

Zain Sarwar

zsarwar865@gmail.com

309-703-5234

[Website](#)

[Google Scholar](#)

[GitHub](#)

RESEARCH INTERESTS

My research focuses on developing more efficient and capable language models, particularly through new neural architectures, improved training methods, and better ways of handling long contexts and reasoning. I am interested in bridging foundational research with practical LLM development, from exploring new approaches to memory, routing, and computation to pretraining and post-training models at scale. My broader goal is to understand and improve how language models learn, reason, and use compute.

EDUCATION

University of Chicago

PhD, Computer Science (GPA 3.91/4.00)

Chicago, IL

Oct 2021 – Sept 2026

Dissertation: Routing, Summarization, and Memory in Neural Architectures for Reasoning and Language

Lahore University of Management Sciences (LUMS)

BSc, Economics and Computer Science (GPA 3.82/4.00)

Lahore, Pakistan

Sept 2016 – June 2020

PROFESSIONAL EXPERIENCE

Capital One

Applied Researcher IV – LLM Core & Agentic AI

New York, NY

Oct 2026 – Present

- Develop and train LLMs for internal enterprise applications and agentic AI products, covering the full model development lifecycle from pretraining to post-training and deployment.
- Design and implement supervised fine-tuning (SFT), preference optimization (DPO), and reinforcement learning (RL) pipelines to improve model reasoning, instruction following, and performance on domain-specific tasks.
- Develop custom training objectives, data pipelines, and evaluation frameworks to adapt open-weight LLMs to enterprise workflows and measure improvements in task performance.
- Build and optimize distributed training infrastructure for large-scale LLM pretraining and post-training, including multi-GPU training, memory optimization, and efficient inference.

RESEARCH EXPERIENCE

University of Chicago

Graduate Research Assistant

Chicago, IL

Oct 2021 – Sept 2026

Differentiable Neural Memory for Long-Context Transformers

- Developed a differentiable neural memory architecture that enables Transformers to store and retrieve information across long contexts without relying on full self-attention.
- Designed hierarchical memory with sparse read/write operations, enabling memory capacity to scale while keeping retrieval costs low and avoiding the quadratic complexity of full attention.
- Validated the architecture through LLM pretraining and fine-tuning of pretrained models, achieving performance close to full-attention Transformers on language modeling and downstream tasks.

Default MoE: Improving Sparse Expert Routing (NeurIPS 2025)

- Developed a training method that improves MoE routing by providing gradient signals for unselected experts using approximated expert outputs, without increasing expert computation.
- Improved expert specialization and model performance over standard Top-K MoE baselines while preserving the computational efficiency of sparse expert activation.

Continual Pretraining and Router Behavior in MoEs (TMLR 2025)

- Investigated router stability, expert specialization, and distribution shift during continual pretraining of MoE language models.
- Analyzed how routing dynamics affect adaptation and robustness, informing effective continual-pretraining strategies for sparse models.

Additional Research in Reasoning and Model Reliability

- Contributed to critique-guided distillation for robust reasoning (ICML 2026), exploring how critique and refinement improve model reasoning.
- Studied limitations of machine-generated text detection, contributing to research published at IEEE S&P 2023.

PREVIOUS INDUSTRY RESEARCH EXPERIENCE

Capital One

Applied Research Intern – LLM Core & Agentic AI

New York, NY

June 2025 – Aug 2025

Structured Thoughts: Reasoning with Context Pruning

- Developed a structured reasoning framework that separates intermediate computation from summarized outcomes, enabling

LLMs to selectively prune reasoning traces while preserving key information.

- Trained pretrained reasoning LLMs using pruning-aware attention, reducing context length and attention computation with minimal degradation in reasoning performance.
- Improved reasoning accuracy by 4–5% on selected benchmarks while enabling more memory-efficient long-form reasoning.

Capital One

Applied Research Intern – LLM Pretraining

New York, NY
June 2024 – Aug 2024

Dynamic Feedforward Networks for Transformer Pretraining

- Designed a Transformer feedforward architecture with token-routed low-rank components, enabling input-dependent specialization without increasing model parameters or compute.
- Improved representation learning and pretraining efficiency over parameter- and FLOP-matched dense Transformers, achieving lower validation perplexity and higher downstream accuracy across multiple model scales.

SELECTED PUBLICATIONS

- [1] **Critique-Guided Distillation for Robust Reasoning via Refinement.** Berkcan Kapusuzoglu, Supriyo Chakraborty, **Zain Sarwar**, et al. *ICML 2026*.
- [2] **Structured Thoughts for Improved Reasoning and Context Pruning.** **Zain Sarwar**, Supriyo Chakraborty, et al. *Preprint*.
- [3] **Dense Backpropagation Improves Routing for Sparsely-Gated Mixture-of-Experts.** Ashwinee Panda, Vatsal Baherwani, **Zain Sarwar**, et al. *NeurIPS 2025*.
- [4] **Continual Pre-training of MoEs: How Robust Is Your Router?.** Benjamin Thérien, Charles-Étienne Joseph, **Zain Sarwar**, et al. *TMLR 2025*.
- [5] **StructMoE: Structured Mixture of Experts Using Low Rank Experts.** **Zain Sarwar**, Supriyo Chakraborty, et al. *NeurIPS 2024 ENLSP Workshop*.
- [6] **Deepfake Text Detection: Limitations and Opportunities.** Jiameng Pu, **Zain Sarwar**, et al. *IEEE Symposium on Security and Privacy (S&P)*, 2023.

TECHNICAL SKILLS

Machine Learning & AI: LLM pretraining and post-training, Transformer architectures, Mixture-of-Experts, neural memory, long-context modeling, supervised fine-tuning (SFT), reinforcement learning (RLHF, GRPO, PPO), preference optimization (DPO), model evaluation and benchmarking.

Distributed Training & Infrastructure: PyTorch, DeepSpeed, FSDP, Megatron-LM, CUDA, Triton, NCCL, Slurm, Kubernetes, Kubeflow, Docker, distributed GPU training, multi-node orchestration, mixed-precision training, GPU profiling and performance optimization.

Programming Languages: Python, C, C++, Java, JavaScript, Go, Haskell, MATLAB