Zhengbang Yang

zyang30@gmu.edu | Zhengbang-Yang/ | in/zhengbang-yang | github.com/Zhengbang-Yang

Research Interests

LLM Alignment: reinforcement learning, model unlearning of undesirable knowledge, and automated red-teaming to identify model vulnerabilities
Agentic AI: systematic design and evaluation of chatbot systems

Education

George Mason University — Aug 2024 – May 2029 (expected)
Ph.D. in Information Technology; Advisor: Dr. Zhuangdi Zhu — *GPA: 4.00/4.00*

University of California - Irvine — Sep 2022 – Dec 2023
Master of Data Science; Advisor: Dr. Weining Shen — *GPA: 3.81/4.00*

Shenzhen University — Sep 2018 – Jun 2022
B. Eng. in Computer Science & B. Econ. in Finance / Outstanding Graduate — *GPA: 3.65/4.00*

Professional Experience

LLM Unlearning — Jun 2025 – Present
NAIRR PILOT Project (PI: Zhuangdi Zhu); First author of the accompanying paper. — *Fairfax, VA*

- Proposed a novel unlearning algorithm for enhancing the **privacy and safety of LLMs**, without relying on **pretraining-data**, which enables fine-grained and robust removal of undesirable knowledge.
- Conducted comprehensive experiments on state-of-the-art LLM unlearning benchmarks, demonstrating superior trade-offs between knowledge forgetting and utility preservation compared to existing methods.
- Designed and analyzed ablation studies to identify the contributions of our algorithm design, showing enhanced **robustness** to fine-tuning data scarcity, heterogeneous data formats.

DUET: Distilled LLM Unlearning from an Efficiently Contextualized Teacher — Jun 2025 – Sep 2025
ICLR 2026. — *Fairfax, VA*

- Developed DUET, an LLM unlearning technique that (i) uses an efficiently contextualized teacher (prompt-conditioned) to **demonstrate refusals on undesirable knowledge** and (ii) **distills this behavior into a student model**, achieving targeted forgetting with minimal utility loss.
- Introduced **Top-K logit alignment** in place of full-vocabulary KL, enabling more precise forgetting with better efficiency; on MUSE, average leakage decreased by 4% (ROUGE-Forget) and utility increased by 10% (ROUGE-Retain/MMLU), while training used about 1/645 of corpus tokens (2,233 vs. 1.44M).
- Demonstrated robustness to reverse prompts and task-format shift (QA $\rightarrow$ continuation) on WMDP and MUSE; under reverse prompts, the in-context teacher's leakage rises from 4.52% to 37.62%, whereas DUET remains around 5.98%$\rightarrow$7.27%.

Developing Engaging AI Chatbots to Enhance Senior Well-being — Nov 2024 – Jun 2025
NAIRR PILOT Project (PI: Zhuangdi Zhu); First author of the accompanying paper. — *Fairfax, VA*

- Designed and implemented ChatWise, a **strategy-guided** AI chatbot that integrates macro-level conversation planning with micro-level utterance generation to support **cognitive health for older adults**, achieving **16.83%** improvement in user verbosity.
- Conducted offline evaluation with **real clinical trial** data, validated via a novel **Strategy Match Percentage** (SMP) metric, achieving near-human alignment in conversational strategies with professional caregivers (SMP over **90%**).
- Developed ablation studies demonstrating that **macro-level** strategy guidance, more than emotion tracking alone, is the key driver of long-term engagement in multi-turn dialogues.

Artificial Intelligence Intern May 2021 – Aug 2021
Guosen Securities, Management System Development Department *Shenzhen, China*

- Employed regression analysis, **time series** forecasting, and cross-sectional analyses to derive significant economic factors with R. Resulting in enhanced risk management and **4%** alpha for the portfolio.
- Predicted a robust MedianAE of **3.1%** and RMSE of **4.5%** stock returns, with a PCA-optimized **LSTM** model.
- Developed a bill OCR system to automate the extraction of information from financial bills by optimizing Paddle's OCR model with **99%** accuracy.

Selected Publications

Full list: Google Scholar

1. **CATNIP: LLM Unlearning via Calibrated and Tokenized Negative Preference Alignment** *arXiv preprint 2026*
   Zhengbang Yang, Yisheng Zhong, Junyuan Hong, Zhuangdi Zhu
2. **DUET: DISTILLED LLM UNLEARNING FROM AN EFFICIENTLY CONTEXTUALIZED TEACHER** *ICLR 2026*
   Yisheng Zhong, Zhengbang Yang, Zhuangdi Zhu
3. **ChatWise: A Strategy-Guided Chatbot for Enhancing Cognitive Support in Older Adults** *KDD 2025 Workshop SciSocLLM*
   Zhengbang Yang, Junyuan Hong, Yijiang Pang, Jiayu Zhou, Zhuangdi Zhu
4. **Hierarchical Federated Unlearning for Large Language Models** *FedKDD 2025*
   Yisheng Zhong, Zhengbang Yang, Zhuangdi Zhu
5. **Sports intelligence: Assessing the sports understanding capabilities of language models through question answering from text to video** *Electronics*
   Zhengbang Yang, Haotian Xia, Jingxi Li, Zezhi Chen, Zhuangdi Zhu, Weining Shen
6. **SPORTU: A Comprehensive Sports Understanding Benchmark for Multimodal Large Language Models** *ICLR 2025*
   Haotian Xia, Zhengbang Yang, Junbo Zou, Rhys Tracy, Yuqing Wang, Chi Lu, Christopher Lai, Yanjun He, Xun Shao, Zhuoqing Xie, Yuan-fang Wang, Weining Shen, Hanjie Chen
7. **Language and Multimodal Models in Sports: A Survey of Datasets and Applications** *arXiv preprint, 2024*
   Haotian Xia, Zhengbang Yang, Yun Zhao, Yuqing Wang, Jingxi Li, Rhys Tracy, Zhuangdi Zhu, Yuan-fang Wang, Hanjie Chen, Weining Shen
8. **SportQA: A Benchmark for Sports Understanding in Large Language Models** *NAACL 2024*
   Haotian Xia, Zhengbang Yang, Yuqing Wang, Rhys Tracy, Yun Zhao, Dongdong Huang, Zezhi Chen, Yan Zhu, Yuan-fang Wang, Weining Shen
9. **Improving Confidence of Uncertain Knowledge Graphs by Crowdsourcing with Limited Budget** *ICPADS 2023*
   Haodi Zhang, Wenxi Huang, Chenyu Xu, Zhengbang Yang, Hongxin Zhou, Fengtian Qi, Chen Zhang, Kaishun Wu

ACADEMIC SERVICES

ICLR Reviewer 2026 2025

Teaching

Mgmt FE 247 VENTURE CAPITAL & PRIVATE EQUITY — The University of California, Irvine Fall 2023
Graduate Teaching Assistant *Irvine, CA*

Skills

Programming Languages: Python, Java, C/C++, C#, SQL, JavaScript, R, HTML/CSS
Frameworks & Tools: AWS, React, Django, Flask, Node.js, Next.js, Docker, Azure, Git, CUDA
Libraries: PyTorch, Transformers, LangChain, OpenCV, Pandas, NumPy, Scikit-learn, Matplotlib
Databases: MongoDB, MySQL, PostgreSQL