After completing my bachelor's degree, I decided to take a gap year to explore my interests and career aspirations. I have been working as a research intern at Microsoft Research Asia for more than a year, where I have found my true passion for research. I enjoy dedicating myself to research projects and bringing them from conception to completion. While 95% of the research journey is challenging and sometimes frustrating, it gives me a constant sense of progress. The remaining 5%—the exciting moments of breakthroughs and rewards—drive me to embrace the next challenge. Research has given me a deep sense of fulfillment and a clear vision for my future.

My research interests primarily focus on (1) **Language Model (LM) Agents in Complex Interactive Environments** and (2) **Data Synthesis and Data Selection**. In the following sections, I will elaborate on my contributions and future aspirations in these areas.

1. LM Agents in Complex Interactive Environments

In my previous work on table-based question answering [1, 2], I observed that LMs perform significantly better with SQL queries than with direct natural language answers. Additionally, in exploring their ability to solve programming competition problems [3], I found that LLMs demonstrate a certain capacity to comprehend complex problem statements, apply algorithmic knowledge, and implement it in code. The efficiency of code in data processing and task-solving reinforced my belief that ***code is one of the most potent tools for language models to effectively manipulate and interact with the world.***

Most existing data analysis benchmarks focus on single-step tasks with no interaction, where data is explicitly provided in the instructions, making these tasks overly simple and unrealistic for real-world applications. To address this gap, I developed a data science benchmark covering tasks such as data wrangling, exploratory data analysis, and machine learning [4]. Each task provides abstract instructions and data stored in the file system, running independently in a sandbox environment. I also introduced an agent framework that integrates Python, SQL, and Bash. LM agents need to use code to interact with Linux environments, make decisions based on goals and feedback, and self-debug using error information. However, even with the most advanced LLMs, only 30% of the tasks can be solved. By analyzing their trajectories, I identified a persistent issue: **the inability of agents to adapt effectively in dynamic environments**. For example, they often fail to revise their code based on feedback, repeatedly producing the same incorrect output without adjustments.

Looking Ahead I wish to focus my future work on creating more efficient, adaptable, and reliable LM agents. Specifically, this includes:

- **Autonomous Learning**: One promising direction is how to enable agents to autonomously adapt to new environments filled with unseen tasks. To achieve this, they must efficiently explore their surroundings, discern meaningful objectives, learn from their experiences, and evolve independently—all of these steps still remain challenging for current agents. To further this research, I am now exploring methods to enhance LM agents' adaptability to diverse tasks and environments by dynamically refining their tool utilization skills through an **evolving tool graph** [5].
- **Safety and Reliability**: Another critical direction is enabling agents to detect malicious actions and ensure they pose no harm to users or society. In digital environments, actions by unreliable agents can lead to data breaches, unauthorized access, and even system failures. Thus, it is imperative to equip agents with the ability to recognize potentially harmful actions. Additionally, exploring effective ways to integrate human supervision to further mitigate safety concerns associated with LM agents is also essential.

2. Data Synthesis and Data Selection

As the saying goes, "You are what you eat." Similarly, in the realm of AI, models are a reflection of what they are trained on. Finetuning on synthetic data has proven effective in improving the instruction-following abilities of LMs. However, synthesizing data for complex problem-solving tasks presents significant challenges in precision, controllability, diversity, and scalability.

To tackle these challenges, I explored two effective **data synthesis frameworks leveraging underlying knowledge and code execution** separately. The first synthesis framework leverages LLMs to generate datasets for mathematical reasoning [6]. Unlike previous approaches that directly mimic existing data—often limiting novelty and diversity—we are the first to utilize the underlying knowledge from existing data to assist in synthesis. Specifically, we extract key points from original problems, analyze their co-occurrence relationships, and recombine them to generate new problems. The second synthesis framework utilizes SQL engines and meticulously predefined rules to generate table-SQL execution tasks [7]. By precisely controlling table properties, answer locations, and the complexity of SQL expressions, it can generate datasets with varying levels of difficulty. This dynamic and scalable data supports the evaluation and enhancement of models' capabilities in symbolic reasoning and long-context comprehension.

Although LLMs have made significant progress in general capabilities, they often struggle with domain knowledge due to the limited availability of high-quality, domain-specific pretraining data. While synthetic data can alleviate data scarcity, it often lacks sufficient diversity and is costly to scale for pretraining. To address this limitation, I am exploring effective methods for **domain-targeted data selection** [8]. For a given datapoint, if the expert model with domain knowledge interprets it better than the general model, it is likely to be more relevant to the target domain. Leveraging the perplexity information from both the base model and the expert model, we can evaluate data relevance for the target domain and select the most suitable data for pretraining, thereby improving LLMs' domain knowledge.

Looking Ahead The learning paradigms of AI fundamentally differ from those of humans, necessitating the development of innovative frameworks for knowledge representation and training data design. I aim to explore data-centric methods to improve learning efficiency and adaptability, leading to a series of profound questions:

- How do we define, combine and re-weight multiple domains of data during pretraining?
- How do we identify "difficult" data for a given model, a crucial step in enabling its iterative self-improvement?
- What is the best way to represent knowledge? How can a concept be grounded across various modalities, such as text, images, videos, and interactive sensory experience?

3. Career Aspirations

My long-term goal is to become a Principal Investigator in an academic research lab, focusing on topics with the potential for lasting impact. I aspire to be an accomplished female computer scientist, making meaningful contributions to the field and inspiring others. I am deeply grateful to my advisors and collaborators for their guidance and support, and I am more than willing to extend the same support to future researchers by mentoring and collaborating with them.

References

[1] Fangyu Lei, Xiang Li, Shizhu He, **Yiming Huang**, Jun Zhao, and Kang Liu. S3HQA: A three-stage approach for multi-hop text-table hybrid question answering. *ACL 2023*.

[2] Tongxu Luo*, Fangyu Lei*, Weihao Liu* **Yiming Huang***, Shizhu He, Jun Zhao, and Kang Liu. TableQAKit: A comprehensive and practical toolkit for table-based question answering. *Preprint*.

[3] **Yiming Huang***, Zhenghao Lin*, Xiao Liu, Yeyun Gong, Shuai Lu, Yelong Shen, Chen Lin, Nan Duan, and Weizhu Chen. Competition-level problems are effective LLM evaluators. *ACL 2024*.

[4] **Yiming Huang***, Jianwen Luo*, Yan Yu, Yitong Zhang, Fangyu Lei, Yifan Wei, Shizhu He, Lifu Huang, Xiao Liu, Jun Zhao, and Kang Liu. Da-code: Agent data science code generation benchmark for large language models. *EMNLP 2024*.

[5] Jianwen Luo*, **Yiming Huang***, Jun Zhao, and Kang Liu. From basic tools to complex skills: Adaptive evolution of tool graphs across diverse tasks. ***To Be Submitted*** *to ACL 2025.*

[6] **Yiming Huang**, Xiao Liu, Yeyun Gong, Zhibin Gou, Yelong Shen, Nan Duan, and Weizhu Chen. Key-point-driven data synthesis with its enhancement on mathematical reasoning. ***AAAI 2025***.

[7] Fangyu Lei*, Qian Liu*, **Yiming Huang***, Shizhu He, Jun Zhao, and Kang Liu. S3Eval: A synthetic, scalable, systematic evaluation suite for large language model. *NAACL 2024*.

[8] **Yiming Huang**, Xiao Liu, Yeyun Gong, and Weizhu Chen. Data selection and refinement for domain-specific continual pre-training. ***To Be Submitted*** *to ICML 2025.*