My research focus revolves around reasoning and acting (a.k.a., two "different" behaviors) of personalized language agents, designed to seamlessly unifing them from tool perspective such as regarding reasoning as **internal cognitive tools** while acting as **external physical tools** instead of treat them in isolation. My long-term objective is to achieve the "impossible triangle" between safety, personalization and autonomy of language agent. To this end, my research seeks to build more helpful, harmless and personalized dialogue agent from the following angles.

- **Dialogue System** — Dialogue, or conversation, has always been one of fundamental interaction between humans and between humans and AI. Therefore, dialogue systems aim to build intelligent conversational agents that generate helpful, harmless and honest responses, e.g., ChatGPT.
- **Tool Learning** — Everything can be seen as a tool. While much focus is on external physical tools (e.g., models, search engines), cognitive tools rooted in cognitive science play an equally or even more critical role. These cognitive tools, such as meta-reasoning and mental models, represent the fundamental reasoning modules of the human mind used to navigate complex environments. By bringing together the strengths of physical and cognitive tools, we can create intelligent systems that bridge the gap between natural and artificial minds, advancing the capabilities of both.
- **Language Agents** — As conversational agents are increasingly integrated into real-world applications, ensuring their safety and personalization – especially as they become more automated – emerges as a critical consideration.

To build a powerful and unified intelligent dialogue model capable of solving diverse real-world tasks, one promising approach is tool learning. As dialogue systems and backbone language models evolve, the boundaries between tasks, models, and even languages are likely to blur. A fundamental capability of such models will be their ability to mimic human cognitive processes and interact effectively with external environments. This raises the critical question of how models can employ cognitive tools internally and physical tools externally. The integration of these tools will be key to creating models that not only reason current state but also act as adaptive problem-solvers in complex real-world contexts.

Internal Cognitive Tools. Cognitive tools refer to specifies a internal cognitive mechanisms that aids systematic or investigative thought (TPE). My research focus on two types of cognitive tools: 1) various conversational strategies such as clarification, hinting, and questioning, play a key role in numerous applications, including tutoring and psychotherapy (Pro-CoT). For example, we could utilize different prompting strategies to reason psychological and emotional state of users and then generate responses (Cue-CoT); and 2) atomic reasoning modules that replicate the reasoning and decision-making processes of the human mind, resulting in o1-like reasoning. We release Open-O1 project (800 stars), which explores how LLMs employ internal cognitive tools like reflection, self-correction, and backward thinking, aiming to achieve System 2 reasoning during inference. I am also open to other insightful ideas from cognitive science to study the underlying cognitive mechanism of LLMs, such as perspective-taking thinking (DualCritique), meta-reasoning theory (AutoPSV).

External Physical Tools. Physical tools refer to external modules that are invoked by a rule or a specific token and whose outputs are incorporated into the context of an augmented language model (LM). These tools include search engines, APIs, databases, robots, and other task-specific external modules. One of typical usages is Retrieval-Augmented Generation (RAG), which leveraging different sources of external knowledge to enrich the contextual information (SAFARI, UniMS-RAG). Moreover, there are several challenges to call these physical tools to fulfill complex user instruction in real-world such as complicated dependency relationship between different function calls, resulting in graph-like execution structure. Another challenge lies in permission management such as which source is authorized to call these tools. To address these challenges, we developed AppBench, a testing platform designed to simulate the iPhone environment.

Frameworks The intergration of internal cognitive tools and external physical tools does not only stands for a novel perspective to build powerful and unified language models, but also demonstrates an robust and flexible framework to combine the intrinsic capabilities of models and functions provided by external environments. On the one hand, I focus on management and unification of these tools by defining different tools as different functions (SelfDC). In detail, SelfDC will decide the action according to confidence signals of used LMs. If the confidence is too low, it will reply on external functions, if it is too high, it will use internal reasoning. On the other hand, there maybe conflicts between internal cognitive tools and external physical tools, i.e., hallucination issue. We provide an comprehensive survey about different types of knowledge conflict (KnowledgeConflicts) and propose SAE-based representation engineering method to control different behavior among them (SpARE). I would like to follow this path to explore more effective and efficient method to combine internal cognitive tools and external physical tools.

Future – To Be Explored I would like to follow this path and explore more effective and efficient ways to build safe, personalized and automated language agent. It is believed that this journey requires not only insights from multiple disciplines, such as data science and cognitive science, but also a novel learning methods from various behavioral data such as dialogues or tool interactions.

- **Dialogue System.** Dialogue systems do not exist in a vacuum; they interact with users whose behaviors, preferences, and needs evolve over time, making it a wonderful testbed to 1) investigate and learn human behaviors, i.e, emotion and psychological states; 2) mimic human behaviors, resulting in human-like perception and reasoning (i.e., memory and meta-reasoning) in different workflows.

- **Tool Learning.** My interest in internal cognitive tools and external physical tools extends beyond dialogue systems to broader applications. In particular, both of these two types of tools can be considered as two different level of behaviors: internal cognitive behaviors and external physical behaviors. I am excited about how dialogue systems (or LLMs) can learn from behavioral data, a.k.a, tool interactions, to refine not just their language abilities, but also their capacity to reflect on and correct their reasoning / acting.

Others I would like to start my postdoctoral scholar on **Aug 2025** after my graduation from The Chinese University of Hong Kong.