Yixiao (Jack) Zeng

rabin.ai | (949) 293-7961| yixiaozeng0208@outlook.com | LinkedIn | GitHub | HuggingFace
Research Topics: AI Infra, LLM/World Model Inference Optimization, RAG System Evaluation, Prompt Generation

EDUCATION

Carnegie Mellon University Aug. 2024 – Dec. 2025
M.S in Intelligent Information Systems (NLP Research Master, GPA: 4.0) *Pittsburgh, PA*
Featured course: 11-711 Advanced NLP, 11-697 Intro to Question and Answering, 11-868 LLM Systems

University of California, Irvine Sep. 2021 – June 2024
B.S. in Computer Science, Specialization in Intelligent System (GPA: 3.803) *Irvine, CA*

WORK EXPERIENCES

Machine Learning Engineer Jan 2026 – Present
XPENG *Santa Clara, CA*

- Integrated custom LLMs and VLMs into the vLLM framework on AI accelerators, optimizing Triton-based linear attention kernels to resolve inference precision issues and boost overall throughput by 80%.
- Invented **X-Cache**, a novel **training-free cross-chunk block caching algorithm** for **few-step interactive autoregressive world models** in autonomous driving simulation; designed **structure-aware and action-aware fingerprinting**, **adaptive per-block thresholding**, and **KV-update protection** to safely reuse DiT residuals across consecutive generation chunks, achieving **71% block skip rate** and **2.6–2.7× DiT wall-clock speedup** with minimal degradation **(PSNR > 50 dB).**
- Worked on evaluation and optimization of **Kimi K2.6 code-agent inference serving** (1000+ internal users) using AIPerf-based multi-user, multi-turn coding-agent workloads to measure TTFT/ITL, goodput, tail latency, and KV-cache behavior. Experimented with multi-node GPU serving using **KubeRay and vLLM/SGLang**, tuning replica placement, autoscaling, and prefill/decode balance under realistic concurrency. Also explored further changes to speculative decoding beyond provided parameters *(still ongoing)*. Achieved **~30% cost reduction compared with the official Kimi API** in internal cost evaluation, with contributions to SGLang.

Software Engineer Internship May 2025 – Aug. 2025
Amazon Web Services (AWS) *Bellevue, WA*

- Performed latency benchmarking on state-of-the-art VLMs and LLMs using inference frameworks such as vLLM, SGLang, and TensorRT with customer-representative datasets, based on internal format and return desired outputs
- Enhanced large-scale encoder-only model inference on AWS servers for unstructured data extraction using VLMs and OCR text + text-only LLMs by extending SGLang **with multi-modal and long-context speculative decoding**, tuning inference server parameters, and **modifying batching strategy** for data parallelism - overcoming prior scalability limitations and achieving a **51%** reduction in average latency with two times throughput compared to baseline.

Undergraduate Researcher Internship Jun. 2023 – Sep. 2023
Microsoft Research Asia *Beijing, China*

- Fine-tuned Llama-2-7B model based on different **PEFT** methods such as **LoRA, P-tuning,** and **Prefix-tuning**
- Tested PEFT models' performances on various NLP tasks based on benchmarks such as MMLU and WMT-22
- Evaluated runtime performance (load time, memory footprint and token generation speed) of LLM edge devices (MacOS, Linux, Windows PC and Android devices) inference tools such as **llama.cpp** and **MLC-LLM**
- Researched on designing heterogeneous inference system to **minimize server-side LLM usage** based on training a small model to judge questions' hardness**. Decrease overall server-side calling frequency by 40% tested on internal real-world user datasets.**

RESEARCH PROJECTS

RAG Synthetic Data Generation and Robustness Evaluation Pipeline Sep. 2024 – Sept 2025
Carnegie Mellon University, Advisor: Prof. Lei Li *Pittsburgh, PA*

- Built a **knowledge-graph-driven synthesis pipeline** that extracts triplets from unstructured datasets and traverses them to automatically generate single- and multi-hop QA with ground-truth chunks and no manual curation.
- Designed a **comprehensive RAG robustness metric** that tests correctness or safe refusal under query/document perturbations and real-world retrieval settings.
- Curated a dataset containing **48,295 questions (6,000 testing QA)** over 527 time-sensitive finance, economics, and policy documents, with diverse multi-hop question patterns (chain, star, inverted-star).

Automatic Instruction Induction via Meta-Learning (PROMPT-MII) Jul. 2025 – Sep 2025

Carnegie Mellon University, Advisor: Prof. Graham Neubig *Pittsburgh, PA*

- Proposed **Prompt-MII**, a GRPO-trained meta policy that converts a few labeled examples + label names into a single, reusable instruction in one pass to steer a frozen follower LLM, using macro-F1 on 20 held-out items as reward-eliminating costly iterative prompt tuning.
- Trained on **3,811** HuggingFace classification datasets (3,430 train / 381 val) and evaluated on **90** unseen tasks across multiple n-shot settings.
- Achieved **+4-9 macro-F1** (around 10-20% relative) and matched 100-shot ICL with **3-13 times fewer tokens.**
- **Outperformed APE/GEPA** iterative prompt optimization while requiring a **single** LLM call yielding higher F1 (**0.405-0.432** vs **0.288-0.358**)

LLM Inference Optimization (Improving Multi-candidate Speculative Decoding) Mar. 2023 – Sep. 2024

UCI Intelligent and Autonomous System Lab, Advisor: Prof. Marco Levorato *Irvine, CA*

- Enhance multi-candidate speculative decoding with a **target-initialized** multi-candidate token tree to better align with the target distribution, a **dynamic sliced topology-aware causal mask** that lets us vary draft length without rebuilding masks, and a lightweight **early-stop decision model** to skip unproductive draft steps
- Up to **27.5% latency speedup** vs. SOTA methods on TriviaQA, Alpaca, and MT-Bench with Llama-2-7B (target) + Llama-68M (draft); the static target-initialized variant achieved the best speed/quality trade-off.

SELECTED PUBLICATIONS

[1] **X-Cache: Cross-Chunk Block Caching for Few-Step Autoregressive World Models Inference**
Yixiao Zeng, Jianlei Zheng, Chaoda Zheng, Shijia Chen, Mingdian Liu, Tongping Liu, Tengwei Luo, Yu Zhang, Boyang Wang, Linkun Xu, Siyuan Lu, Bo Tian, Xianming Liu
Technical Report, arxiv 2604.20289

[2] **RARE: Retrieval-Aware Robustness Evaluation for Retrieval-Augmented Generation Systems**
Yixiao Zeng, Tianyu Cao, Danqing Wang, Xinran Zhao, Zimeng Qiu, Morteza Ziyadi, Tongshuang Wu, Lei Li
In processing to ICML 2026

[3] **Prompt-MII: Meta-Learning Instruction Induction for LLMs**
Emily Xiao, **Yixiao Zeng**, Ada Chen, Chin-Jou Li, Amanda Bertsch, Graham Neubig
Accepted by ICLR 2026

[4] **HARDTESTGEN: A High-Quality RL Verifier Generation Pipeline for LLM Coding**
Zhongmou He, Yee Man Choi, Kexun Zhang, Ivan Bercovich, Jiabao Ji, Junting Zhou, Dejia Xu, Aidan Zhang, **Yixiao Zeng**, Lei Li
Accepted by ICLR 2026

[5] **Improving Multi-candidate Speculative Decoding**
Yixiao Zeng*, Xiaofan Lu*, Feiyang Ma, Zixu Yu, Marco Levorato
Accepted by NeurIPS 2024 ENLSP-IV