

Yucheng Du

Los Angeles, CA | (626) 654-9800 | yuchengdu.tech@gmail.com
linkedin.com/in/bryandu001 | github.com/yucheng-du

EDUCATION

University of Southern California

M.S. in Computer Science | Graduate GPA: 3.85/4.00

Los Angeles, CA

Jan 2026 - Expected Dec 2027

University of California, Los Angeles

B.A. in Linguistics & Computer Science | Honors: Dean's List

Los Angeles, CA

Sep 2021 - Jun 2023

PUBLICATIONS

Yucheng Du, Xiyang Hu. "Recognition-Refusal Misalignment in LLMs: Why Models Answer Structurally Unanswerable Questions." To appear in Proceedings of EMNLP 2026 (Main Conference). [arXiv:2608.29109](https://arxiv.org/abs/2608.29109) | [code and data](#)

Yucheng Du. "Geometric Deviation as an Unsupervised Pre-Generation Reliability Signal: Probing LLM Representations for Answerability." Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026), ACL 2026. [ACL Anthology](#) | [arXiv:2605.03196](https://arxiv.org/abs/2605.03196)

RESEARCH EXPERIENCE

Graduate Researcher | FORTIS Lab, University of Southern California

Mar 2026 - Present

Recognition-Refusal Misalignment in LLMs

EMNLP 2026 (Main Conference)

Mentored by Xiyang Hu, Assistant Professor, Arizona State University

- Led the project from hypothesis formation through implementation, experiments, analysis, and writing, studying why LLMs answer structurally impossible questions using a 4,800-prompt matched benchmark.
- Developed a low-capacity residual-stream probe that detected pre-generation impossibility signals with 0.939 mean AUC on held-out prompts across 11 open-weight LLMs (1.7B-70B) and 22 math/code settings.
- Separated structural-impossibility recognition from the canonical safety-refusal direction (mean cosine 0.087) and established causal relevance; bidirectional activation steering produced 33-52 pp shifts beyond matched random controls in the strongest structural settings, supporting a routing-failure account.

Ongoing Project | AI Auditing

Jun 2026 - Present

Supervised by Yue Zhao, Assistant Professor, University of Southern California

- Leading implementation, experimentation, and analysis for an ongoing project auditing AI execution traces.

Independent Research | LLM Reliability via Representation Geometry

Jan 2026 - Jul 2026

- Developed an unsupervised, single-pass pre-generation answerability signal from hidden-state geometric deviation; achieved ROC-AUC 0.78-0.84 across Llama 3.1-8B, Qwen 2.5-7B, and Mistral-7B-Instruct.
- Built matched math/fact/code experiments, 5,000-permutation and layer-wise analyses, and reproducible Python/PyTorch/Hugging Face pipelines; released [code and data](#) on GitHub.

PROFESSIONAL EXPERIENCE

Huachuan Die Casting | Digital Solutions Lead

Dec 2023 - Dec 2025

Sichuan, China

- Led requirements, system architecture, and technical direction for a REST API-based factory-wide manufacturing execution system (MES) spanning production, quality, equipment, inventory, and piece-rate payroll; guided a team of two full-stack engineers through implementation and UAT.
- Launched the platform in Dec 2025 for 100+ internal users across 10 departments; it averaged 16K+ shop-floor completion records and 1.8M+ adjusted production units per month over its first seven production months.

ByteCommerce LLC | Founder & Computer Science Specialist

May 2023 - Nov 2023

Los Angeles, CA

- Led educational software projects and taught Python (UCLA CS 30) to student clients.

TECHNICAL SKILLS

Programming & CS: Python, C++, Java, SQL; Data Structures & Algorithms, Object-Oriented Programming

Machine Learning & Artificial Intelligence: Large Language Models (LLMs), model evaluation, statistical modeling

ML Frameworks: PyTorch, Hugging Face Transformers, scikit-learn, NumPy, pandas

Systems & Tools: REST APIs, PostgreSQL, Redis, MQTT, Git, Linux/Bash; AI coding tools (Claude Code, Codex, Cursor)