ALEKSANDAR VASIC, STAFF AI ENGINEER

(541) 412-6567 • aleksandar.vasic.ai@gmail.com • CA, United States

AI Research Engineer with deep understanding in AI, deep learning, NLP and background of both development and research. Expertise in transforming ideas into fully operational AI solutions and further refining them into market-ready products.

EXPERIENCE

DTxPlus – Remote (Princeton, New Jersey) 01/2023 – Present

Founding AI Engineer

- Responsible for design and implementation of an end-to-end LLM system pipeline that prioritized user data security, achieving 100% compliance with HIPAA regulations while enhancing the reliability of healthcare AI assistant Carrie through Guardrails AI integration
- Focused on Test Driven Development (TDD) integrating LLMOps technologies such as automated testing using CircleCI and adopted several approaches like embedding adapter, re-rank to solve hallucination problem
- Created several low-latency agentic solutions based on API-based/local models using several frameworks such as LangChain/LangGraph, CrewAI, AutoGen, and refined more than 20+ prompt sets for smooth agent operation (OpenAI GPT-4o, ChatGPT, Claude, Gemini, Llama, ChromaDB, Django for back-end, software engineering)
- Inspired reliable, explainable, multi-modal, and accurate RAG systems leveraging query translation methods like multi-query, RAG-fusion, decomposition, HyDE, and effective retrieval approaches, achieved CAG methodologies independently (HNSW, knowledge graph, Neo4j, Agentic AI)
- Constructed an agile experimentation system driven by the Comet framework that enhanced optimization processes for machine learning models improving evaluation speed by 100%
- Explored and pioneered the implementation of MemGPT using the Letta framework to enhance patient assistant memory management, enable proactive disease detection while removing response time from 1.2s to 0.8s, utilizing PydanticAI to ensure the precise and structured outputs from agents

CypherLearning – Remote (Plano, Texas) 09/2021 – 12/2022

Staff AI Engineer

- Led AI career assistant agent development team and played a key role of building LLM-based application workflow
- Trained several domain specific LLMs using RLHF by Kubeflow in Google Cloud's Vertex AI.
- Crafted a comprehensive career support data schema and processing pipeline, incorporating synthetic data generation via LLMs and human-labeled data and managed data storage and retrieval in PostgreSQL deployed on Azure, leveraging pgvector for efficient handling of high-dimensional embedding. (Python, serverless Azure Function, SQLalchemy, Docker, Kubernetes)
- Collaborated with data engineers to create benchmarks to evaluate career assistant agent's performance in the perspective of guidance relevancy on freshmen's career choices, representing relationship between quiz result and careers, adapting concept of Holland code and other personality measures (Data-centric AI, Pinecone for knowledge base, CoT, ReAct)
- Built end-to-end LLM-based application development workflow concentrating on data ingestion/processing pipeline using unstructured (Python library) for unstructured data management in VertexAI environment and leveraged VertexAI Notebook and BigQuery and effective data visualization (OCR, PySpark)

Infinite Computer Solutions (Rockville, Maryland) 09/2018 – 08/2021

AI Research Engineer

- Conducted in-depth research on advanced LLM serving techniques, including continuous batching, load-balancing, key-value (KV) caching for latency reduction, and prompt tuning, parameter efficient fine-tuning like LoRA and multi-LoRA (Python, Tensorflow, PyTorch, Hugging Face)
- Trained and fine-tuned LLMs following several architectures (BERT, T5, GPT) for medical keyword detection, multi-modal chat applications and increased in the evaluation scale like ROUGE or BERTScore (Conversational AI)
- Designed and developed multiple LLM applications utilizing Amazon Bedrock focusing on model selection ensuring highest performance by evaluation performance on different use cases, integration to deliver scalable (Generative AI)

- Leveraged AWS SageMaker for model training and deployment, EC2 for compute optimization, and integrated additional AWS services such as Lambda, S3 and CloudWatch to ensure seamless performance, monitoring and cost efficiency

Amazon (Portland, Oregon) 05/2018 – 08/2018

Deep Learning Research Intern

- Researched effective architecture of self-supervised learning on large scale text dataset for pre-training LLMs (NLP)
- Took part in research group of LLM serving techniques and clustering algorithms (C, C++, Python, DBSCAN)

EDUCATION

Stanford University**,** Stanford, CA 09/2016 – 08/2018

Master of Science (MSc) in Computer Science

Comenius University in Bratislava**,** Bratislava, SK 09/2012 – 08/2016

Bachelor of Science (BSc) in Computer Science (Minor: Mathematics)

SKILLS

Programming Languages C C++ Python ML frameworks TensorFlow PyTorch Keras Transformers ML platforms MLflow Prompt Engineering RAG Adaptive RAG GraphRAG CoT ReAct LLM BERT T5 GPT Llama LLM serving techniques continuous batching KV-caching LLM pre-training & fine-tuning instruction tuning LoRA prompt tuning NLP LLM System Design MLOps LLMOps GenerativeAI AgenticAI Data-centric AI Conversational AI LLM application frameworks LangChain LangGraph LlamaIndex CrewAI AutoGen AWS SageMaker Lambda EC2 S3 CloudWatch VertexAI VertexAI Notebook BigQuery Comet CircleCI Guardrails AI VectorDB Pinecone ChromaDB FAISS PostgreSQL GraphDB neo4j Sentence-BERT LlaVa OpenAI embedding Django Agile Scrum Kanban Docker Kubernetes