Sushant Karki | Beverly, MA | 978-810-6462

00ber.dev | sushantkarkiii@gmail.com | linkedin.com/in/sushant-karki/ | github.com/00ber

Objective

Machine Learning Engineer interested in the intersection of Neuroscience and AI, seeking to reverse engineer the brain's computational principles to develop novel brain-inspired AI architectures.

Skills

Generative AI	Transformer Models	Large Language Models (LLMs)	Neurosymbolic AI
Langchain	Vector Databases	Retrieval Augmented Generation (RAG)	Knowledge Graphs
Text Classification	BERT/RoBERTa	Natural Language Processing	Sentiment Analysis
Supervised Finetuning	PEFT(LoRA)	Quantization/Optimization/ONNX	PyTorch
DevOps	AWS Administration	CI/CD	Containerization/Docker

Work Experience

Principal Investigator & Machine Learning Engineer Oct 2024 - Present
Citizen Codex LLC (Arlington, VA)

- Principal Investigator on NSF SBIR research initiative - conceptualized novel technical approach and authored successful Project Pitch and subsequent Phase I research proposal.
- Conducting comprehensive literature reviews of state-of-the-art hallucination mitigation approaches, and developing novel integration strategies using advanced retrieval and knowledge representation techniques.
- Leading the development of Govskills.io, leveraging AI to streamline complex job search processes for federal positions.

Machine Learning Engineer Intern June 2024 - Sep 2024
Lamini (Menlo Park, CA)

- Conducted comprehensive literature review and led research initiative exploring advanced LLM fine-tuning approaches for automated unit test generation.
- Performed systematic evaluation of leading open-source models for code generation, benchmarking their performance on complex C testing scenarios.
- Engineered automated feedback mechanisms that achieved 30% improvement in test pass rates through systematic iteration and optimization.

Machine Learning Engineer Intern March 2024 - May 2024
Citizen Codex LLC (Arlington, VA)

- Led research and development of Federal Regulation Explorer, conducting literature review of state-of-the-art RAG techniques and implementing them for regulatory text analysis.
- Iteratively enhanced baseline RAG system through systematic experimentation with advanced retrieval strategies (re-ranking, multi-stage retrieval, and fusion techniques), achieving 20% improvement in RAGAS metrics.
- Designed and implemented systematic evaluation methodology incorporating multiple RAGAS metrics (context relevance, retrieval precision, faithfulness), enabling quantitative analysis of each architectural enhancement's impact.

Tech Lead & Site Reliability Engineer (Dual Role) May 2015 - Aug 2022
Minma, Inc. (Tokyo, Japan)

- Doubled platform throughput by designing a high-availability AWS infrastructure, mirroring the demands of deploying scalable machine learning models.
- Reduced incident response times by 60% by implementing a comprehensive monitoring solution (ELK stack, Cloudwatch) for a 100+ node distributed system, critical for the continuous operation of ML applications.
- Reduced deployment cycles by 50% (-25min) developing an automation framework with Ansible & Terraform, streamlining configurations and deployments, essential for agile development and machine learning workflows.
- Enabled dynamic scaling of core services by spearheading the transition to containerized applications using Docker and AWS ECS, crucial for agile machine learning model development and testing.
- Reduced search latency from 6 secs to under 1 sec by strategic caching and switching from Postgres to Elasticsearch, optimizing query performance and user experience.

Projects

Style Transfer and Mixing for Personalization | *SFT, Representation Learning* Ongoing

- Project Lead conducting research on style transfer and style mixing strategies for personalized applications using Large Language Models.
- Exploring methods to adapt and combine stylistic elements to achieve tailored outputs in text generation.

Better Recommendations with RAG and LLMs | *Langchain, Mistral, RAG, Pinecone, Databricks* Nov 2023

- An independent project aimed at enhancing search and recommendation systems using LLMs and RAG.
- Implemented an interactive chatbot that provides recommendations for local businesses in Maryland based on user query and chat history.

Scholarships/Awards/Honors

Evans Scholarship Jan. 2010 - Dec. 2014
Scholarship covering full-tuition, room and board. *Missouri Southern State University Honors Program*

Magna Cum Laude Dec. 2014
Missouri Southern State University

5th Place Nationwide in Database Design and Application Programming Jun. 2012
FBLA-PBL National Leadership Conference *San Antonio, TX*

1st Place in Database Design and Application Programming Jan. 2012
Missouri FBLA-PBL State Leadership Conference *Lake of the Ozarks, MO*

1st Place in Computer Concepts Jan. 2012
Missouri FBLA-PBL State Leadership Conference *Lake of the Ozarks, MO*

Phi Eta Sigma National Honor Society Induction Jan. 2011
Missouri Southern State University Chapter *Joplin, MO*

Education

University of Maryland at College Park College Park, MD
M.S. in Machine Learning (GPA: 4.0/4.0) *Aug. 2022 – May 2024*

Missouri Southern State University Joplin, MO
B.S. in Computer Information Science and Mathematics (GPA: 3.819/4.0) *Aug. 2010 – May 2014*

Ryukoku University Kyoto, JP
Study Abroad (Japanese Culture and Language Program) *Apr. 2013 – Mar 2014*

Language Proficiency

Nepali
Native/Bilingual

English
Native/Bilingual

Japanese
Full Professional Proficiency

Hindi
Limited Working Proficiency