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.

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.