Keshav Kapur

AI Engineer & Software Engineer specializing in Large Language Models and LLMOps

Actively seeking Full-Time Opportunities and Fall 2026 Internships

College Station, Texas | kekapur2001@gmail.com | (979) 574-8292 | linkedin.com/in/keshav-kapur-123kk | GitHub

Education

Texas A&M University, Master of Science in Computer Science — May 2027
- **GPA:** 4.0/4.0

Manipal Institute Of Technology, BTech in Computer Engineering — Jul 2019 – Jul 2023
- **CGPA:** 9.1/10.0, Department Rank: 7/139. **Activities:** Head of AI & Automation at Mars Rover Manipal.

Work Experience

LinkedIn (Microsoft), AI Intern | *Python, PyTorch, Java, MySQL, Qwen 2.5, LoRA, Flyte* — May 2026 – Present
- Spearheaded an LLM fine-tuning initiative using LoRA on Qwen **2.5**, outperforming out-of-the-box SOTA models.
- **Increased** model accuracy to **70.4%** (compared to **49.8%** with GPT-5.3-Codex and **58%** with Claude 4.6 Opus).
- Decreased production inference costs from **$14.00–$25.00 down** to **$0.30 per million tokens**, achieving a **98% decrease** in expenditure.
- Engineered a robust, cross-functional tech stack integrating Python, PyTorch, Java and MySQL for high-performance machine learning workloads.
- Optimized production ML pipelines via Flyte orchestration, maximizing system reliability and model training throughput.

Texas A&M University, Graduate Researcher | *Python, PyTorch, LLMs, DPO, GRPO, SFT* — Aug 2025 – May 2026
- Implemented bias mitigation techniques (DPO, GRPO, SFT) and a multi-step story generation pipeline to reduce LLM social bias via neutral explanation generation.
- Evaluated model performance, improving StereoSet ICAT scores from **67.5** to **76.97** while reducing the Stereotype Score (SS) to **59.88** (Target: **50**).
- Validated bias reduction on the BOLD benchmark, significantly dropping gender bias variance from **0.206** to **0.044** and political ideology bias from **0.515** to **0.183**.

Shell, Associate Software Engineer | *C#, .NET, GitHub Actions, Terraform, Azure, Vector Databases, RAG* — Aug 2023 – Aug 2025
- Automated CI/CD pipelines leveraging GitHub Actions and Terraform; developed C#/.NET backend services that accelerated developer onboarding by **35%**.
- Engineered a Proof-of-Concept employing Retrieval Augmented Generation (RAG) with GPT-**3.5**-turbo and Vector Index hosted on Microsoft Azure.

Projects

Recurrent Attention Model (RAM) | *Python, PyTorch, Reinforcement Learning, RNN* — GitHub
- Architected a visual attention model with a hybrid training pipeline (REINFORCE and supervised learning) to simulate human visual focus.

PixelQuery: Multi-Modal Image Retrieval | *OpenAI CLIP, FAISS, Vector Databases* — GitHub
- Built a system bridging the gap between text and vision utilizing OpenAI's CLIP for joint embeddings and FAISS for efficient vector similarity search.
- Enabled lightning-fast image retrieval based on natural language queries, implementing advanced cross-modal alignment techniques.

Technical Skills

Software Engineering: C++, C#, Java, Python, JavaScript, Terraform, Shell Scripting, Docker, FastAPI, ReactJS, .NET Framework, Git, MySQL, SQL Server, PostgreSQL, Grafana.

AI: MCP, RAG, LoRA, PyTorch, NLP, Sentiment Analysis, Deep Reinforcement Learning, LLM, LLMOps, AI Foundry.

Cloud/DevOps: Azure, AWS, GitHub Actions, Linux, Kubernetes, Azure DevOps, Flyte.

Certifications: Terraform Associate (003), Microsoft Certified: Azure Fundamentals, Green Software Practitioner.