Jiaqi Cao

(+86) 177-8839-9523 · maximus.cao@outlook.com · Wechat@maxmcao
Homepage @ mxmcao.github.io

Summary

My research focuses on parametric memory architectures and continual learning for large language models (LLMs). I have published 2 papers as a (co-)first author at **NeurIPS 2025** and **ICLR 2026**. My representative works, **Memory Decoder** and **MLP Memory**, introduce a novel parametric memory paradigm and provide a preliminary exploration of decoupling reasoning capabilities from long-tail knowledge. I am well-versed in LLM pretraining, embedding models, and various RAG architectures, and am dedicated to mitigating LLM hallucinations and enabling continual learning.

Education

Shanghai Jiao Tong University · M.E. Sep 2024 – Present
GPA: 3.79 / 4.0 · Published 2 papers at top ML venues; Expected graduation: Mar 2027

Shanghai Jiao Tong University · B.E. Sep 2020 – Jun 2024
GPA: 3.86 / 4.0 · Dean's Scholarship (Rank 1/86, $7,500)

Publications

(denotes equal contribution)*

Memory Decoder: A Pretrained, Plug-and-Play Memory for Large Language Models ***NeurIPS 2025***
Continual Learning · Parametric Memory · Knowledge Decoupling · Domain Adaptation · *60+ stars*
Jiaqi Cao*, Jiarui Wang*, Rubin Wei, Qipeng Guo, Kai Chen, Bowen Zhou, Zhouhan Lin

MLP Memory: A Retriever-Pretrained Memory for Large Language Models ***ICLR 2026***
Knowledge-Intensive QA · MLP Memory Module · Factual Knowledge Memorization · LLM Hallucination · *40+ stars*
Rubin Wei*, **Jiaqi Cao***, Jiarui Wang, Jushi Kai, Qipeng Guo, Bowen Zhou, Zhouhan Lin

Experience

Shanghai AI Lab | Research Intern Apr 2025 – Aug 2025

- **Research Vision:** Addressed the root cause of LLM hallucinations — the entanglement of knowledge and reasoning — by proposing parametric memory modules that explicitly decouple long-tail knowledge from the base model.
- **Memory Decoder:** Employed a parametric model to approximate the distribution of a non-parametric retriever, enabling a plug-and-play memory module for domain adaptation (legal, medical, financial, etc.) that injects domain-specific knowledge without modifying base model weights, thereby avoiding catastrophic forgetting.
- **MLP Memory:** Simplified the full decoder architecture into a stack of MLPs that directly learn the mapping from hidden states to output distributions. Achieved significant accuracy improvements on multiple QA benchmarks while attaining 2.5× inference speedup over conventional RAG pipelines.

Microsoft Cloud + AI (C+AI) | LLM Algorithm Intern Mar 2024 – Aug 2024

- **AKS Intelligent Operations Copilot:** Led the design and development of an intelligent diagnostic agent for Azure Kubernetes Service (AKS), covering high-frequency failure scenarios such as pod crashes and node anomalies. Built an end-to-end pipeline from natural language to interactive diagnosis (intent understanding → knowledge retrieval → kubectl command generation and execution guidance).
- **RAG Retrieval Optimization:** Implemented a query rewriting mechanism for intent alignment; restructured the QA knowledge base with systematic organization, boosting Top-3 retrieval recall to 85%.
- **Latency Reduction:** For multi-turn diagnostic sessions, designed key-field extraction and dynamic summarization to compress long contexts; developed an adaptive retrieval strategy to skip redundant RAG calls, combined with parallel API invocations to optimize the LLM call chain, reducing end-to-end latency by 800ms.

Awards & Honors

- **Dean's Scholarship** (Rank 1/86, $7,500 — highest departmental scholarship) Sep 2023
- **VEX Robotics World Championship Champion** (Team Captain) Apr 2023
- **National Olympiad in Informatics in Provinces (NOIP) First Prize** Nov 2018
- **Mathematical Contest in Modeling (MCM) Meritorious Winner** Jan 2021

Skills

- **Languages:** English (TOEFL: 109); strong proficiency in academic writing and technical presentation.
- **Programming:** `Python`, `C/C++`; experienced with PyTorch, Transformers, FAISS, and related frameworks.