Statement of Purpose

Haofei Yu *haofeiy@cs.cmu.edu*

We are living in a multisensory world, encompassing diverse real-world modalities like thermal, touch, vision, audio, text, and movement. In recent times, there has been remarkable growth within two key research communities: the natural language processing community, which focuses on language modality, and the multimodal community, which focuses on integrating audio, language, and visual modalities. Both large language models and multimodal models are pivoting from the objective of improving on benchmarks to delivering better performance in real-world interactive tasks. Concurrently, the scope of research is expanding from the analysis of individual models to the construction of complex systems. These systems, likely to be called agents, are composed of multiple interconnected components, such as tools, the internet, and memory. Based on that, my research interest is in **crafting and training a multisensory agent for real-world applications**. My ultimate research goal is to connect abstract exploration of consciousness theory with a pragmatic AI agent framework and build an agent capable of social intelligence, digital intelligence, and physical intelligence. Moving toward the realization of such an agent, two critical challenges are poised:

(1) How to design a training framework for existing language agents to push real-world performance to a useable level?

(2) How to design a multisensory agent architecture that is capable of handling all types of real-world tasks?

I am fortunate to have conducted some initial research in these areas under the guidance of Prof. Yue Zhang, Prof. Zhenzhong Lan, Prof. Graham Neubig, Prof. Ruslan Salakhutdinov, and Prof. Louis-Philippe Morency. Additionally, my industrial experience in **Apple** and **Tencent** pushes me one step forward to achieving my goal. I am passionate to continue studying these problems in my Ph.D. Fig 1 explains my research line related to the agent and my current progress on that.

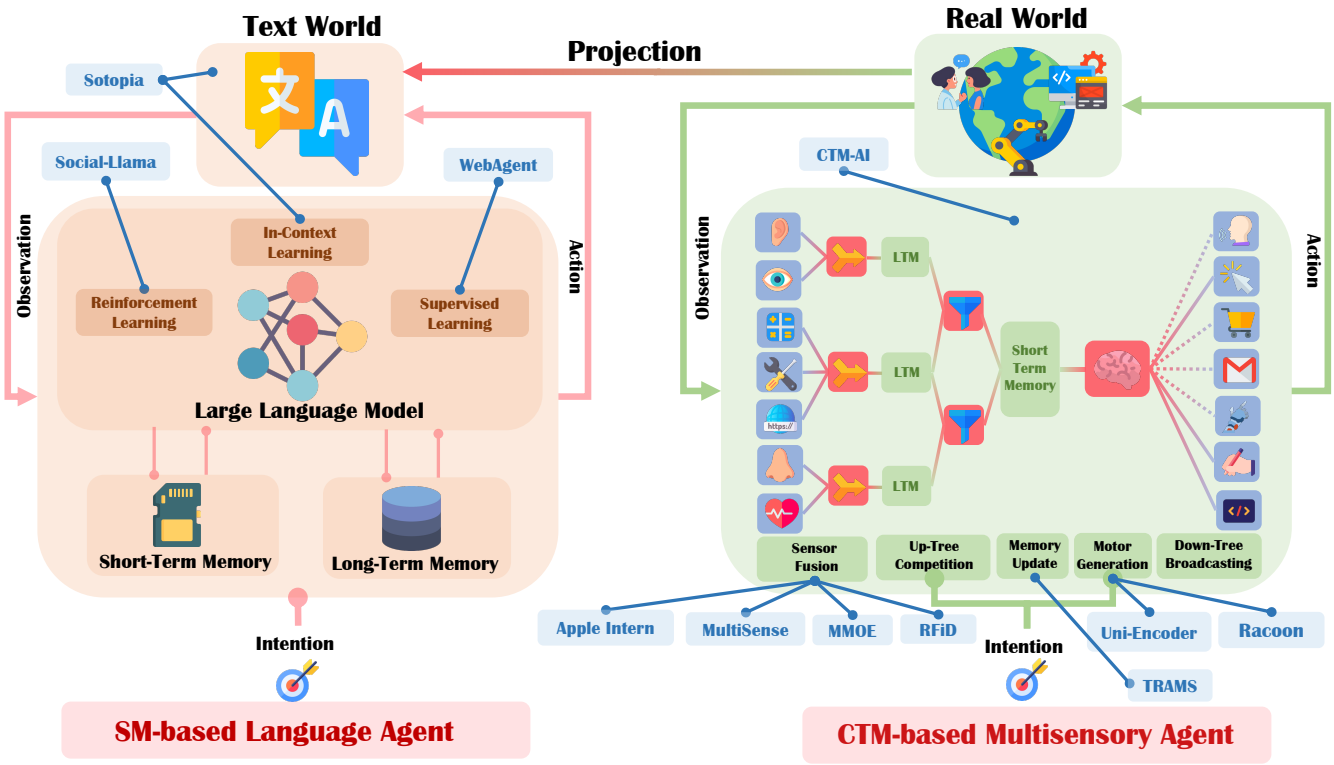

Figure 1: Research Line and current progress on two topics: (1) language agent training on existing agent architecture based on Socratic Model (left) (2) multisensory agent architecture motivated by Consciousness Turing Machine (right). Both accepted work and on-going work are labeled in blue.

Current Progress on Language Agent Training Framework

Currently, a lot of language agents have been proposed [3]. However, most of them are based on in-context learning with an absence of defined training approaches for their agent task. Apart from in-context learning, as illustrated in the left side of Fig 1, my current progress proves that **both supervised training and RL-based training can boost the performance of agent on target task** including agent-agent social interaction and agent-environment web navigation.

Part1.Supervised training helps agents explicitly learn sub-task skills. The fundamental premise of supervised training is to split the agent task into sub-parts, collect data for each sub-task, and train agents on specific skills. Collaborating with Prof. Ruslan Salakhutdinov, I sought to put this insight into the web navigation task [**?**]. By dissecting the overarching web navigation objective into relevant sub-steps, we were able to utilize collected human trajectory data for supervised training of the web navigation agent. This setup allowed agents to automatically plan and engage different sub-tasks throughout a navigation action sequence, ultimately achieving a higher success rate across extended action sequences. This endeavor further underscores the potential of expert-based training in enhancing an agent's efficacy in interactive tasks.

Part2.RL-based training helps agents implicitly learn generalizable skills. While supervised training equips agents with specialized skills through focused training, acquiring high-quality supervised training data remains a challenge, particularly for real-world agent tasks where data collection can be arduous. An alternative approach to enhance agents' generalizable skills entails employing an observation-feedback loop based on rewards during interaction with the environment. A significant advantage of this method is the potential for virtually infinite training data through reinforcement learning. Inspired by this prospect, I contributed to the development of SOTOPIA [6], an innovative evaluation framework for social interaction assessment, submitted to ICLR 2024. Based on the SOTOPIA benchmark, I further demonstrated that reinforcement-learning-based self-training among agents, guided by a GPT-4-based reward function, could enhance general social abilities across diverse social tasks such as negotiation and collaboration. Furthermore, this training approach showcased commendable generalization to unseen social scenarios [**?**].

Current Progress on Multisensory Agent Architecture

To build a multisensory agent that can tackle diverse real-world input and output, a new architecture needs to be developed. The Consciousness Turing Machine (CTM) model [1], motivated based on theoretical computer science, provides one possibility. Inspired by that, I contributed to the proposal of the pragmatic CTM-AI architecture [**?**] that targets handling diverse real-world modalities. **CTM-AI includes 5 stages: sensor fusion, up-tree competition, memory update, motor ranking, and down-tree broadcasting.** Both the input and output of CTM-AI architecture should include diverse processors with 100+ types. More details related to CTM-AI will be discussed.

Part1.Multimodal sensors can be fused in multiple interaction types. The idea of sensor fusion is to discuss how to handle interaction between information from multiple processors. For **unimodal** sensor fusion, I proposed RFiD [2] and proved that adding contrastive learning loss between multiple sensors improves the cross-document understanding of the Fusion-in-Decoder model. This discovery laid a foundation for my subsequent internship at Apple, where I proposed a hierarchical prompting framework that groups and combines outputs of multiple LLMs. This enables it to have more comprehensive information and provide a more satisfying answer to ambiguous and challenging questions. Extending it to **multimodal** data, the multimodal fusion process is more complicated. The interaction between different modalities can be classified into agreement, disagreement, synergy, redundancy, and uniqueness. Therefore, I propose MMOE, a multimodal mixture of experts, to help separately handle different types of multimodal interaction and get performance gain on the Mustard dataset. To push one step further, I help propose a multisensory foundation model [**?**] that incorporates heterogeneous data including IoT, vision, language, and music into single language model outputs based on multimodal adapters that utilize representation-based fusion.

Part2.Motor ranking and generating can be learned simultaneously. A multisensory agent should not act on all outputs to the environment and there should be a motor ranking stage for the final output. An intuitive solution was to incorporate an additional ranker during generation. In collaboration with Prof. Zhenzhong Lan, I contributed to the design of Uni-Encoder [4], which proficiently handled response ranking, showcasing exemplary performance on Ubuntu v1, Ubuntu v2, PersonaChat, and Douban. This success led to a natural inquiry: could ranking information be integrated directly without needing an extra ranker? To explore this, I worked with Prof. Graham Neubig and proposed Racoon [**?**] that focuses on code generation tasks and proves that under the guidance of contrastive ranking loss, a generative model could be trained to accomplish ranking and generating in a multi-task fashion.

Part3.Short-term memory updating can be done automatically. Following the research on the input and output facets of the agent framework, the central part of CTM-AI is the short-term memory that is dynamically updated at each timestep. There should be an automatic, intension-independent memory update strategy for this. Based on this, I found that performing a query-independent automatic memory selection operation can help improve the overall performance on language modeling benchmarks including character-level enwik8 and word-level wikitext-103. This endeavor, termed as training-free memory selection (TRAMS) [5], offers valuable insights into executing an intention-independent automatic memory selection process for effective working memory management.

Future Plan on Multisensory Agent

There are three main future works I want to focus on discussing during my Ph.D. to build a better multisensory agent for real-world applications:
(1) how to scale up the multisensory architecture to handle all real-world modalities
(2) how to make such a multisensory agent trainable when given a typical task
(3) how to incorporate the world model into the current multisensory agent architecture
For the first question, I hope to explore the possibility of crafting agents to do complicated multimodal robotics or social tasks. For the second question, I believe a better training method can be proposed for the training of a complicated system including multiple components. For the final question, the importance of the world model should be emphasized.

Why School

References

[1] Lenore Blum and Manuel Blum. A theory of consciousness from a theoretical computer science perspective: Insights from the conscious turing machine. *Proceedings of the National Academy of Sciences*, 119(21):e2115934119, 2022.

[2] Cunxiang Wang*, **Haofei Yu***, and Yue Zhang. Rfid: Towards rational fusion-in-decoder for open-domain question answering. *arXiv preprint arXiv:2305.17041*, 2023.

[3] Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023.

[4] Chiyu Song, Hongliang He, **Haofei Yu**, Huachuan Qiu, Pengfei Fang, and Zhenzhong Lan. Panoramic-encoder: A fast and accurate response selection paradigm for generation-based dialogue systems. *arXiv preprint arXiv:2106.01263*, 2021.

[5] **Haofei Yu***, Cunxiang Wang*, Yue Zhang, Wei Bi, et al. TRAMS: Training-free memory selection for long-range language modeling. *arXiv preprint arXiv:2310.15494*, 2023.

[6] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, **Haofei Yu**, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*, 2023.