Mapping the Design Space of Interactions in Human-AI Text Co-creation Tasks

Zijian Ding
College of Information Studies
University of Maryland, College Park
USA

Joel Chan
College of Information Studies
University of Maryland, College Park
USA

ABSTRACT

Large Language Models (LLMs) have demonstrated impressive text generation capabilities, prompting us to reconsider the future of human-AI co-creation and how humans interact with LLMs. In this paper, we present a spectrum of content generation tasks and their corresponding human-AI interaction patterns. These tasks include: 1) fixed-scope content curation tasks with minimal human-AI interactions, 2) independent creative tasks with precise human-AI interactions, and 3) complex and interdependent creative tasks with iterative human-AI interactions. We encourage the generative AI and HCI research communities to focus on the more complex and interdependent tasks, which require greater levels of human involvement.

KEYWORDS

Large Language Models, Analogy, Creativity Support Tools

ACM Reference Format:
Zijian Ding and Joel Chan. 2023. Mapping the Design Space of Interactions in Human-AI Text Co-creation Tasks. In *CHI 2023 Workshop on Generative AI and HCI), Apr 28, 2023, Virtual.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3411764.3445618

1 INTRODUCTION

Large Language Models (LLMs), such as Generative Pre-trained Transformer 3 (GPT-3) [3], have garnered significant attention from researchers and practitioners for their ability to generate text content. The rapid success of ChatGPT[1] - reaching 100 million monthly active users just two months after its launch and setting a record for the fastest-growing consumer application in history[2] - highlights not only the potential and capabilities of Generative AI for producing precise and personalized text content, but also the critical role of interface and interaction in communicating with AI. Since ChatGPT is a variant of GPT-3 fine-tuned for conversational tasks, the technical foundation remains similar; instead, the primary difference seems to be the shift in human-AI interaction paradigms, from prompt programming, parameter tuning and autocompleting in OpenAI playground or API[3], to interactive conversations in ChatGPT.

The wide deployment of these language models beyond academic research projects carries risks, but has also uncovered a broader sense of the potential applications of these models: from foundation models for traditional NLP applications such as text classification, summarization, and information extraction, to generative applications such as analogy generation and even complex creative tasks such as fiction creation. What might this wave of progress enable for augmenting — rather than automating — human creativity?

To help make sense of this question, in this position paper we sketch out a possible **design space of human-AI co-creation**, focusing on the role of humans and their interactions and collaboration with Generative AI. Specifically, we synthesize previous research on human-AI interactions in text generation applications and tasks onto a spectrum of task complexity and creativity: 1) fixed-scope text curation tasks, 2) atomic creative tasks and 3) complex and interdependent creative tasks; and propose a mapping of this spectrum to a taxonomy of existing design patterns of human-AI interaction that can address the requirements and challenges of each point in the spectrum, as shown in Figure 1. Finally, we suggest future avenues for further exploring this spectrum, especially towards more complex and interdependent creative tasks.

2 TYPES OF HUMAN-AI INTERACTIONS FOR TEXT GENERATION

Our discussion of human-AI interaction for text generation draws on the taxonomy of five common human-AI interactions proposed by Cheng et al. [6]: 1) *guiding model output*, 2) *selecting or rating model output*, 3) *post-editing*, 4) *interactive editing* (initiated by AI) and 5) *writing with model assistance* (initiated by human).

The former three interaction types do not involve rounds of iterations between human and AI, referred to as *precise human-AI interactions*. In contrast, the latter two types involve rounds of iterations between human and AI, referred to as *iterative human-AI interactions*. In the following sections, we discuss how these interaction types can be usefully mapped to a spectrum of text-based human-AI co-creation tasks that range from low to high complexity and creativity.

[1]https://chat.openai.com/chat
[2]https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/
[3]https://platform.openai.com/overview

CHI '23 Workshop on Generative AI and HCI, Apr 28, 2023, Virtual

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/3411764.3445618

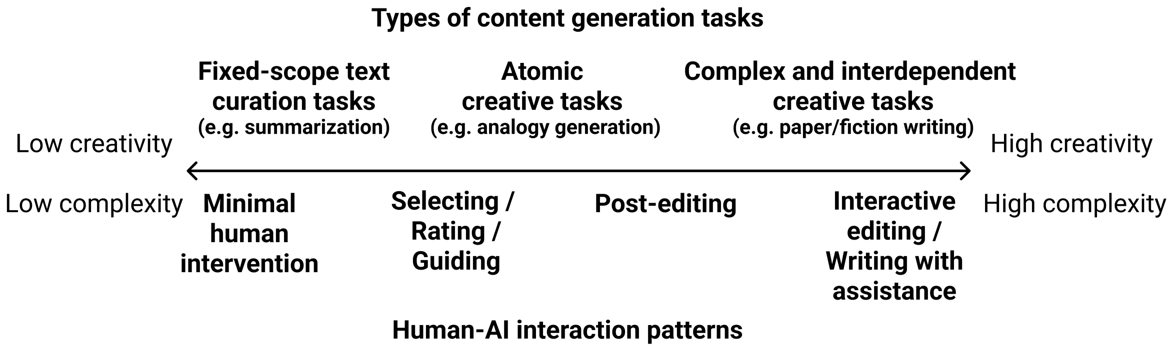

Figure 1: Spectrum of human-AI co-creation tasks and corresponding human intervention complexity. The upper half describes the spectrum of text-based co-creation tasks from low to high creativity and complexity; the bottom half proposes a mapping of the points on these spectrum to human-AI interaction patterns from the taxonomy in [6].

3 SPECTRUM OF TEXT-BASED HUMAN-AI CO-CREATION TASKS

3.1 Minimal human-AI interactions in fixed-scope content curation tasks

The first point in our spectrum can be described as fixed-scope content curation tasks. Previous research has indicated that Large language models (LLMs) can effectively handle well-defined, content curation tasks such as text summarization [9, 15], content refinement [17], and code explanation [18, 19]. In these tasks, existing knowledge and information are summarized and presented in a cohesive manner, but no new knowledge is generated. In these fixed-scope tasks, advanced LLM models such as GPT-3 DaVinci have already produced satisfactory results with no human intervention on the outputs [3]. According to a study conducted by Clark et al. [7], text generated by GPT-3 exhibited such a high degree of linguistic sophistication that it was almost impossible for human evaluators to discern whether it had been authored by a machine or a human, and carefully designed frameworks are required to scrutinize different types of human and machine errors and determine the authorship of a piece of text [11]. The aforementioned evidence serves as a testament to the high quality of machine-generated text, raising the possibility that in the future, the amount of human involvement required for content curation tasks could be significantly reduced.

3.2 Precise human-AI interactions in atomic creative tasks

The second point in our spectrum can be described as atomic creative tasks. By creative, we mean outputs that are both novel and useful [22, 23], including generating analogies / metaphors / analogous design concepts [2, 16, 26, 29], slogans [8] and inspirations for tweetorials (short technical explanations on Twitter) [14]. The broad coverage of existing knowledge by LLMs such as GPT-3 can help to create novel connections, known as "creative leaps" [5, 20, 24]. However, generating truly creative, inspiring, and insightful content often requires domain-specific knowledge, including subtle and implicit knowledge, which may not be present in the training data for LLMs. To compensate for this lack of knowledge, LLMs must be guided by carefully crafted prompts and examples, and their outputs must be selected or edited by humans to ensure their quality. In other words, these specific creative tasks demand precise human-AI interactions.

For instance, when generating analogous problems, classic analogous problems created or selected by humans, such as Duncker and Lees' [12] radiation problem, can be applied to *guide* LLMs in the analogy generation process. The generated analogies must then be *selected* or *rated*, even *post-edited* by humans to avoid any biases, illegal, or inappropriate content. In some recent experiments to explore the current performance of LLMs on atomic creative tasks, we generated 120 analogous problems with GPT-3 *text-davinci-002* model and the Duncker and Lees' analogous problem for one-shot learning, and asked participants to use them to reformulate the original problem [10]. Our results showed that the AI-generated analogous problems were frequently perceived as helpful (with a median helpfulness rating of 4 out of 5) and led to observable changes in problem formulation in approximately 80% of cases. However, we also found that up to 25% of the outputs were potentially harmful, mostly due to potentially upsetting content that was not biased or toxic. Our findings demonstrate the potential of using LLMs for atomic creative tasks, but also highlight the need for human intervention. Below is an example of the LLM-generated analogous problem and how participant used the analogy to stimulate reformulation of the original problem:

> *Original problem*
> Stakeholder: owners of travel agency
> Context: the restriction of pandemic has been mitigated and people are willing to travel again
> Goal: reopen their traveling business
> Obstacle: cannot find enough employees because people have left the travel industry during the pandemic
> *GPT-3 generated analogous problem*
> Stakeholder: a farmer
> Context: the restriction of the use of pesticides has been mitigated
> Goal: use pesticides to increase crop yield

> Obstacle: the farmer cannot afford to buy pesticides
> *Participant's response to reformulation question*
> **It's not that the travel agency can't find employees, it's that they can't afford to pay employees to work for them after being closed for so long**, thus causing a feedback loop of: not enough employees -> less money ->cant afford to hire employees -> not enough employees.

3.3 Iterative human-AI interactions in complex and interdependent creative tasks

Our third and final point in the spectrum can be described as complex and interdependent creative tasks. These are larger scale creative tasks containing a set of subtasks interdependent to each other, such as storytelling [1, 4, 13, 21, 25, 27, 28]. Those tasks go beyond just the combination of atomic creative tasks and cannot be decomposed into atomic creative tasks. Those tasks require not only domain-specific knowledge but also the ability to plan, reason, delve into ideas, and retain context over time in order to generate new and coherent concepts, knowledge and stories.

For these tasks, we argue that humans and large language models (LLMs) must work closely together and iteratively refine the text content, and different types of interactions are needed depending on the iteration stages. The co-creation of a story, for example, may involve a series of iterations of guidance, selection, and post-editing by both humans and LLMs [28]. And fixed-scope text curation tasks and specific creative tasks can serve as building blocks for more complex and comprehensive creative tasks, such as including a literature review for scientific paper writing. Expert human review, justification, and post-editing are crucial to ensure the originality and logic of the AI-generated content and its alignment with other elements.

4 FUTURE DIRECTIONS FOR TEXT-BASED HUMAN-AI CO-CREATION

Our belief and hope is that iterative human-AI interactions in complex and interdependent creative tasks will become a focus of future research on human-AI text co-creation, due to their complexity, potential, and the need for intensive human-AI interaction. Current LLM-powered tools show potential for supporting that vision of human-AI collaboration in creative tasks, but there is still room for improvement in certain areas.

For example, chatbots like ChatGPT are capable of supporting multiple rounds of interactions but currently only offer one interaction paradigm - guiding the model output. This limitation may reduce the efficiency and overall experience of creative writing with the tool. Within the context of human-AI co-writing, Yuan et al. [28] also focused primarily on exploring various formats of guiding model outputs, such as continuation, elaboration, story seeding, and infilling. While their work sheds important light on the collaborative aspects of story writing, there is potential for artificial intelligence to play an even more proactive role through interactive selecting and editing paradigms.

There is also a difficult set of challenges around *evaluation*. For example, NLP progress has benefited substantially from well-defined benchmarks. This can work well for accelerating progress for fixed-scope content curation tasks, but is a poor fit for atomic and complex and interdependent creative tasks with no single "correct" reference output. Crowdsourced human evaluations may also index only surface-level linguistic coherence vs. more substantive dimensions of quality without more specific (task-specific) instructions or domain expertise: for example, Clark et al [7] reported that crowd workers mostly relied on form vs. content heuristics to make their judgments about human-likeness of LLM-generated text. Community-based and participatory research methods may be needed to address these challenges.

A final challenge concerns the question of how to integrate domain knowledge and expertise. For example while novice users may have the most to gain from human-AI co-creation tools, they may need domain expertise to effectively control the outputs generated by AI. We believe a promising direction is to explore the design of co-creation tools that integrate the generative strengths of LLMs with sources of domain knowledge (e.g., heuristics, design patterns, knowledge bases, access to other peers and experts for feedback/validation); for example, the new Bing search tool[4] integrates question-answering and summarization loops with API calls to knowledge bases.

We believe these open problems are difficult but tractable, and look forward to exploring solutions to these problems with the human-AI interaction community.

REFERENCES

[1] Shlomo Berkovsky, Yoshinori Hijikata, Jun Rekimoto, Margaret Burnett, Mark Billinghurst, and Aaron Quigley. 2018. How Novelists Use Generative Language Models: An Exploratory User Study. In *23rd International Conference on Intelligent User Interfaces*. ACM, Tokyo Japan.

[2] Bhavya Bhavya, Jinjun Xiong, and Chengxiang Zhai. 2022. Analogy Generation by Prompting Large Language Models: A Case Study of InstructGPT. http://arxiv.org/abs/2210.04186 arXiv:2210.04186 [cs].

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]* (June 2020). http://arxiv.org/abs/2005.14165 00030 arXiv: 2005.14165.

[4] Chen Cao. 2023. Leveraging Large Language Model and Story-Based Gamification in Intelligent Tutoring System to Scaffold Introductory Programming Courses: A Design-Based Research Study. http://arxiv.org/abs/2302.12834 arXiv:2302.12834 [cs].

[5] Joel Chan, Pao Siangliulue, Denisa Qori McDonald, Ruixue Liu, Reza Moradinezhad, Safa Aman, Erin T. Solovey, Krzysztof Z. Gajos, and Steven P. Dow. 2017. Semantically Far Inspirations Considered Harmful?: Accounting for Cognitive States in Collaborative Ideation. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition (C&C '17)*. ACM, New York, NY, USA, 93–105. https://doi.org/10.1145/3059454.3059455

[6] Ruijia Cheng, Alison Smith-Renner, Ke Zhang, Joel Tetreault, and Alejandro Jaimes-Larrarte. 2022. Mapping the Design Space of Human-AI Interaction in Text Summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 431–455. https://doi.org/10.18653/v1/2022.naacl-main.33

[7] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. *arXiv:2107.00061 [cs]* (July 2021). http://arxiv.org/abs/2107.00061 00008 arXiv: 2107.00061.

[8] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on

[4] https://www.bing.com/new

Slogans and Stories. In *23rd International Conference on Intelligent User Interfaces*. ACM, Tokyo Japan, 329–340. https://doi.org/10.1145/3172944.3172983

[9] Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. 2022. Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries. https://doi.org/10.1145/3526113.3545672 arXiv:2208.09323 [cs].

[10] Zijian Ding, Arvind Srinivasan, Stephen MacNeil, and Joel Chan. 2023. Fluid Transformers and Creative Analogies: Exploring Large Language Models' Capacity for Augmenting Cross-Domain Analogical Creativity. *arXiv preprint arXiv:2302.12832* (2023).

[11] Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah Smith, and Yejin Choi. 2022. Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 7250–7274. https://doi.org/10.18653/v1/2022.acl-long.501

[12] Karl Duncker. 1945. On problem-solving. *Psychological Monographs* 58, 5 (1945), i–113. https://doi.org/10.1037/h0093599

[13] Jonas Freiknecht and Wolfgang Effelsberg. 2020. Procedural Generation of Interactive Stories using Language Models. In *International Conference on the Foundations of Digital Games*. ACM, Bugibba Malta, 1–8. https://doi.org/10.1145/3402942.3409599

[14] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for Science Writing using Language Models. In *Designing Interactive Systems Conference*. ACM, Virtual Event Australia, 1002–1019. https://doi.org/10.1145/3532106.3533533

[15] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News Summarization and Evaluation in the Era of GPT-3. http://arxiv.org/abs/2209.12356 arXiv:2209.12356 [cs].

[16] 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 Bernstein, and Percy Liang. 2022. Evaluating Human-Language Model Interaction. http://arxiv.org/abs/2212.09746 arXiv:2212.09746 [cs] version: 2.

[17] Zhicheng Lin. 2023. *Why and how to embrace AI such as ChatGPT in your academic life*. preprint. PsyArXiv. https://doi.org/10.31234/osf.io/sdx3j

[18] Stephen MacNeil, Andrew Tran, Arto Hellas, Joanne Kim, Sami Sarsa, Paul Denny, Seth Bernstein, and Juho Leinonen. 2022. Experiences from Using Code Explanations Generated by Large Language Models in a Web Software Development E-Book. http://arxiv.org/abs/2211.02265 arXiv:2211.02265 [cs].

[19] Stephen MacNeil, Andrew Tran, Dan Mogil, Seth Bernstein, Erin Ross, and Ziheng Huang. 2022. Generating Diverse Code Explanations using the GPT-3 Large Language Model. In *Proceedings of the 2022 ACM Conference on International Computing Education Research - Volume 2*. ACM, Lugano and Virtual Event Switzerland, 37–39. https://doi.org/10.1145/3501709.3544280

[20] Lior Noy, Yuval Hart, Natalie Andrew, Omer Ramote, Avi Mayo, and Uri Alon. 2012. A quantitative study of creative leaps, Mary Maher, Kristian Hammond, Alison Pease, {Rafael Pérez, Dan Ventura, and Geraint Wiggins (Eds.). 72–76. http://computationalcreativity.net/iccc2012/wp-content/uploads/2012/05/072-Noy.pdf

[21] Melissa Roemmele. 2016. Writing Stories with Help from Recurrent Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence* 30, 1 (March 2016). https://doi.org/10.1609/aaai.v30i1.9810

[22] Mark A. Runco and Garrett J. Jaeger. 2012. The Standard Definition of Creativity. *Creativity Research Journal* 24, 1 (Jan. 2012), 92–96. https://doi.org/10.1080/10400419.2012.650092

[23] R. Keith Sawyer. 2012. *Explaining creativity: the science of human innovation* (2nd ed.). Oxford University Press, New York.

[24] Pao Siangliulue, Joel Chan, Krzysztof Gajos, and Steven P. Dow. 2015. Providing timely examples improves the quantity and quality of generated ideas. In *Proceedings of the ACM Conference on Creativity and Cognition*. https://doi.org/10.1145/2757226.2757230

[25] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman. 2022. Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence. *ACM Transactions on Computer-Human Interaction* (Feb. 2022), 3511599. https://doi.org/10.1145/3511599

[26] Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2022. Emergent Analogical Reasoning in Large Language Models. http://arxiv.org/abs/2212.09196 arXiv:2212.09196 [cs].

[27] Daijin Yang, Yanpeng Zhou, Zhiyuan Zhang, and Toby Jia-Jun Li. [n.d.]. AI as an Active Writer: Interaction strategies with generated text in human-AI collaborative fiction writing. ([n. d.]).

[28] 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*. ACM, Helsinki Finland, 841–852. https://doi.org/10.1145/3490099.3511105

[29] Q. Zhu and J. Luo. 2022. Generative Pre-Trained Transformer for Design Concept Generation: An Exploration. *Proceedings of the Design Society* 2 (May 2022), 1825–1834. https://doi.org/10.1017/pds.2022.185 Publisher: Cambridge University Press.