Akshat Vasisht

615-484-2887 | avasisht2@wisc.edu | linkedin.com/in/akshat-vasisht | github.com/akshatvasisht | akshatvasisht.com

Education

University of Wisconsin Aug 2023 – May 2026
BS in Computer Science and Data Science *Madison, WI*
Relevant Coursework: Artificial Intelligence, ML Theory, Data Science Algorithms, Statistical Modeling, Probability Theory, Big Data Systems, Database Systems, Advanced DSA, Software Engineering, Systems Programming (OS, Hardware, Networks)
Extracurricular: Wisconsin AI Safety Initiative — Technical Scholar, Boxing Club

Experience

Machine Learning Intern (Capstone) Jan 2026 – Present
Qualcomm *Madison, WI*

- Engineered a multimodal inference pipeline on NPUs using ONNX Runtime, achieving real-time generative recommendations through hardware-aware INT8 quantization and model pruning.
- Architected a context-aware sensor fusion engine processing asynchronous streams (GPS, accelerometer, camera) to fine-tune Large Vision Models (LVMs) on-device, enabling adaptation to diverse environmental conditions.
- Eliminated cloud dependency by deploying pre-trained models entirely on-device, optimizing memory footprint and thermal efficiency for continuous offline operation.

Machine Learning Research Assistant Sep 2025 – Present
UW-Madison SSEC, Atmospheric Motion Vector Lab *Madison, WI*

- Engineered an optical flow pipeline in PyTorch using RAFT (Recurrent All-Pairs Field Transforms) to process multispectral satellite imagery, generating 3D atmospheric wind vectors for operational weather forecasting.
- Streamlined HPC workflows using Bash scripting and Slurm job scheduling, implementing MLOps best practices (Git, Docker) to ensure reproducible experiments across compute nodes.

Software Engineering Intern May 2025 – Aug 2025
Techbaton *Remote*

- Delivered a Learning Management System MVP within a cross-functional Agile team, managing sprints in Jira and designing scalable PostgreSQL schemas as database owner.
- Architected backend microservices using Django and RESTful APIs for course management and progress tracking; integrated Llama 3 (Groq API) with prompt engineering to automate quiz generation and grading.

Research Intern May 2022 – Jul 2022
Vanderbilt University Medical Center Biostatistics Department *Nashville, TN*

- Developed an R package for drug toxicity analysis, implementing statistical validation metrics and Total Least Squares (TLS) regression to assess drug synergy and experimental reproducibility.
- Performed statistical analysis and data visualization to identify promising drug combinations, collaborating with Prof. Amir Asiaee's research group to present functionality and clinical insights.

Projects

Weather Betting Platform | *Python, Java, FastAPI, Spring Boot, XGBoost, scikit-learn, React, MySQL, Docker*

- Engineered a time-series forecasting API using XGBoost quantile regression for risk-adjusted odds, automating hyperparameter optimization (Optuna) and feature selection (RFECV, TimeSeriesSplit) across 12 models for three weather targets.
- Built an automated ETL pipeline ingesting historical weather data (NOAA/NWS), engineering 90+ features (cyclical encodings, rolling statistics, lags, derived meteorological indicators), with backtesting validation for dynamic odds pricing.

Visual Agent Orchestrator | *Python, FastAPI, LangGraph, React, Docker*

- Architected a VS Code extension enabling visual multi-agent orchestration through user-defined DAGs compiled into LangGraph state machines, providing deterministic agent coordination with real-time state streaming.
- Engineered a dynamic model selection layer using LiteLLM to route tasks to specialized APIs based on reasoning complexity, achieving cost optimization across providers.

Semantic Audio Codec | *Python, C++, PyTorch, FastAPI, Next.js, WebSockets*

- Architected an audio codec achieving 300 bps bitrate (98% bandwidth reduction vs. VoIP), enabling a 158× cost reduction for satellite/IoT communications by reconstructing speech from semantic metadata.
- Optimized inference latency for consumer CPUs by implementing 8-bit quantization and thread-safe model management, reducing memory footprint while maintaining real-time full-duplex performance without GPU acceleration.
- Engineered a prosody extraction layer using pitch and energy dynamics to condition generative synthesis, preserving speaker emotion and identity within the semantic token stream.

Prompt Tuner | *TypeScript, React, Plasmo, Tailwind CSS, Playwright*

- Engineered a privacy-first, local-only inference engine using Chrome's built-in LLM (Edge AI) to eliminate cloud API latency and token costs, preventing data leakage by managing session caching on the client.
- Built a fault-tolerant ETL pipeline using GitHub Actions to orchestrate headless browser scraping, distilling documentation into JSON heuristics to keep the local model synchronized with current best practices.

Skills

Programming Languages: Python, Java, C++, C, JavaScript, TypeScript, SQL (PostgreSQL, MySQL), R, Bash, Swift, HTML/CSS
Machine Learning: PyTorch, TensorFlow, scikit-learn, XGBoost, ONNX, PySpark, Pandas, NumPy, Matplotlib
Backend: FastAPI, Django, Flask, Spring Boot, Hibernate, RESTful APIs
Tools & Platforms: Docker, Git, Github, Linux, Slurm, AWS, Vim, Power BI
Certifications: Oracle Cloud Infrastructure — AI Foundations, Generative AI Professional, Data Science Professional