We interact with text-based AI systems (e.g., Google search autocomplete and Gmail's smart compose) dozens of times a day without even thinking about it. Behind these systems, however, are models that are primarily trained and evaluated *in isolation*, in the sense that they are trained on a static dataset and subsequently evaluated on a static benchmark. This kind of model development neglects that ultimately, these models will be used to *interact with humans*, complement human capabilities, and potentially improve performance based on human feedback over time. As a result, when building interactive systems, it is unclear how these models will behave in interactive settings and which models are more desirable than others. This gap raises a question at the heart of my research: **How can we evaluate and develop models that can interact with humans?**

I strive to (1) **evaluate models based on their ability to complement humans** and (2) **develop models with human needs in mind**. When evaluating models, I first observe how humans interact with models to understand the benefits humans gain from interaction, measure the interactability of models based on this understanding, and then find ways to incentivize researchers to build models that can augment human capabilities. Similarly, when developing models, I identify human needs and extend AI capabilities to provide assistance.

My research area is in natural language processing (NLP), with a focus on language models (LMs) and how they interact with humans. I draw inspiration from human-computer interaction (HCI) and publish in both NLP and HCI conferences (e.g., ACL, NAACL, NeurIPS, and CHI). This year, my work has been recognized with an Honorable Mention Award at CHI 2022, featured in various media outlets including *The Economist*, and further **used by professional copywriters, story writers, high school teachers, a journalist, and a movie scriptwriter**. In the future, I am eager to investigate how AI systems will change the way we write and communicate.

Evaluating AI in interactive settings

To evaluate AI systems' ability to complement humans, AI experts must understand how humans interact with the systems and how systems can augment human capabilities. I help experts evaluate AI in interactive settings by developing tools and methods for collecting and analyzing the process of interaction.

Capturing human-AI interaction. I built **CoAuthor, a platform for collecting human-AI interaction traces** at a keystroke level along with timestamps, which can be replayed precisely (Figure 1) [Lee et al., 2022a]. With this platform, I captured 1445 writing sessions between 63 crowd workers and four instances of GPT-3. Unlike prior work that only studied human-AI interaction in specific domains or with specific user groups, this work showed that it is useful to collect a large interaction dataset to understand how LMs perform in diverse interactive contexts.

Figure 1: I built CoAuthor (a text editor with real-time suggestions generated by GPT-3) to observe human-AI interaction and evaluate AI systems' capabilities in diverse contexts.

Interaction traces provide insights into model behavior that are hard to infer from model outputs alone. For instance, I found that the collaboration patterns vary greatly across users, topics, and models. Model performance varied across topics, but overall, the sentences written by humans and AI together had fewer errors and more diverse vocabulary than either alone. A particularly exciting outcome of this work is that I was able to quantitatively measure instances where AI supported humans through *ideation*, i.e., suggesting new ideas that humans later adopt into their writing. With this work, I demonstrated **the value of interaction traces for analyzing model behavior when working with humans**.

Measuring AI's ability to interact with humans. My most recent work builds on interaction traces and proposes a new benchmarking framework for evaluating human-AI interaction (Figure 2) [Lee et al., 2022b]. To

measure AI's ability to interact with humans, I constructed a framework that breaks down evaluation into three dimensions—targets, perspectives, and criteria—and **emphasized human-centered aspects for each dimension** as follows: (1) targets include more than just the final output, and cover the entire interaction process (e.g., user queries and edits); (2) perspectives are not limited to third parties, but the users who interact with AI to capture first-person experiences; and (3) criteria include not only quality (i.e., measures of targets, such as accuracy), but also preference (i.e., attitudinal measures of humans, such as enjoyment).

With this framework, I designed five tasks and their interactive solutions—social dialogue, question answering, crossword puzzles, text summarization, and metaphor generation—and ran user studies with 1000+ crowd workers and four LMs. I led 15 Postdocs, Ph.D. students, and Master's students for a year, supervising and supporting the overall effort. One important takeaway from our experiments is that **better non-interactive performance does not correspond to better human-AI interaction**. For example, LMs optimized for quality tend to generate more generic outputs, making them less preferable in creative tasks.

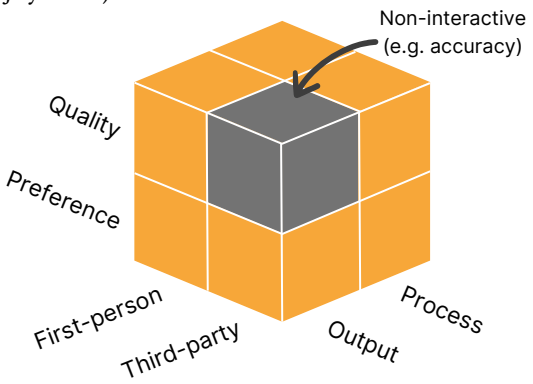

Figure 2: I propose a new benchmarking framework to evaluate human-AI interaction. Standard, non-interactive benchmarking covers only one cell of the cube (gray), whereas our framework considers all cells.

To my knowledge, this is **the first work to underscore human-centered aspects in the context of benchmarking across multiple domains**, while showing the trade-offs between those dimensions. By doing so, I emphasized the critical need to shift our current benchmarking practice to consider interactive settings.

Incentivizing AI experts to build AI for humans. After identifying ways to measure desirable properties in human-AI interaction, how can we incentivize AI experts to build AI that complements humans? Consider the traditional NLP task of finding synonyms for a word in a sentence (i.e., lexical substitution). Previous benchmarks were collected based on the synonyms humans could quickly recall, and therefore, lacked coverage of the synonyms. As a result, LMs designed to perform well on these benchmarks mimic human recall capabilities, and cannot produce synonyms that would be most helpful to humans.

To encourage AI experts to build LMs that can augment human capabilities to find synonyms, I built a new benchmark, **Stanford Word Substitution Benchmark (Swords), with 4.1x higher coverage and 1.5x higher quality synonyms** [Lee et al., 2021]. This benchmark is significantly less biased by what humans can remember on the spot, as the data collection process used human judgment rather than recall to assess the appropriateness of synonyms. By evaluating state-of-the-art LMs and commercial systems (e.g., Wordtune and Thesaurus.com) on the benchmark, I found that previous LMs tailored for this task performed poorly, possibly because they were designed to perform well on the previous benchmarks. On the other hand, GPT-3 and Wordtune surpassed humans' ability to remember synonyms, indicating these systems could already augment human recall capabilities. Still, none of them outperforms humans' ability to judge the appropriateness of synonyms, given a list of options.

This work demonstrated **the need to build benchmarks with an understanding of the collaborative roles of humans and AI**. Building such benchmarks provides a mechanism to incentivize experts to build AI that can complement humans, and ultimately improve human-AI interaction.

Leveraging AI to support humans

To develop AI systems that can augment human capabilities, AI experts must identify human needs and extend AI capabilities to provide desirable assistance. However, human needs can be latent, or change over time. Even when the needs are explicit, there could be a mismatch between the ways humans and models operate. How can we address these challenges and support such needs?

Extending AI's capability to support humans. Most LMs are trained to predict what the next word should be, thereby generating left to right. However, humans do not always write sequentially; they iteratively go back and edit their writing. Given this, how can we take existing LMs and make them support the editing process?

I proposed **a simple and computationally efficient way to enable pre-trained LMs to fill in the blanks** (i.e., infill) [Donahue et al., 2021]. Previously, AI experts had to build bespoke models for the task or train models from scratch. Unlike prior work, this approach manufactures training data from plain text and fine-tunes an existing model, which results in a model that can perform infilling while retaining its ability to do language modeling. In the user study, humans had difficulty identifying sentences infilled by our approach as machine-generated.

With this approach, I demonstrated a practical way to extend LM capabilities to support human writing. To my knowledge, this is **the first work to infill with pre-trained LMs**, spawning a new direction of research that led to several follow-up papers. Nowadays, infilling is one of the most widely used modes of writing with LMs. In fact, my work has become a de facto method for infilling: it was adapted by OpenAI and implemented as part of OpenAI Playground.

Modeling human behaviors to train complementary AI. One of the common ways to train a model is to collect data from human annotators and optimize the model to mimic the patterns in the data (i.e., supervised learning). However, there are scenarios where collecting data is impossible or suboptimal. For instance, consider building an autocomplete system that takes keywords as input and generates a full sentence as output. Because humans have a natural tendency to adapt to a system over time (e.g., become more efficient by using fewer keywords), it is challenging to annotate data while accounting for the potential interaction and change in behavior.

To address the challenge, I modeled human behaviors to train complementary AI by **framing the interaction as a cooperative communication game between a human and a system** [Lee et al., 2019]. In this game, there are two competing goals: minimizing effort (i.e., the human wants to type as few keywords as possible) and minimizing error (i.e., given the keywords, the system wants to guess the sentence as accurately as possible). I took a human-centered approach to training a model by formulating these two goals into mathematical objectives and optimizing the model for both objectives. The user study showed that the resulting system generated usable completions 90% of the time, while reducing the time for typing by nearly half, compared to typing full sentences. With this work, I demonstrated **a way to train a model that can take into account changing behaviors of humans in interactive settings**. Such training approaches can enable us to build systems that are robust to human adaptation.

Future research

I aim to further spur the development and evaluation of AI systems that can interact with humans and augment human capabilities. Meanwhile, I aspire to understand how AI systems will change the way we write and communicate. Concretely, I envision pursuing the following lines of research in the next 3-5 years.

Writing in the real world. I have explored various writing contexts in controlled settings [Lee et al., 2019, Donahue et al., 2020, Lee et al., 2021, Lee et al., 2022a, Lee et al., 2022b]. However, much more needs to be done to understand **the impact of long-term use and the challenges humans encounter in the real world**. For instance, when I interviewed copywriters, a journalist, and a script writer, it became apparent that they have domain-specific knowledge that is difficult to support with current general-purpose models. Likewise, from organizing the First Workshop on Intelligent and Interactive Writing Assistants, I learned that professional poets and musicians have vastly different writing strategies and preferences than crowd workers. I hope to empower such professionals by taking a holistic approach to writing, encompassing not just language generation, but also other aspects of writing, such as brainstorming and revising.

Homogenization. Great potential comes with great risks. For instance, there is huge potential for increased productivity through the use of LMs for sentence completion. However, **will it homogenize writing outcomes?** In other words, if everyone uses the same LM, **will we see a loss of creativity and individual voice**? What

other risks might arise if human communication becomes highly mediated by AI systems? I want to investigate the effects of LMs on writing, and how we can use them to improve writing quality while preserving individual voice. I plan to build on my work on observing human-LM interaction [Lee et al., 2022a] to answer these questions.

AI in teaching and learning writing. AI for writing is yet to be integrated in education. This year, I provided CoAuthor [Lee et al., 2022a] for Stanford AI and Teaching Writing Project Design Workshop and plan to provide it for teachers at Stanford Online High School next year. Through this, I want to **understand what teachers look for in AI tools**, such as how they want to use AI to interact with students, what content and types of feedback educators find valuable, and how AI can support culturally responsive writing pedagogy. For students, I have an ongoing collaboration with Korea Advanced Institute of Science & Technology (KAIST) that investigates **the impact of AI writing tools on non-native English speakers**. This work aims to shed light on the role of AI writing tools in supporting the needs of language learners.

Writing in the economy. Little is known about the long-term economic impact of LMs. Although I have discussed the potential impact of LMs on the economy in terms of productivity, wage inequality, and centralization [Bommasani et al., 2021], I hope to **obtain a quantitative understanding of what types of writing tasks will be most affected, how they will be affected, and what the economic implications are**. As part of an ongoing collaboration with Stanford Digital Economy Lab, I have been conducting randomized experiments to compare performance between human-only, machine-only, and human-machine collaborative settings. I will continue these field experiments to understand how the economy changes with the rapid advancement of AI.

References

[Lee et al., 2022a] **Mina Lee**, Percy Liang, and Qian Yang. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *Conference on Human Factors in Computing Systems (**CHI**) 2022*.

[Lee et al., 2022b] **Mina Lee**, Megha Srivastava, Amelia Hardy, John Tickstun, 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. *Preprint*.

[Lee et al., 2021] **Mina Lee**[*], Chris Donahue[*], Robin Jia, Alexander Iyabor, and Percy Liang. Swords: A Benchmark for Lexical Substitution with Improved Data Coverage and Quality. In *North American Chapter of the Association for Computational Linguistics (**NAACL**) 2021*.

[Donahue et al., 2020] Chris Donahue, **Mina Lee**, and Percy Liang. Enabling Language Models to Fill in the Blanks. In *Association for Computational Linguistics (**ACL**) 2020*.

[Lee et al., 2019] **Mina Lee**, Tatsunori Hashimoto, and Percy Liang. Learning Autocomplete Systems as a Communication Game. In *Neural Information Processing Systems (**NeurIPS**) Workshop on Emergent Communication* 2019.

[Bommasani et al., 2021] Rishi Bommasani and 113 others, including **Mina Lee** (§2.5 Interaction: Joon Sung Park, Chris Donahue, **Mina Lee**, Siddharth Karamcheti, Dorsa Sadigh, and Michael Bernstein; §5.5 Economics: Zanele Munyikwa, **Mina Lee**, and Erik Brynjolfsson). On the Opportunities and Risks of Foundation Models. *Preprint*.