

Longtian Bao

lobao@uchicago.edu | github.com/LongtianBao | Google Scholar

Education

University of Chicago	Chicago, IL
Ph.D. in Computer Science	2026 – Present
University of California, San Diego	La Jolla, CA
M.S. in Computer Science (GPA: 4.0/4.0)	2024 – 2026
B.S. in Computer Engineering (GPA: 3.8/4.0)	2020 – 2024

Publications

* denotes equal contribution.

- [1] **Longtian Bao***, Jianyou Wang*, Yang Zhang*, Youze Zheng*, Ramamohan Paturi. “Question Begets Question: Self-Evolving Curriculum for Reinforcement Fine-Tuning on Competition Mathematics.” *arXiv preprint arXiv:2608.01522*, 2026. [Paper]
- [2] Zhongkai Yu, Jinya Jiang, **Longtian Bao**, Junxia Cui, Jingyi Zhang, Zhengan Cheng, Zaifeng Pan, Jingbo Shang, Yufei Ding. “Multi-Agent LLM Serving Is a Full-Stack Problem: A Small Step Towards Algorithm–System Co-Design.” *Under review*, 2026.
- [3] Youze Zheng*, Jianyou Wang*, Yuhan Chen, Matthew Feng, **Longtian Bao**, Hanyuan Zhang, Maxim Khan, Aditya K. Sehgal, Christopher D. Rosin, Umber Dube, Ramamohan Paturi. “DeepImagine: Learning Biomedical Reasoning via Successive Counterfactual Imagining.” *arXiv preprint arXiv:2604.23054*, 2026. [Paper]
- [4] Jianyou Wang*, Youze Zheng*, **Longtian Bao**, Hanyuan Zhang, Qirui Zheng, Yuhan Chen, Yang Zhang, Matthew Feng, Maxim Khan, Aditya K. Sehgal, Christopher D. Rosin, Ramamohan Paturi, Umber Dube, Leon Bergen. “CT Open: An Open-Access, Uncontaminated, Live Platform for the Open Challenge of Clinical Trial Outcome Prediction.” *arXiv preprint arXiv:2604.16742*, 2026. [Paper]
- [5] Weili Cao*, Jianyou Wang*, Youze Zheng*, **Longtian Bao***, Qirui Zheng, Taylor Berg-Kirkpatrick, Ramamohan Paturi, Leon Bergen. “Single-Pass Document Scanning for Question Answering.” *Conference on Language Modeling (COLM)*, 2025. (**Oral Spotlight, Top 2%**). [Paper] [Code]
- [6] Jianyou Wang*, Weili Cao*, **Longtian Bao**, Youze Zheng, Gil Pasternak, Kaicheng Wang, Xiaoyue Wang, Ramamohan Paturi, Leon Bergen. “Measuring Risk of Bias in Biomedical Reports: The RoBBR Benchmark.” *Empirical Methods in Natural Language Processing (EMNLP)*, 2025. [Paper] [Code]

Research Experience

Ph.D. Student — Department of Computer Science, University of Chicago	Chicago, IL
Reinforcement fine-tuning for LLM reasoning; multi-agent LLM serving	Mar. 2026 – Present
• Designed Question-begets-Question (QbQ), a scalable data-generation procedure in which a teacher model transforms competition-math problems into diverse variants probing the same underlying skills, enabling reinforcement fine-tuning from problem statements and final answers alone — no oracle reasoning traces [1]. • Proposed a self-evolving curriculum that re-seeds QbQ each round from problems the current checkpoint can mostly solve; under an identical data budget this breaks the static-training ceiling, lifting Qwen2.5-Math-7B from 5.6% to 16.5% pass@1 on AIME with no saturation after 20 rounds [1]. • Introduced a unified typed-graph representation that casts multi-agent LLM systems as analyzable serving workloads, and developed static runtime patterns and dynamic accuracy–efficiency strategies achieving up to a 5× speedup while using 54–89% fewer tokens on real GPUs — while showing that decode dominance means prefill-only wins do not survive end to end, making phase-aware algorithm–system co-design essential	

[2].

Research Assistant — Laboratory for Emerging Intelligence, UC San Diego

La Jolla, CA

Advisors: Prof. Leon Bergen and Prof. Ramamohan Paturi

Apr. 2024 – Mar. 2026

- Co-developed CT Open, an open-access, uncontaminated, live platform for the open challenge of clinical trial outcome prediction, enabling rigorous evaluation of LLMs on prospective (post-training-cutoff) trials [4].
- Contributed to DeepImagine, which teaches LLMs biomedical reasoning by training them to infer how observed clinical-trial results change under controlled perturbations (dosage, outcome measures, study arms), combining supervised fine-tuning on strict counterfactual pairs with RL from verifiable rewards [3].
- Re-architected a Mamba-2 State-Space Model with a classification head, creating a subquadratic model that preserves global context for efficient sentence retrieval in documents up to 256k tokens, and designed a link-based synthetic data pipeline exploiting natural long-range dependencies for long-context QA training [5].
- Constructed a fine-grained sentence-retrieval benchmark from 41 existing QA datasets via an LLM-assisted annotation-validation pipeline [5].
- Developed the RoBBR benchmark for risk-of-bias assessment in biomedical research from expert judgments on 500+ studies, designing subtasks that disentangle the retrieval and reasoning abilities of language models and building a human-validated GPT-4 annotation pipeline for ground truth [6].

Teaching Experience

Instructional Assistant — UC San Diego

Spring 2021 – Winter 2024

CSE 12 (Data Structures), CSE 11 (Accelerated Intro to Programming), CSE 30 (Computer Organization), CSE 15L (Software Tools & Techniques), across eight selected quarters.

Projects

UCSD Autograder Platform

2023 – 2024

Engineered and maintained a full-stack educational platform serving 1,000+ students across multiple departments, with a containerized autograder, a real-time ticket-based office-hours queue, and a database holding millions of rows of student data.

Skills

Programming: Python (advanced), C/C++, Java, JavaScript/TypeScript, Bash

ML & NLP/IR: PyTorch, Hugging Face Transformers, reinforcement fine-tuning (GRPO), State-Space Models (Mamba-2), long-context QA & sentence retrieval, LLM evaluation & benchmarking

Systems & DevOps: Docker, Linux/SSH, Git/GitHub, AWS (EC2/S3), Distributed Training, Weights & Biases

Backend & Data: Flask/FastAPI, PostgreSQL, REST APIs, React

LLM Tooling: OpenAI/Anthropic/Google APIs