Hongyi **Du**

Research Assistant · Undergraduate Researcher

905B W Hill St. 61801 Urbana, IL, United States

(+1) 217-778-5314 | hongyid4@illinois.edu

Summary

I am a Junior student at University of Urbana-Champaign(UIUC). I am also an undergraduate research assistant at the Blender Lab, under the direction of Professor Heng Ji. I have over 2 years of experience in developing large language model (LLM) agents, familiar with the complete process of training LLMs and VLMs, including preprocessing, finetuning, postprocessing, evaluating, and benchmarking, etc. Currently, my research focuses on LLM Theory of Mind, Multi-Agent System, and LLM Tool Learning.

Education

University of Illinois at Urbana-Champaign

Illinois, United States

B.S. in Computer Science; GPA:3.83/4.00 — *Sep. 2022 - May. 2026*

- Achieved top grades (A+, A) in rigorous coursework: Advanced Natural Language Processing, Computational Photography, System programming, Algorithms, Numerical Methods, Computer Architecture, Data Structures, Probability and Statistics, and Linear Algebra. **Especially, Earn prefect A+(100/100) in Graduate course Advanced Natural Language Processing.**

Publications

ModelingAgent: Bridging LLMs and Mathematical Modeling for Real-World Challenges

Cheng Qian, **Hongyi Du**, Hongru Wang, Xiusi Chen, Yuji Zhang, Avirup Sil, Chengxiang Zhai, Kathleen McKeown, Heng Ji — *May 2025*

- **arXiv**: arxiv.org/abs/2505.15068

MultiAgentBench: Evaluating the Collaboration and Competition of LLM Agents

Kunlun Zhu, **Hongyi Du**, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, Jiaxuan You — *Mar. 2025*

- **arXiv**: arxiv.org/abs/2503.01935
- **ACL 2025 Main Conference**

EscapeBench: Towards Advancing Creative Intelligence of Language Model Agents

Cheng Qian, Peixuan Han, Qinyu Luo, Bingxiang He, Xiusi Chen, Yuji Zhang, **Hongyi Du**, Jiarui Yao, Xiaocheng Yang, Denghui Zhang, Yunzhu Li, Heng Ji — *Dec. 2024*

- **arXiv**: arxiv.org/abs/2412.13549
- **ACL 2025 Main Conference**

Research Experience

MultiAgentBench: Evaluating the Collaboration and Competition of LLM agents

Champaign, United States

Undergraduate Research Assistant at Blender Lab - Supervised by Prof. Heng Ji and Prof. Jiaxuan You — *Aug 2024 – Feb 2025*

- Proposed and developed a comprehensive benchmark framework for LLM-based multi-agent systems, focusing on core characteristics such as planning, memory, and collaborative abilities across various domains, including coding, gaming (e.g., Minecraft), and research tasks.
- Independently developed a Werewolf game framework (*Werewolf Arena*) using LLM prompting and tools to simulate multi-agent interactions, evaluating the model's second-order theory of mind reasoning, decision-making, and collaboration capabilities.
- Collected and analyzed game data to assess LLMs' abilities such as understanding implicit signals, leadership, cooperation, and learning from others' behaviors in complex game scenarios. Evaluations were performed using in-game performance metrics (e.g., game outcomes, key moments) and detailed chain-of-thought analysis.
- Introduced the Werewolf game as a novel multi-agent evaluation benchmark, providing unique insights into the cognitive and cooperative abilities of LLMs in decentralized social simulations. Notably, discovered that high-intelligence foundation models tend to develop mutual distrust, resulting in internal friction and reduced collaboration efficiency.
- Accepted by **ACL 2025 Main Conference** (co-first author).
- **Code**: github.com/MultiagentBench/MARBLE | **arXiv**: arxiv.org/abs/2503.01935

ModelingBench & ModelingAgent: Real-World Mathematical Modeling Benchmark

Champaign, United States

Undergraduate Research Assistant at Blender Lab — supervised by Prof. Heng Ji — *Feb. 2025 – Present*

- Led development of the **DataAgent** and **SimulationAgent** modules, enabling automatic data acquisition (web/API) and Monte-Carlo/ODE simulation pipelines that ground LLM reasoning in quantitative evidence.
- Implemented task-agnostic **baseline** workflows and reusable **baseline-tools** (web search, Python execution, file I/O) to benchmark GPT-4o, Claude 3, Gemini 1.5, etc. across 30+ ModelingBench tasks.
- Curated and formalized new real-world problems—traffic flow, renewable scheduling, crypto trading, ecosystem planning—into JSON schemas, expanding benchmark coverage by 40%.
- Co-designed multi-dimensional metrics (tool-use efficiency, structural coherence, realism, presentation) and integrated automated logging for reproducible evaluation.
- Submitted to **EMNLP 2025 Main Conference** (second author).
- **GitHub**: github.com/qiancheng0/ModelingAgent | **arXiv**: arxiv.org/abs/2505.15068

Proflie Bench: personalized diologue benchmark

Champaign, United States

Undergraduate Research Assistant at Blender lab - Supervised by Prof. Heng Ji — *Aug. 2024 – Current*

- Developed a benchmark to evaluate LLMs' ability to generate engaging and personalized dialogue content based on user profiles. The benchmark tests whether the model can retain details about the user's profile and recall previous plans or arrangements when mentioned in later conversations.
- Created a profile dataset using GPT-4o with LLM tools and generated dialogues based on these profiles, ensuring realistic and dynamic changes in the conversations to test the model's memory and engagement capabilities.
- Developed performance evaluation methods to assess how well the LLM adapts to changes in user profiles remember user's command long ago, and maintains engagement in the dialogue.

Openhands-Codeact: Code-Execution-based Preference Annotation

Champaign, United States

Undergraduate Research Assistant at Blender lab - Supervised by Prof. Heng Ji — *Feb. 2024 – Aug. 2024*

- Designed and tested a framework to enhance existing LLM's evaluation ability on problems difficult for humans to evaluate.
- Performed prompt engineering and established a code-execution-based agent system to enhance LLMs' proficiency in distinguishing optimal solutions from suboptimal ones, advancing model evaluation capabilities.
- This project secured $5 million in seed funding and received 32k stars on GitHub: https://github.com/All-Hands-AI/OpenHands.

Project Experiences

CS497: Individual Research

Champaign, United States

Researcher - Supervised by Prof. Jiawei Han — *Jul 2024 – Feb 2025*

- Integrated Knowledge Graphs (KG) as retrieval sources into the RAG system, replacing unstructured text.
- Enhanced the LLM's capability to handle QA tasks requiring complex knowledge backgrounds.
- Tested different KG creation methods to enhance the model's ability to handle QA tasks.
- Proposed a novel QA solution paradigm: retrieving relevant article summaries from a large-scale database, then constructing a more targeted KG for each specific QA task, which improves the RAG pipeline's accuracy and reasoning capabilities.
- Earned an A in this individual research course as an undergraduate student, and am currently preparing this work for submission to ACL 2025.

Skills

Languages	Python, C++, Java, MATLAB
Tools	ChatGPT, Git/GitHub, Bash, Linux, SQL Server, AWS, Hugging Face, PostgreSQL, JIRA
Frameworks	PyTorch, Flask, Vue.js, Vits, Whisper, Rapid Paraformer, BERT, OpenCV, Qwen, Yolo, Resnet, OpenHands, XAgent, AgentNet, VillagerAgent, WereWolf
Methodologies	LLM Agent Development, LLM Fine-tuning, LLM Benchmarking, LLM Tool Learning, Chain of Thought Reasoning, Prompt Engineering, Multi-agent Systems Developing and Evaluating, Multi-agent framework construction, data collection, data analysis, paper writing