Research Statement Katy Ilonka Gero

I am a Human-Computer Interaction researcher with a focus on Human-Centered AI.

Writing is a deeply human activity: we strive to make ourselves understood by others, to be heard and seen. I study how language technologies can support writers in communicating complex ideas, because our ability to communicate complex thought should not prevent us from sharing these thoughts accurately and engagingly. While AI systems can generate fluent and convincing text, they lack the communicative intent that people bring. I design ethical AI systems that writers feel are authentic and meaningful. My research focuses on questions that combine AI innovation with cognitive models of how we think and write:

1. How can we **design novel interactive systems** that support complex writing tasks, like generating new metaphorical explanations or explaining technical topics to a general audience?
2. How can we address the **sociotechnical issues of writing support** and consider writer agency, authenticity, and (mis)understandings of system capabilities?

I draw on quantitative and qualitative methods to answer these questions. I have published in premier venues for HCI such as CHI and CSCW, as well as NLP venues such as ACL and INLG, resulting in 7 first author full conference papers and 8 workshop or poster papers, including a comment in Nature Machine Intelligence. My work received a Best Paper award at CHI (top 1% of submissions) and I have written successful grant proposals that resulted in funding from the National Science Foundation, the Brown Institute for Media Innovation, and Amazon Science.

Coherent Suggestions for Complex Writing

Generating metaphors. Early work on generating suggestions for writing demonstrated the potential to support writers, but most suggestions failed to be relevant to writers [Roemmele 2015, Roemmele 2016, Clark 2018]. I investigated this myself with GPT-2, and replicated these results: suggestions typically did not conform to the relevant context [Calderwood 2020]. My first foray in writing support solved this problem of generating suggestions that were coherent to context.

Figure 1: I built Metaphoria to show how targeted writing support could provide more coherent suggestions to writers.

I designed Metaphoria, a system that generated metaphorical connections between two nouns (Figure 1) [Gero 2019]. Metaphoria focused on a more specific writing task than previous work. By thinking about writing as a cognitive process, I understood that writing goals exist on a spectrum from high level ("write a paper") to low level ("find a more formal word"). Prior work focused on quite high level goals. To design a system to be more coherent to context, I focused on a lower level goal: writing a metaphor. A metaphor is a complex, nuanced, and creative rhetorical device used across domains from poetry to journalism. Metaphoria used a unique algorithm I developed which leveraged knowledge graphs and word embeddings. It performed better than existing metaphor generation algorithms of the time, and the suggestions were found to be coherent to context a majority of the time by both undergraduates and professional poets.

Metaphoria inspired a range of research by other labs, from the system Metaphorian, which generates extended metaphors using language models [Kim 2023], to using metaphor generation as a language model evaluation task [Lee 2022]. In a recent class on Research Topics in HCI, graduate students re-tested Metaphoria and anecdotally found it still produced more creative and interesting metaphors than GPT-4.

Metaphoria was an early demonstration that by combining relevant AI technologies with a deep understanding of writing and rhetoric, we can design systems that support writers with complex writing tasks. This work was funded by my NSF Graduate Research Fellowship, where I proposed studying poetic devices to fuel language technology innovation.

Explaining technical concepts. As language models were improving, I performed early investigations into their capability to support complex writing. I studied a popular science writing format called 'tweetorials,' which are Twitter threads of about 500 words that explain a technical concept to a general audience. Tweetorials are an accessible way for anyone to explain their work to the public, but require making the concept interesting to the average reader.

In collaboration with a science journalist, I examined the tweetorial format and identified the common rhetorical devices used [Gero 2021]. Grounded in this, I developed a series of prompts based on expository and narrative theory, such as "One application of {topic} in the real world is…" and "For example, {topic} can cause…", where quality prompt completions would provide meaningful ideas about how to make a topic interesting to the reader.

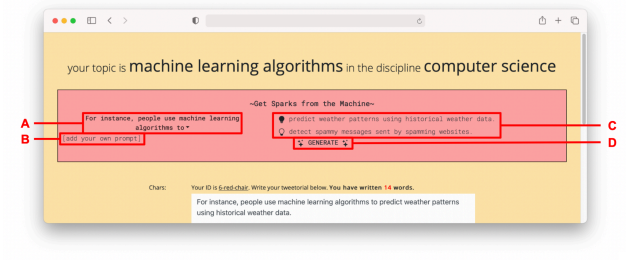

Figure 2: I built Sparks to understand how language models could support science writing, which requires both creative ways of making a topic engaging while also requiring knowledge about the topic itself.

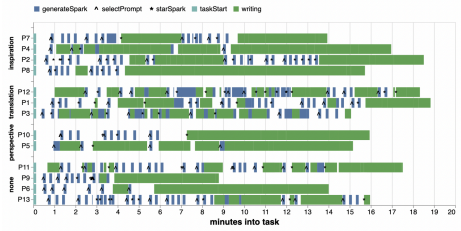

Figure 3: By analyzing when writers engaged with suggestions (marked blue) versus when they engaged in writing (marked in green), distinct patterns of usage can be seen which align with qualitative evaluations.

I developed a system called Sparks which provided ideas about how to make a technical topic interesting to the average reader [Gero 2022]. The problem with using off-the-shelf language models was that the generated sentences tended to be generic, vague, and repetitive. I developed a custom decoding algorithm which upweighted the likelihood of words with higher specificity and resulted in sentences that were more relevant and factually accurate than existing decoding methods. In a user study with STEM graduate students writing about their own work, Sparks was used not only for idea generation but also for speeding up the writing of short and clear explanatory sentences.

High-impact writing tasks come with ethical concerns: miscommunication about science can have harmful consequences. Given the potential for misuse, I am planning to investigate how to prevent harm and describe my plans in the Future Work section.

Sparks demonstrated that even when dealing with a writing task that required technical knowledge, by combining an understanding of the writing process with AI innovation we can develop improved systems that support writers with difficult and important writing tasks. This work was funded by the Brown Institute for Media Innovation, where I proposed studying tweetorials as an increasingly popular form of science communication on social media.

Addressing Sociotechnical Issues

The social dynamics of creative writing support. In both Metaphoria and Sparks, there were always outlier participants in the user studies who found the system was not useful. I wanted to understand why, as it hinted at underlying factors that existing research on writing assistants was failing to reveal. I performed a qualitative study with 20 creative writers from a variety of genres, disciplines, and educational backgrounds, probing when and why a writer might turn to a computer, versus a peer or mentor, for support [Gero 2023a]. This resulted in a two-level taxonomy as well as associated themes and a graphical model of how these three elements interact.

This model demonstrates why we need to align our AI systems with human values, and can explain why some writers value systems and others don't, despite seeing similar outputs. For instance, some writers value independence for idea generation, so any system that supports idea generation will be problematic for them. Other writers value execution over idea generation, and may cognitively engage in even mediocre systems to find some kind of utility.

This research inspired me to write a public-audience article for Wired, titled AI Reveals the Most Human Part of Writing, which was subsequently referenced by the Venture Capitalist firm Sequoia Capital in their writing on how to invest in AI-powered writing assistants.

The social dynamics of support model (Figure 4) helps researchers and designers build human-centered AI systems. It demonstrates the unique opportunities computers can bring to writing support, while also outlining what writers are worried about when using computers for support. Writers are not a monolith, and their differing values dictate which part of the writing process they want support with.

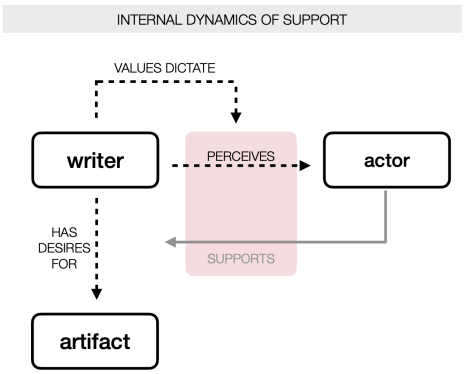

Figure 4: Writers have values about the kinds of support interactions they would like to have, which impacts who or what they turn to for support.

Characterizing language model outputs. While there have been huge improvements in text generation, it remains difficult to make sense of model capabilities, choose between models or prompts, or select an ideal output from a set of potential possibilities. Existing systems typically either show users one output at a time, or engage in extensive annotation. In my work as a Postdoctoral Fellow at Harvard University, I explored how to support people in making use of and characterizing 10s to 100s of language model outputs. This has applications in writing support, where writers may want to select from 10s of outputs and system designers may want to inspect 100s of outputs as part of prompt engineering. It also has implications for additional language model applications.

By drawing on knowledge of how we read, skim, and compare text, I designed a series of highlighting visualizations and evaluated these in a controlled user study and a series of open-ended case studies [Gero *under review*]. I found that supporting the skimming of many language model outputs allows people to more quickly pick out optimal outputs or portions of outputs, determine more differences between models, and more quickly identify if a model is achieving the desired result.

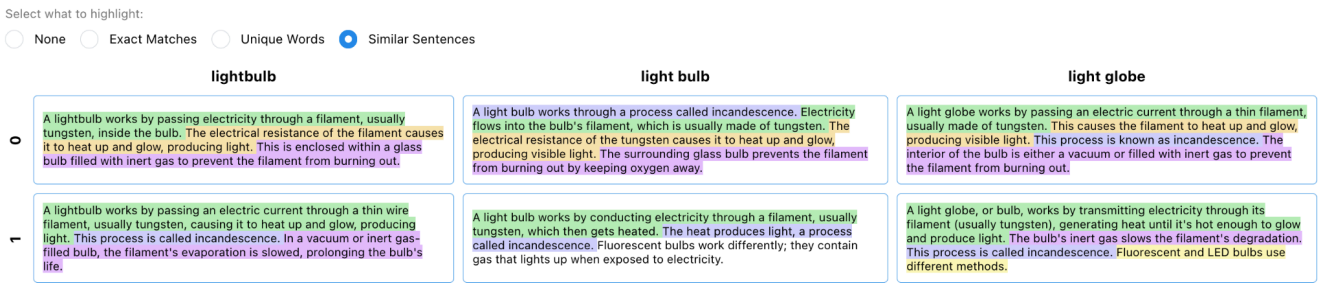

Figure 5: This system allows users to compare responses from different models or prompts. Here, the prompt is “How does a {item} work?” and shows variation across different ways to say ‘lightbulb’ in the prompt. By highlighting similar sentences across repeated responses, users can easily infer the impact of the prompt on the response, as well as identify different content elements.

Understanding and characterizing language models is an important piece of building systems atop these models. For example, system designers must select the best pre-trained or fine-tuned model, and automated benchmarks fail to capture aspects of interest which are often unique to particular goals or tasks. Making sense of language models is also a key component of auditing them, and creating tools that support this auditing by end-users (not just researchers or software developers) supports ethical development.

Characterizing language model outputs supports human sensemaking; if we want to build human-centered AI, then people must have the tools to understand the systems they are interacting with. This work was partially funded by an Amazon Research Award, where I proposed developing novel visualization techniques to support improved utility of large language models.

Future Work

As AI language technologies have improved rapidly, a number of pressing questions have opened up about aligning them with human and societal values. Here I present two lines of work which are particularly key to ensuring safe and meaningful deployment of these technologies. Each grows from my existing work and will have a large impact on how we use AI in the coming years and even decades.

Measuring people's mental models of large language models. Joint human-AI abilities improve when people have a better mental model of the AI they are collaborating with [Gero 2020]. However, little work has rigorously investigated how to elicit or measure people's mental models of large language models. Such measurement is the first step in improving joint human-AI abilities and preventing harms based on misunderstandings of model capabilities.

I propose drawing on experimental paradigms from psychology, cognitive science, and safety engineering to design new experiments that can measure people's understanding of three critical language model capabilities: (1) stochasticity and regeneration, (2) conceptual consistency of outputs, and (3) sensitivity to prompt engineering. I have already developed an initial proposal for this research which will be submitted to the NSF.

Human-centered design of ethical language models. Datasets fuel the abilities of large language models, and rigorous data work requires deep technical understanding (and often technical innovation) separate from model innovation [Gero 2023b]. I am currently starting a project to engage a variety of literary communities—amateur and fan fiction writers, literary writers with small and large presses, pulp fiction writers, and self-published writers—to understand under what conditions, if any, they would want to have their work included in a language model. For instance, how would they want to be credited and/or compensated? Based on community engagement, I will propose a new kind of licensing agreement, and build a 40GB contemporary creative writing dataset, which would allow me to train an open-source language model the size of GPT-2.[1] This work will be done in collaboration with the Library Innovation Lab, housed in Harvard's Law Library, and with the collaboration of legal scholars.

Such a project would have social as well as technical implications. In light of the recent labor concerns about language models (such as AI writing being a core component of the Writer's Guild of America contract negotiations, and the current copyright lawsuits being levied against OpenAI) this project would model improved labor practices and support certain legal arguments. Additionally, when considering preferences about credit and compensation, I will explore novel technical methods for tracing model reliance on specific data points when generating text. Influence functions have proved capable of such 'tracing' for understanding word embedding bias [Brunet 2019] and classification fairness [Sattigeri 2022] and I believe could be fruitful in this context.

Conclusion

Writing is an ancient technology, but a technology nonetheless. Newer technologies, especially large language models, are changing what writing is, means, and may become. I believe in technically rigorous and inventive work, but I also believe that we must engage in the social and psychological aspects of technology in order to create positive and meaningful innovation. Long term, I am interested in continuing to do technical work with generative AI, while further exploring community engagement and the psychology of interacting with generative models. I'd like my work to model how we can be excited about technology while simultaneously being alert to its potentials for misuse and harm. I hope to be a responsible steward, ushering in changes with alertness and care.

[1] 40GB is about half a million books; although this sounds like a large number, anywhere from one to four million books are published in the USA each year.

References to my publications

[Gero 2019] **Katy Ilonka Gero** and Lydia B. Chilton. “Metaphoria: An Algorithmic Companion for Metaphor Creation.” *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ACM, 2019, pp. 1–12. *DOI.org (Crossref)*, https://doi.org/10.1145/3290605.3300526.

[Gero 2020] **Katy Ilonka Gero**, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel, and Wei Zhang. “Mental Models of AI Agents in a Cooperative Game Setting.” *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ACM, 2020, pp. 1–12. *DOI.org (Crossref)*, https://doi.org/10.1145/3313831.3376316.

[Calderwood 2020] Alex Calderwood, Vivian Qiu, **Katy Ilonka Gero**, Lydia B. Chilton. “How Novelists Use Generative Language Models: An Exploratory User Study.” *23rd International Conference on Intelligent User Interfaces*, ACM, 2018.

[Gero 2021] **Katy Ilonka Gero**, Vivian Liu, Sarah Huang, Jennifer Lee, Lydia B. Chilton. “What Makes Tweetorials Tick: How Experts Communicate Complex Topics on Twitter.” *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, Oct. 2021, pp. 1–26. *DOI.org (Crossref)*, https://doi.org/10.1145/3479566.

[Gero 2022a] **Katy Ilonka Gero**, Vivian Liu, and Lydia B. Chilton. “Sparks: Inspiration for Science Writing Using Language Models.” *Designing Interactive Systems Conference*, ACM, 2022, pp. 1002–19. *DOI.org (Crossref)*, https://doi.org/10.1145/3532106.3533533.

[Gero 2023a] **Katy Ilonka Gero**, Tao Long, and Lydia B. Chilton. “Social Dynamics of AI Support in Creative Writing.” *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ACM, 2023, pp. 1–15. *DOI.org (Crossref)*, https://doi.org/10.1145/3544548.3580782.

[Gero 2023b] **Katy Ilonka Gero**, Payel Das, Pierre Dognin, Inkit Padhi, Prasanna Sattigeri, and Kush R. Varshney. “The Incentive Gap in Data Work in the Era of Large Models.” *Nature Machine Intelligence*, vol. 5, no. 6, June 2023, pp. 565–67. *DOI.org (Crossref)*, https://doi.org/10.1038/s42256-023-00673-x.

[Gero *under review*] **Katy Ilonka Gero**, Chelse Swoopes, Ziwei Gu, Jonathan Kummerfeld, and Elena Glassman. “Visualizing and Characterizing Many Language Model Outputs at Once.” *Under review.*

References to publications by others

[Roemmele 2015] Melissa Roemmele and Andrew S. Gordon. “Creative Help: A Story Writing Assistant.” *Interactive Storytelling*, edited by Henrik Schoenau-Fog et al., vol. 9445, Springer International Publishing, 2015, pp. 81–92.

[Roemmele 2016] Melissa Roemmele. “Writing Stories with Help from Recurrent Neural Networks.” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016. https://ojs.aaai.org/index.php/AAAI/article/view/9810.

[Clark 2018] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. “Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories.” *23rd International Conference on Intelligent User Interfaces*, ACM, 2018, pp. 329–40. *DOI.org (Crossref)*, https://doi.org/10.1145/3172944.3172983.

[Brunet 2019] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. “Understanding the Origins of Bias in Word Embeddings.” *Proceedings of the 36th International Conference on Machine Learning*.

[Sattigeri 2022] Prasanna Sattigeri, Soumya Ghosh, Inkit Padhi, Pierre Dognin, Kush R. Varshney. "Fair infinitesimal jackknife: Mitigating the influence of biased training data points without refitting." *Advances in Neural Information Processing Systems* 35 (2022): 35894-35906.

[Lee 2022] 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. *Evaluating Human-Language Model Interaction*. arXiv:2212.09746, arXiv, 20 Dec. 2022. *arXiv.org*, http://arxiv.org/abs/2212.09746.

[Kim 2023] Jeongyeon Kim, Sangho Suh, Lydia B Chilton, and Haijun Xia. “Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing.” *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, ACM, 2023, pp. 115–35. *DOI.org (Crossref)*, https://doi.org/10.1145/3563657.3595996.