My research goal is to develop language models that utilize dynamically available information in real time. However, current models often struggle to incorporate new information introduced after pretraining, resulting in hallucinations and misinterpretations of user contexts. Specifically, I am interested in (i) leveraging external documents through **retrieval-augmented generation** and (ii) tailoring model behaviour through **human interaction**. Through these approaches, I aim to equip language models with the capability to address more complex and ever-changing tasks.

Retrieval-Augmented Generation While recent works on retrieval-augmented generation systems have shown promise in enabling language models to stay up-to-date, my research aims to identify failure modes that emerge when these systems adopt naive approaches in complicated scenarios.

While generating responses, retrieval-augmented language models should be capable of utilizing multiple documents that have complementary pieces of knowledge. However, this can lead to **knowledge conflicts** between parametric and non-parametric knowledge or within non-parametric knowledge itself. In my COLM paper, I proposed a question-answering benchmark to assess how language models distinguish between entities sharing the same ambiguous name across multiple documents. I categorized long-form answers into five groups based on accuracy and factuality, enabling a systematic analysis of failure modes across models. For example, I demonstrated that current models often provide incorrect answers, merging information from distinct entities. This issue becomes more prevalent in long-context models, where the potential for conflicts increases with more information in the prompt. In the future, I seek to build retriever-augmented language models that can better comprehend and extract knowledge from **multiple and long documents**.

Meanwhile, the **retriever also acts as a bottleneck**, as retrieval errors propagate to the generation step, hindering optimal performance. Motivated by this challenge, I have focused on building more robust retrievers through task-specific adaptations. In my recent EMNLP Findings paper, I trained task-specific retrievers in a zero-shot setting (i.e., only documents are available) by generating synthetic queries with explicit search intents. This approach eliminates the reliance on few-shot examples from prior methods, which can be costly and unreliable when query-document relationships are not well-defined. In another paper under review, I showed that retrievers can be trained to leverage in-context examples to dynamically adjust their embeddings, analogous to in-context learning in decoder-only models. As a future work, it would be interesting to devise advanced training objectives beyond relevance, such as ensuring diversity of documents or jointly training the retriever and language model with a shared objective.

Once capable language models and powerful retrievers are established, I plan to develop methods for **augmenting the set of retrieved documents**. Rather than simply prepending documents to the prompt, strategic modification could compensate for irrelevant information. For instance, agents could intelligently filter out distractors that confuse language models or refine the content by paraphrasing or adding relevant information. The retrieved set could also be divided into multiple subsets, each optimized for addressing specific aspects of a query, thereby maximizing the model's ability to reason over complex and multifaceted tasks.

Interactive Language Models Humans now interact with large language models (e.g., ChatGPT) more naturally and frequently than ever before. This presents a huge potential for models to become more personalized and better suited to individual needs.

I am deeply interested in understanding the cases in which **people exhibit different interaction styles** and the challenges they encounter. Current large language models are built to cater to a single, generalized user, limiting their ability to accommodate diverse users. My research experience in HCI has prepared me to design and conduct thorough human experiments, including iterative pilot studies and qualitative analyses. During my internship, I conducted user studies examining how people with varying language proficiency interact differently with language models during collaborative writing tasks. I found that non-native speakers were more likely to request and accept model-generated drafts without careful consideration, whereas native speakers tended to collaborate and brainstorm more actively with language models. These real-world observations of such discrepancies have inspired me to design language models that can provide tailored responses, **adapting to individual interactions**.

One possible solution is to make language models **anticipate user needs and take proactive actions**. For instance, models could ask clarifying questions to resolve ambiguous queries, narrowing the answer space rather than providing all possible options. In my ongoing project, I am exploring this problem using real-world Korean insurance data. Since insurance terms are highly sensitive to subtle variations in user queries, I am developing a clarification module that precisely determines when and what to ask. Another proactive approach could involve models suggesting contextually relevant follow-up questions, guiding users to explore additional helpful information. I also see great potential in language model-based agents, which can automate decision-making to identify user goals and execute versatile actions using external tools.

Another promising direction lies in language models **learning user-specific information through feedback**, progressively aligning responses over multiple turns. Latent features in feedback, such as preferences and personas, provide valuable clues for tailoring responses to user intent. However, these intents are often not explicitly stated, making it challenging for models to determine appropriate responses. Looking forward, I plan to enhance language models to recognize and learn user intents effectively through multi-turn interactions.

Why Princeton? The Ph.D. program at Princeton is a great fit for me to accomplish my research goals. I am eager to collaborate with Professor **Danqi Chen** to advance the knowledge acquisition and representation of language models through retrieval-augmented generation. I am also interested in working with Professor **Karthik R. Narasimhan** to explore how language models can interact with individuals, enabling dynamic operation and personalization. With Professor **Sanjeev Arora**, I hope to work on deep learning methods for effective language model adaptation, including in-context learning and synthetic data generation.

After completing my Ph.D., I aspire to become a professor and continue conducting research in academia. Inspired by the mentorship I have received, I am committed to fostering the growth of future scholars and cultivating their research skills. With my strong preparation and clear research vision, I am confident in my ability to conduct innovative and impactful research in natural language processing, and I look forward to contributing to Princeton NLP community.