Aravind Kannappan

New York City, NY
1-408-591-5449
aravinds.kannappan@gmail.com
LinkedIn | GitHub | Website
Open To Relocation
US Citizen

EDUCATION

New York University
MS in Applied Statistics
Specialization in Machine Learning
Aug 2024 – Expected Aug 2026

Baylor University
BS in Statistics
Minor in Biology
Aug 2020 – Aug 2024

Activities:
BU Chess Club, Honors College

TECHNICAL SKILLS

Languages:
Python, TypeScript, SQL, R, C++
AI/ML Frameworks:
LangChain, LangGraph, PyTorch, Hugging Face, vLLM, Scikit-learn, RAG, Prompt Engineering, Tensorflow, CV
Frontend: React, REST APIs
Cloud & DevOps:
AWS (S3, EC2, Bedrock), GCP, Docker, Kubernetes, CI/CD, Git

COURSEWORK

- NLP, Deep Learning, Data Engineering, Distributed Systems, Big Data, Bayesian Methods, Biostatistics
- Data Structures & Algorithms, Linear Algebra, Calculus, Database Systems, System Design, Operating Systems

TECHNICAL WRITING BLOG

- Wrote posts on optimization, generalization, and probabilistic modeling from first principles.
- Built reproducible Python notebooks, training scripts, and benchmarks.
- Studied failure modes (leakage, shift, calibration) and proposed fixes.

PUBLICATIONS

- **Multiple Myeloma Gene Signatures (The Oncologist, 2024)**
  Built statistical models in Python for mutation, pathway, and survival analysis, identifying gene signatures. Link
- **Agentic AI Systems for Clinical Reasoning (Book Chapter, 2026)**
  Developed frameworks for integrating clinical & social data to improve decision-making and patient-centered care. Link

SUMMARY

AI/ML Engineer with production experience designing LLM-powered systems, RAG pipelines, and multi-agent orchestration frameworks across healthcare, sports analytics, and AI safety research. Deployed microservices on AWS and GCP with CI/CD automation, HIPAA-aligned security, and ETL pipelines processing large scale datasets. Strong foundation in ML systems; comfortable driving full-stack AI delivery in fast-moving environments.

WORK EXPERIENCE

Machine Learning Engineer Apr 2026 – Present
AI Whistleblower Initiative | CORDA AI Fellow *Remote*

- Architected an RLHF pipeline over a LLM, collecting annotations from 10+ experts and converting pairwise ratings into reward model, improving precision by 30% over baseline zero-shot prompting
- Built a drift detection layer computing KL-divergence across consecutive model output distributions, triggering workflows when shift exceeded a tuned threshold across 500+ weekly sessions
- Constructed a semantic clustering pipeline vectorizing annotated session transcripts with sentence-transformers and indexing into FAISS, reducing retrieval latency by 60%
- Evaluated prompt engineering methods on expert-labeled datasets and chose CoT, reducing hallucinated legal citations by 22% on held-out jurisdiction-specific tests.
- Feedback finetuned a loop by ingesting annotated preference pairs into a LoRA adapter training workflow, cutting full model retraining cost by 80% while maintaining within 2% on domain evals

Machine Learning Engineer Feb 2026 – May 2026
Supervised Program for Alignment Research (SPAR) *Remote*

- Investigated AI safety properties of frontier LLMs using Bayesian methods to model AI consciousness; designed experiments across 3 models using Python and PyTorch for downstream analysis.
- Fine-tuned LLMs to generate 6 semantically equivalent question variants per indicator; measured cross-variant output distributions to stress-test alignment stability, reducing variance by 30%.
- Developed mathematical models characterizing ground truth drift in LLM judgment across model generations; analyzed how scaling and RLHF-driven training affect model behavior.
- Built statistical analysis over response distributions including Likert histograms, cross-model EDA, and semantic analysis of chain-of-thought justifications across 50+ safety-relevant indicators.
- Open-sourced full evaluation framework on GitHub for reproducible AI safety research; designed for longitudinal alignment drift tracking and integration with human expert survey data.

Software Engineer, AI/ML Website Sep 2025 – Apr 2026
Replays AI | AI-Powered Immersive Post-Game Recaps *New York, NY*

- Designed ingestion pipeline writing 5M+ structured play-by-play events into PostgreSQL as system of record; stored raw video in GCP referenced by pointer, cutting compute costs by 40%.
- Preprocessed using Python to handle schema validation, player disambiguation and deduplication; anchored video segmentation to event timestamps before Computer Vision inference.
- Orchestrated 3 parallel agentic inference pipelines spanning event feature extraction, CV play classification, and task-split LLM summarization; reduced inference latency by 35%.
- Decoupled storage and inference from the frontend using REST APIs; implemented recap caching and fast paths for live rankings sustaining sub-second response times
- Built React and TypeScript mobile app rendering live rankings, game recaps, and CV-scored highlight reels; reduced mean load time by 28% via component memoization.

Founder and ML Engineer Website Jul 2023 – Jul 2025
Synthure | Care Clarity AI *New York, NY*

- Built RAG retrieval pipeline over Snowflake for medical code classification and claim denial routing; reduced policy misinterpretation errors by 94% across production clinical workflows.
- Engineered typed intermediate representation and data quality gate validating schema, deduplication, and entity extraction confidence before downstream inference
- Designed multi-agent orchestration with task-split LLM routing; directed high-volume entity tagging to lighter models and plain-language generation to frontier LLMs using Claude APIs.
- Implemented gated controlled generation constraining outputs to source-grounded entity rewrites and defined output schemas; reduced hallucinated content across 60K+ records.
- Deployed vLLM inference at under 1.8s p95 latency on AWS within MongoDB and Redis microservices; enforced HIPAA-aligned JWT access control and input validation across all services.

PROJECTS

CSCI 2271 Computer Vision Github

- Trained diffusion world model on 737K+ frames with VAE, CNN reward model and PPO agents

RobustSight: Advancing AI Safety and Alignment Github

- AI Safety framework investigating the intersection of adversarial robustness and interpretability