Dimensions for Designing LLM-based Writing Support

Frederic Gmeiner
HCI Institute
Carnegie Mellon University
Pittsburgh, PA, USA
gmeiner@cmu.edu

Nur Yildirim
HCI Institute
Carnegie Mellon University
Pittsburgh, PA, USA
yildirim@cmu.edu

1 INTRODUCTION

Advances in large language models (LLMs) have enabled a myriad of unprecedented capabilities over the past few years: computers can write code, translate between languages, and generate conversations. Among the potential use cases, *writing* has been a domain that continually fascinated researchers. Nowadays, as LLMs move from research labs to the real world, there is a growing interest in exploring the capabilities of these systems to support writing tasks across domains such as fiction writing [3, 11], scientific writing [7], poetry [4], and theatre scripts and screenplays [8].

As HCI researchers and designers, we have been "playing with" GPT to understand its capabilities as a *design material* to support writing. A common pitfall when designing AI experiences is the tendency to envision *"holy grail use cases"*: places that require near-perfect AI performance for delivering high-quality outputs that match human intelligence or creativity [5, 10]. For instance, our initial experiments mostly focused on prompting the model to come up with novel content: writing stories, podcast scripts, characters, and plots. Similar to others [8, 11], we were soon disappointed by how generic and bland the outputs were. However, over time we learned to lower our expectations and focus on low hanging fruit: less complex writing tasks where average quality LLM output could be useful. These included recommending word associations, listing things (e.g., places, names, objects) as inspiration, or simply reformatting writing (e.g., creating an outline based on prose).

In our experience, there are three key considerations when designing LLM experiences for writing support: *LLM capabilities, task complexity* and *output quality*. We propose that these dimensions are likely to complement common taxonomy dimensions coming from human-centered and socio-technical perspectives, such as feasibility, value co-creation, usability, etc. [2]. In this position paper, we argue that a taxonomy of writing assistants capturing these dimensions holistically could scaffold the process of designing experiences that writers find valuable. The remainder of this paper details each dimension and how these could inform the exploration of LLM's design space.

2 LLM CAPABILITIES AS DESIGN MATERIAL

Gaining an understanding of AI's capabilities and limitations is a major challenge for designers and end users who do not have a technical background [10]. One of our goals throughout our experiments with chatGPT was to gain a better understanding of what it can do, and what it can do *reasonably well*. At a high level, LLMs are "trained to predict the most likely next word given a textual prompt" [1]. However, this high-level description does not capture the wide range of things LLMs can do.

We found that explicating distinct LLM capabilities was a good starting place to understand LLMs as a design material. We first reviewed prior literature and looked for emerging UX patterns for prompting [3, 11]. We then curated a subset of example prompts demonstrating LLM capabilities, such as *text summarization, paraphrasing, dialog generation, elaboration, story seeding, sentence completion, rewriting in a tone, rewriting in a style*, etc. (Table 1). As we played around with each capability, we tried to assess the quality of outputs – however, this was dependent on the use case context. For example, we prompted chatGPT to *"write a rap battle between Harry Potter and Lord Voldemort"* or *"write the lyrics of a song where Rousseau and Voltaire argue on the nature of mankind"*. While these prompts resulted in reasonably well outputs (i.e., coherence and rhyme), we did not find them particularly useful. We wondered whether we could produce a podcast script instead, which could be immediately useful. However, this seemed to be a complex writing task. Our trials resulted in generic scripts that professional podcasters would not find useful. In search of a target user group, we thought the scripts could provide value for content creators on Youtube who review products. A draft script for product review based on a few bullet points might be better than having no script.

Table 1: Non-exhaustive list of LLM capabilities and example writing tasks where they might be useful.

LLM Capabilities	Writing Tasks
Text summarization	Reviewing, Reflection
Paraphrasing	Refining, reviewing
Elaboration	Detailing, scene setting
Dialog generation	Writing scripts, screenplay
Story seeding	Unblocking
Sentence completion	Detailing plots, dialog, etc.
Rewriting in a tone	Reviewing, characters' speech
Rewriting in a style	Reviewing, conveying setting, time, mood
Listing	Detailing places, characters, etc.
Formatting	Prose to outline
Keyword association	Inspiration, ideation

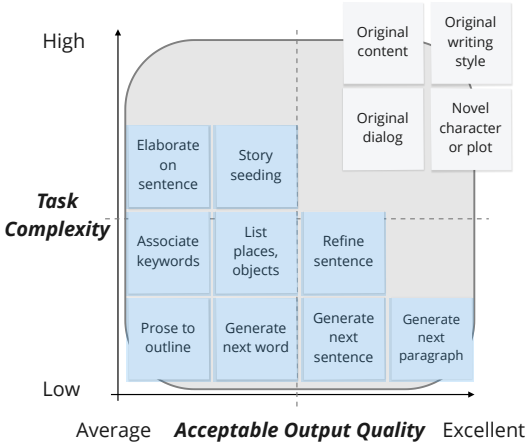
Through these experiments, we became aware of the interplay between *the task complexity* and *output quality* for a given LLM capability. In the next section, we detail how these dimensions are likely to impact the end user experience for writing support.

3 TASK COMPLEXITY AND OUTPUT QUALITY

Writing tasks vary vastly: some are trivial, some are more challenging. Some tasks are tedious, while others have a great impact on writers' enjoyment and ownership [11]. We asked two questions as we tried to gain a better sense of writing tasks:

(1) How complex is this writing task for a user to do? Does it require a high level of expertise or a deep contextual understanding?
(2) What is the level of quality for LLM outputs to be perceived as useful? Does the task require particular qualities to be acceptable? (e.g., factuality, consistency, etc.)

Figure 1: Task Complexity-Output Quality matrix for LLM capabilities. Our exploration revealed that writing tasks where average quality outputs are acceptable, provide a rich design space for LLM experiences.



We started to map the writing tasks we have been thinking about based on the task complexity and required output quality. Prior work delineated writing tasks based on the writing goals for specific writing processes (i.e., planning, translation, reviewing) and noted that each part of the writing process might impose different levels of constraints [6]. In our case, we focused on the quality of output as a key dimension that makes or breaks the usefulness of LLMs for writing support.

We realized that our initial explorations mostly focused on complex writing tasks that required high quality outputs: things such as generating original story plots, podcast content, characters, or styles. These tasks were part and parcel of the writing process where writer expertise and involvement were high (Figure 1). On the other hand, we found many writing tasks that were relatively less complex. Things such as asking for inspirational keywords, word or sentence completion, listing names or places, and rewriting sentences to be longer or more concise – these were places where vague, unrelated, or even inconsistent outputs could be useful. For example, we prompted chatGPT to play a word association game that could help us describe a mood and feel in a story. We found the listed words useful even though we did not necessarily use them, as they sensitized us to better think and reflect on our story.

We view the task complexity-output quality mapping as a valuable perspective to navigate LLM's design space. Can researchers, designers, and technologists think of low complexity writing tasks where average quality LLM outputs could be useful? What are the ways LLMs can support high complexity writing tasks without providing original, factual, or consistent outputs? Similar to others [1], we suspect that more contained and tedious writing tasks lend themselves better for LLM support and AI-assisted writing in general.

From a machine learning perspective, the classification of LLM capabilities within a complexity-output quality matrix could also serve as guidance for model developers for improving the support of specific writing tasks – for example, when part of a reinforcement learning from human feedback approach. Furthermore, for developing writing assistive tools that can adapt to users' expertise and context, a complexity-output quality matrix might help determine which capabilities best support different users' needs and expectations.

4 PROMPTS FOR WORKSHOP DISCUSSION

Below, we highlight two open questions as starting points for workshop discussion:

- ***What dimensions should a taxonomy of writing assistants capture?*** Our exploration focused on dimensions that are critical for finding use cases for LLM-based writing support. However, this is a partial perspective; there are many critical dimensions for designing LLM experiences, including factuality, bias, stereotyping, and homogenization. Recent work that proposed a taxonomy of LLM risks [9] could provide a starting point for further dimensions to consider.
- ***How to assess the success of LLM-based writing support?*** Evaluating writing assistants is an open research area. The workshop could expand on dimensions that could be used to assess LLM experiences.

REFERENCES

[1] 2022. Wordcraft Writers Workshop. https://g.co/research/wordcraft
[2] Robert P Bostrom and J Stephen Heinen. 1977. MIS problems and failures: A socio-technical perspective. Part I: The causes. *MIS quarterly* (1977), 17–32.
[3] Alex Calderwood, Vivian Qiu, Katy Ilonka Gero, and Lydia B Chilton. 2020. How Novelists Use Generative Language Models: An Exploratory User Study.. In *HAI-GEN+ user2agent@ IUI*.
[4] Tuhin Chakrabarty, Vishakh Padmakumar, and He He. 2022. Help me write a poem: Instruction Tuning as a Vehicle for Collaborative Poetry Writing. *arXiv preprint arXiv:2210.13669* (2022).
[5] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX design innovation: Challenges for working with machine learning as a design material. In *Proceedings of the 2017 chi conference on human factors in computing systems*. 278–288.
[6] Katy Gero, Alex Calderwood, Charlotte Li, and Lydia Chilton. 2022. A design space for writing support tools using a cognitive process model of writing. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*. 11–24.
[7] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for science writing using language models. In *Designing Interactive Systems Conference*. 1002–1019.
[8] Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. [n. d.]. Co-writing screenplays and theatre scripts alongside language models using Dramatron. ([n. d.]).
[9] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 214–229.
[10] 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*. 1–13.
[11] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *27th International Conference on Intelligent User Interfaces*. 841–852.