-08-04.

[54] OpenAI. 2023. New AI classifier for indicating AI-written text. https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text Accessed: 2023-08-04.

[55] Davis Creative Publishing Partners. 2023. DCPP Artificial Intelligence (AI) Policy. https://creativepublishingpartners.com/ai-policy/ Accessed: 2023-08-04.

[56] QuillJS. 2021. QuillJS: a Rich Text Editor. https://quilljs.com Accessed: 2022-02-04.

[57] Eric D. Ragan, Alex Endert, Jibonananda Sanyal, and Jian Chen. 2016. Characterizing Provenance in Visualization and Data Analysis: An Organizational Framework of Provenance Types and Purposes. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 31–40. https://doi.org/10.1109/TVCG.2015.2467551

[58] Jen Rogers and Anamaria Crisan. 2023. Tracing and Visualizing Human-ML/AI Collaborative Processes through Artifacts of Data Work. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 837:1–837:22. https://doi.org/10.1145/3544548.3580819

[59] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. 2017. Vega-Lite: A Grammar of Interactive Graphics. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 341–350. https://doi.org/10.1109/TVCG.2016.2599030

[60] Ben Shneiderman. 2022. *Human-Centered AI*. Oxford University Press, Oxford, United Kingdom.

[61] Sarah Sterman, Evey Huang, Vivian Liu, and Eric Paulos. 2020. Interacting with Literary Style through Computational Tools. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376730

[62] Chris Stokel-Walker. 2022. AI bot ChatGPT writes smart essays - should professors worry? *Nature: News Explainer* (Dec. 2022). https://doi.org/10.1038/d41586-022-04397-7

[63] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. In *Proceedings of the ACM Symposium on User Interface Software and Technology*.

ACM, New York, NY, USA, 1:1–1:18. https://doi.org/10.1145/3586183.3606756

[64] Johnny Torres, Sixto García, and Enrique Peláez. 2019. Visualizing Authorship and Contribution of Collaborative Writing in E-Learning Environments. In *Proceedings of the ACM Conference on Intelligent User Interfaces*. ACM, New York, NY, USA, 324–328. https://doi.org/10.1145/3301275.3302328

[65] Frank van Ham, Martin Wattenberg, and Fernanda B. Viégas. 2009. Mapping Text with Phrase Nets. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 1169–1176. https://doi.org/10.1109/TVCG.2009.165

[66] Mitchell Vásquez-Bermúdez, Cecilia Veronica Sanz, María Alejandra Zangara, and Jorge Hidalgo. 2021. Visualization Tools for Collaborative Systems: A Systematic Review. In *Proceedings of the International Conference on Technologies and Innovation (Communications in Computer and Information Science, Vol. 1460)*. Springer Publishing, New York, NY, USA, 107–122. https://doi.org/10.1007/978-3-030-88262-4_8

[67] Fernanda B. Viégas and Martin Wattenberg. 2008. Tag clouds and the case for vernacular visualization. *Interactions* 15, 4 (2008), 49–52. https://doi.org/10.1145/1374489.1374501

[68] Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. 2004. Studying cooperation and conflict between authors with *history flow* visualizations. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 575–582. https://doi.org/10.1145/985692.985765

[69] Fernanda B. Viégas, Martin Wattenberg, and Jonathan Feinberg. 2009. Participatory Visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 1137–1144. https://doi.org/10.1109/TVCG.2009.171

[70] Dakuo Wang, Judith S. Olson, Jingwen Zhang, Trung Nguyen, and Gary M. Olson. 2015. DocuViz: Visualizing Collaborative Writing. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1865–1874. https://doi.org/10.1145/2702123.2702517

[71] Martin Wattenberg and Fernanda B. Viégas. 2008. The Word Tree, an Interactive Visual Concordance. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (2008), 1221–1228. https://doi.org/10.1109/TVCG.2008.172

[72] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models. *CoRR* abs/2112.04359 (2021), 64 pages. arXiv:2112.04359

[73] Kanit Wongsuphasawat, Daniel Smilkov, James Wexler, Jimbo Wilson, Dandelion Mané, Doug Fritz, Dilip Krishnan, Fernanda B. Viégas, and Martin Wattenberg. 2018. Visualizing Dataflow Graphs of Deep Learning Models in TensorFlow. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 1–12. https://doi.org/10.1109/TVCG.2017.2744878

[74] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 385:1–385:22. https://doi.org/10.1145/3491102.3517582

[75] Soobin Yim, Dakuo Wang, Judith Olson, Viet Vu, and Mark Warschauer. 2017. Synchronous Collaborative Writing in the Classroom: Undergraduates' Collaboration Practices and Their Impact on Writing Style, Quality, and Quantity. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, New York, NY, USA, 468–479. https://doi.org/10.1145/2998181.2998356

[76] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *Proceedings of the ACM Conference on Intelligent User Interfaces*. ACM, New York, NY, USA, 841–852. https://doi.org/10.1145/3490099.3511105