Vivek Teja Sayyaparaju

vivek.teja.raju@gmail.com • (437) 980-3574 • aifiction.org • Vancouver BC Canada

AI research Engineer with a physics background and 7 years experience bridging research and production AI systems. Led claude-based LLM support agent at AWS (500K+ case hours saved), designed ResNet architectures at INL ($30M+ fuel cost reduction), with expertise in model explainability, reinforcement learning, and distributed high performance training across 20M+ simulations.

PROFESSIONAL EXPERIENCE

AMAZON WEB SERVICES [AWS] VANCOUVER BC — Dec 2024 - Present

- Re-architected and deployed a **general guidance LLM agent** reducing annual AWS support costs by **$100M+**, saving **500K+ support hours** per year.
- Improved **first-response resolution by 35%**, **decision accuracy by 26%**, and reduced agent hallucinations by **~95%** in 2500 cases evaluated by front line support.
- Designed **LangChain-based RAG pipelines**, optimized via **HyDE**, and applied **DSPy (MIPRO)** for systematic prompt optimization.
- Built a **scalable, production-grade architecture** using **API Gateway, Lambda, Fargate, and Amazon Bedrock**.
- Expanded supported services to **40+ AWS services**, increasing agent deflection to **~2% of all support cases** at scale.

AMAZON WEB SERVICES [AWS] NYC -> TORONTO — Feb 2022 – Dec 2024

- **Founded and led** development of **AWS Well-Architected Profiles**, scaling from **0 → 20,000+ users**.
- Reduced customer effort by **30,000+ hours**, influencing **$200M+ in downstream AWS revenue**.
- Led system design, implementation, and **team growth from 1 → 6 engineers**.
- Deployed Trusted Advisor consoles into **air-gapped, top-secret ADC regions**, meeting strict compliance, security constraints and under tight deadlines.

BLUE WAVE AI LABS, WEST LAFAYETTE INDIANA USA — May 2019 – Feb 2022

- Nuclear

- Designed and deployed **ResNet-based Convolutional Nueral Networks** to predict nuclear reactor eigenvalues, outperforming legacy models by **orders of magnitude 100x** in prediction accuracy.
- Models supported **fuel planning for 10+ U.S. nuclear reactors**, generating **$30M+ in cost savings**.
- Collaborated with domain scientists to integrate ML models into safety-critical workflows.

- Defense

- Selected as **youngest flight lead** in a DARPA [**U.S. Department of Defense**] program; ranked **top-3 in 2/3 performance metrics**.
- Built large-scale simulations on **INL [Idaho National lab] HPC clusters**, executing **20M+ simulations** using **SLURM and Singularity**.
- Trained reinforcement learning models achieving a **3% win rate in asymmetric strategy simulations**, then performed post-hoc analysis using **model explainability techniques such as SHAP and LIME**.
- Bridged simulation, ML, and decision-theoretic analysis for defense-grade research.

EDUCATION AND SKILLS

Interest

- Designing agent architectures that scale from single interactions to millions of queries while maintaining interpretability a nd reliability. Applying physics-inspired optimization and formal verification to build trustworthy agent systems for high-stakes domains.

EDUCATION

- **Purdue university Physics Bachelors: 2015 - 2019**
- **Purdue university visiting scholar: 2019 – present**

Skills

- Model architectures: ResNet, Transformers, - Training: PPO, A3C, RLHF with SFT - Explainability: SHAP, LIME, attention visualization - Optimization: DSPy, HyDE, RAG systems

- AWS: Bedrock, Lambda, Fargate, SageMaker - Distributed Computing: SLURM, Singularity, Docker - Databases: DynamoDB, Neptune, Elasticsearch Languages: Python, Java, TypeScript