Abeer Kashar

akashar@uwaterloo.ca | linkedin.com/in/abeer-kashar | github.com/tmnin

Education

University of Waterloo — Waterloo, ON
Candidate for Bachelor of Applied Science in Computer Engineering — *Sep. 2023 – Exp. 2028*

Experience

Incoming Software Engineer Intern — May 2027 – Aug 2027
Stripe — *South San Francisco, CA*

Applied AI/ML Engineer Intern — May 2026 – Aug 2026
Microsoft AI — *Redmond, WA*

- Fine-tuned and benchmarked Phi-4 14B, Qwen, and other open-weight LMs using LoRA across query-generation, ranking, and personalization workloads, including 40K examples for query generation
- Built LLM evaluation infrastructure in Azure ML, developing query-level metrics for correctness, freshness, intent preservation, and negative-query behavior and enabling offline side-by-side model comparisons
- Evaluated quality, latency, and inference-cost tradeoffs across LLMs and SLMs, identifying task-specific drop-in replacements for GPT-class models in a large-scale recommendation and content-generation pipeline
- Containerized fine-tuned SLMs and worked with Azure Kubernetes Service (AKS) to deploy models on GPU-backed clusters, evaluating inference latency and serving performance

AI Engineer Intern — Sep 2025 – Dec 2025
Cresta — *Toronto, ON*

- Built privacy-preserving inference pipelines for real-time PII detection and redaction across 50K+ daily voice and text interactions
- Achieved sub-100 ms end-to-end latency using RoBERTa-based NER models combined with lightweight rule-based classifiers under strict production constraints
- Led fine-tuning, deployment, and monitoring of production LLM services, building GPU-backed serving infrastructure and staged rollout systems

Undergraduate Research Assistant — Sep 2025 – Present
CompLING Lab, University of Waterloo — *Waterloo, ON*

- Constructed rigorously controlled multilingual lexical datasets across English, French, Hindi, and Chinese, removing morphological and etymological confounders to enable large-scale study of sound-meaning relationship
- Developed a cross-linguistic analysis framework combining articulatory phonological representations, Dynamic Time Warping, and contextualized language-model embeddings, uncovering a novel locality effect in sound-meaning systematicity across languages

NLP and Computational Linguistics Research Intern — Jan 2025 – Apr 2025
NorwAI, NTNU — *Trondheim, Norway*

- Executed large-scale evaluation workloads on the LUMI supercomputer using CUDA, MPI and distributed Python
- Built data-generation and evaluation infrastructure for multilingual LLM reasoning, generating synthetic benchmarks across Germanic languages to probe temporal and linguistic generalization
- Supported the EU TrustLLM project through multilingual benchmark creation and large-scale LLM evaluation

Publications

- **Co-first author**. *Computational Linguistics* (ACL flagship journal, MIT Press), 2026
- **Co-author**. *NeurIPS* (Evaluations & Datasets Track, Poster), 2026
- **Third author**. *Natural Language & Data Bases (NLDB)*, 2026

Skills

Languages: Python, C++, Java, SQL
ML: PyTorch, Hugging Face Transformers, LoRA/PEFT, LLM Fine-Tuning & Evaluation, GNNs, scikit-learn
Systems: Kubernetes (AKS), Docker, Azure ML, Linux, CUDA, MPI, Slurm, Ollama