XINGJIAN (SCOTT) ZHANG

Pittsburgh, PA ◇ xingjiaz@andrew.cmu.edu

GitHub ◇ LinkedIn ◇ Google Scholar (77 citations) ◇ Homepage

Research focus: efficient and reliable LLM inference – KV-cache compression and reuse, long-context memory, RAG.

Sole author of four papers accepted in 2026, incl. EACL 2027 (main) · Seeking a Summer 2027 internship

EDUCATION

Carnegie Mellon University — Pittsburgh, PA

M.S. in Artificial Intelligence Engineering – Electrical & Computer Engineering — Aug 2026 – Dec 2027 (expected)

New York University — New York, NY

B.A. in Computer Science, Minors in Mathematics & Economics; Dean's List — May 2026

PUBLICATIONS

- **X. Zhang**. AuthorityKV: Preventing Authority Reversals in KV-Cache Compression. *EACL 2027*, to appear [paper] [code].
- **X. Zhang**. ProofKV: Compiling Complete Evidence Chains into Reusable KV Memory for Structured-Evidence QA. *Findings of EACL 2027*, to appear [paper] [code].
- **X. Zhang**. SAFE-KV: Causal Suffix Repair for Behavior-Preserving Non-Prefix KV Cache Reuse in RAG. *Findings of IJCNLP-AACL 2026* [paper] [arXiv].
- **X. Zhang**. ASCENT: Scale-Complementary External State for Frozen Long-Context Language Models. *Array* 32:101399 (2026) [paper].
- X. Yuan, **X. Zhang**, et al. Tracing LLM Reasoning Processes with Strategic Games. *arXiv:2506.12012*, 2025.
- Journals: **X. Zhang** et al., *Sci. Rep.* 15:36467 (2025) [paper] [code]; **X. Zhang** et al., *Results Eng.* 29:108543 (2026) [paper] [code]; Z. Guo, ..., **X. Zhang** (co-first), et al., *Knowl.-Based Syst.* 301:112286 (2024) [paper] [code]; J. Jin, ..., **X. Zhang** (co-first), et al., *Sci. Rep.* 15:35099 (2025) [paper] [code].

RESEARCH EXPERIENCE

Independent Researcher (sole author) – *efficient and reliable LLM inference* — 2026 – Present

- **AuthorityKV** *(EACL 2027)* [code]: built IH-CacheBench and a training-free authority-tier wrapper for NVIDIA KVPress: instruction-hierarchy reversals cut from 2,055 to 0 of 7,200 decisions (20–40% KV); System-IFEval accuracy +36–46 pts; prompt-injection success from 7/64 to 0/64 (StreamingLLM).
- **ProofKV** *(EACL 2027 Findings)* [code]: built an evidence-to-KV compiler for long-context RAG with a vLLM paged-KV route: +19.3 F1 over LLMLingua-2 at the same token budget; +35.6 pts recall of complete multi-hop evidence chains over a reranker (MuSiQue); 1.16–1.84× end-to-end H100 speedup with a reusable index.
- **SAFE-KV** *(AACL 2026 Findings)* [paper] [code]: built causal suffix repair for non-prefix KV-cache reuse in RAG: 99.8%/98.4% of full-prefill answers preserved vs. 47.2%/37.8% for unrepaired reuse (HotpotQA/2Wiki, Qwen2.5-7B), with changed-wrong answers cut from 241 to 1; 135 unit tests.
- **ASCENT** *(Array 2026)* [paper] [code]: built external memory for frozen LLMs that lifts five 7–8B models by 67–86 pts on 16K-token BABILong (Qwen2.5-7B: 33%→100%) with 78–84× fewer reader FLOPs than full context (SmolLM2).
- Further sole-author papers under review on LLM evaluation and test-time compute.

Efficient Sequence Modeling – *with Prof. Eunsol Choi & Yilun Kuang (NYU)* — May 2025 – Jan 2026

- Designed SubMLR, a multi-level low-rank attention architecture between quadratic attention and constant-memory linear RNNs that beat strong baselines on Zoology MQAR; built chunk-compression models on linear attention (∼10% higher accuracy at matched FLOPs) and a FLOPs-matched train/eval pipeline on Slurm (PyTorch, Triton, flash-linear-attention).

Research Collaborations – *Prof. Haohan Wang (UIUC); Prof. Qingyun Wang (W&M)* — Feb 2025 – Jan 2026

- Co-built AdvGameBench [paper], a strategic-game benchmark of LLM planning and revision under budgets.
- Added nine chart-type experts to the ChartMoE VLM; built verified Program-of-Thought data for SFT and DPO/GRPO.

Transport Safety Modeling – *collaboration with Dr. Faan Chen (Harvard)* — Jan 2024 – Sep 2025

- Built hybrid ML + multi-criteria decision models (tuned DBSCAN/K-means, GMM clustering) on a curated 2010–2023 road-safety panel: 4 first/co-first-author journal papers (60+ citations) and 4 open-source reproduction packages.

SKILLS

ML & LLM: PyTorch, JAX, Hugging Face, vLLM, KVPress, Triton, flash-linear-attention; RAG, LoRA, SFT, DPO/GRPO
Programming & data: Python (pandas, NumPy, SciPy), C++, Java; SQL (PostgreSQL, Spark SQL), PySpark
Infrastructure: Linux, Docker, Git, Slurm/HPC; cloud GPUs (A100/H100), GCP, AWS
Evaluation & statistics: benchmark design, LLM-as-judge audits, pre-registered experiments, bootstrap CIs, paired tests