Youze "Hargen" Zheng

hargenzheng.com | hargen.zheng@gmail.com | github.com/hgnzheng | Google Scholar

Education

University of California, San Diego (UC San Diego) *Expected Jun. 2026*

B.S. Data Science (Major GPA: 4.00) · B.S. Mathematics-Computer Science (Major GPA: 3.97) *La Jolla, CA*

Graduate-level Coursework: LLM Systems, Deep Learning, ML Algorithms, ML for Music, Convex Optimization

Publications

*denotes equal contribution.

[1] 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]

[2] 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

Research Assistant | Laboratory for Emerging Intelligence, UC San Diego *Apr. 2024 – Present*

Advisors: Prof. Leon Bergen and Prof. Ramamohan Paturi *La Jolla, CA*

- Trained a 1.3B-parameter Mamba-2 for long document retrieval via novel synthetic data; outperformed strong embedding baselines and approached full-context LLMs at far lower computational cost [1]
- Co-developed RoBBR, an expert-grounded risk-of-bias benchmark disentangling LLMs' retrieval and reasoning capabilities; Llama-3-8B fune-tuning yielded consistent gains across subtasks [2]

Student Researcher | Halıcıoğlu Data Science Institute, UC San Diego *Sep. 2025 – Present*

Advisors: Prof. Hao Zhang *La Jolla, CA*

- Senior Capstone: "Open LLM Training, Inference, and Infrastructure" (On-going)

Student Researcher | Faculty Mentor Program, UC San Diego *Oct. 2023 – Apr. 2024*

Advisor: Prof. Justin Eldridge *La Jolla, CA*

- Led a 2-person project on EEG-based depression classification using mined discriminative features and presented a poster at UC San Diego's 2024 Online Undergraduate Research Symposium

Teaching

Teaching Assistant, CSE 151A: *Introduction to Machine Learning* *Spring 2025*

Teaching Assistant, CSE 151B: *Deep Learning* *Winter 2025*

Instructional Assistant, LIGN 167: *Deep Learning for Natural Language Understanding* *Fall 2024*

Teaching Assistant, DSC 10: *Principles of Data Science* *Fall 2024*

Teaching Assistant, DSC 20: *Data Structures for Data Science* *Fall 2023, Winter/Spring/Summer 2024*

Selected Projects

Language Model Training, Inference, and Alignment | Stanford CS 336

- Implemented byte-level BPE tokenizer and a pre-norm Transformer; trained on TinyStories and OpenWebText
- Built a math-reasoning pipeline on MATH with batched vLLM generation; explored r1-style zero-shot prompting, SFT on chain-of-thought, Expert Iteration (reward-filtered sampling), and GRPO
- Ran instruction-tuning and DPO on Llama-3.1-8B; evaluated on MMLU, GSM8K, and SimpleSafetyTests

ML Systems Implementation & Optimization | UC San Diego CSE 234

- Wrote autodiff operators and a graph executor from scratch (NumPy and limited PyTorch functions)
- Engineered fused kernels (MatMul+LayerNorm and MatMul+Softmax operators) and a Triton epilogue computing $D = \text{ReLU}(A@B + C)$ with tiling/cooperative fetching; achieved 1.51× speedup vs PyTorch baseline on T4 GPU
- Derived optimal scaling-law parameters under fixed budgets and GPU options; analyzed DeepSeek-V3 training cost

Applied NLP Mini-Projects

- Intent classification on Amazon MASSIVE with SimCSE/SupCon and classification head with BERT-330M
- MBTI personality classification with BERT-110M; deployed with Streamlit and demoed during showcase
- ABC-notation music generation with LSTM; compared against RNN and ran ablation on hidden neurons sizes

Honors & Awards

Conference Travel Funding, Halıcıoğlu Data Science Institute — *Jul. 2025*

Conference Travel Support, Sixth College at UC San Diego — *Jul. 2025*

Provost's Honors — *Fall 2022 - Spring 2025*

Academic Services

Community Mentor for Deep Learning Specialization, DeepLearning.AI — *Jan. 2024 – Apr. 2024*

Student Mentor, Mentor Collective Program, UC San Diego — *Sep. 2023 – Mar. 2024*

Note-Taking Volunteer, Office for Students with Disabilities, UC San Diego — *Aug. 2022 – Dec. 2023*

Technical Skills

Languages/Frameworks: Python, PyTorch, Triton, Transformers, vLLM, PostgreSQL, PySpark

NLP/IR: Stanza, NLTK, SentencePiece, OpenAI/Anthropic/Google APIs (sync/async multithreading requests)

Systems/Tooling: Distributed training, Weights & Biases, Docker, AWS (EC2/S3), Linux, SSH, Cursor