Kavya Sree Kammari

Bangalore, India | kavyasreekammari2090@gmail.com | +91 9550394139

LinkedIn | GitHub | Google Scholar | AI Blog

Summary

Platform Engineer at SAP Labs specializing in agentic workflows, distributed backend services, and cloud-native AI systems, with a Computer Science and Engineering (AI) background and silver medalist distinction. Strong foundations in Java, Python, machine learning, and distributed systems, backed by published Q1 research and competitive AI/ML projects. Experienced in building scalable AI platforms, reliable multi-agent frameworks, and production-oriented agent infrastructure.

Skills

- **Languages/ Frameworks:** Python, Java, SQL, JavaScript, PyTorch, TensorFlow, Spark, Google ADK
- **AI / GenAI:** LangChain, LangGraph, Agentic AI, Multi-Agent Systems, RAG, LLM Evaluation, Natural Language Processing, Computer Vision, Reinforcement Learning, Machine Learning
- **Cloud & Infrastructure:** GCP, AWS, Azure, Docker, Kubernetes, Git, CI/CD, REST APIs, Microservices
- **Core Concepts:** Distributed Systems, System Design, AI Orchestration, State Management, Evaluation Pipelines

Experience

Platform Engineer (Agentic Engineering), SAP Labs July 2026 - Present

- Designed and implemented a checkpointing mechanism for Pydantic AI agents to enable short-term memory and thread state persistence across agent interactions, with TTL-based eviction to prevent unbounded memory growth in long-running processes.
- Built an in-process caching layer for MCP tool-list calls to eliminate redundant round-trips in agentic loops; implemented LRU eviction and per-entry TTL expiry with support for caller-triggered cache invalidation.

Internships

AI Intern, SAP Labs Jan 2026 - July 2026

- Built a 4-step LangGraph orchestration workflow coordinating two agents, Joule (natural language input) and a downstream AI agent (workflow orchestration and tool-calling), with bidirectional sync to the frontend canvas, integrating natural language-driven data pipeline management into SAP's enterprise SaaS platform.
- Implemented inter-agent state synchronization ensuring Joule-generated workflow updates reflected in the canvas in real time while incorporating manual user edits back into the agent's context for accurate workflow understanding.

Cloud Migration Intern, SAP Labs May 2025 - July 2025

- Built a Flask-based summarization tool using Ollama LLM to extract and condense project context from PQA/DED data and backup assignments, enabling migration experts to onboard onto projects without reviewing full documentation.
- Reduced manual project context discovery by replacing end-to-end document review with agent-generated summaries, supporting faster handovers and decision-making during enterprise cloud migrations at SAP.

Education

Amrita Vishwa Vidyapeetham, Bengaluru - Bachelor of Technology in Computer Science and Engineering (Artificial Intelligence), CGPA: 9.56 — 2022 - 2026

Sri Chaitanya College, Hyderabad - Senior Secondary Education (PCM), Percentage: 85% — 2020 - 2022

Sri Chaitanya Techno School, Kurnool - Secondary School, GPA: 10 — 2018 - 2020

Certifications

CS50 Cybersecurity, Harvard	Nov 2026
Google Cloud Certified - Generative AI Leader	Oct 2026
SAP Certified - SAP Generative AI Developer	Aug 2026
Google Cloud Gen AI Academy APAC, Google Cloud	Jul 2026
CS50AI, Harvard	Jan 2025
AWS Certification, AWS Academy	Aug 2024
CS50X, Harvard	Jun 2023

Volunteering

CSR Volunteer, SAP & Shristi Foundation Aug 2026 – Present

- Contributing to SAP's CSR initiative by developing dashboards and websites for NGOs, enabling Shristi Foundation to better manage and communicate their social impact.

Chair/President, IEEE Computational Intelligence Society (CIS) Nov 2023 – July 2025

- Founded the CIS chapter and grew it to a 40-member team; organized AI and RAG workshops (100+ attendees) and the AI Arena hackathon (150+ participants), mentoring 5 members and leading technical project initiatives.

Technical and Non-Technical Member, Google Developers Student Club Aug 2023 – Jul 2024

- Contributed to conducting sessions, assisted in event coordination of Dev Fests, Code-Labs and Meet-ups, created video content, and supported commuwnity learning activities.

Teaching Assistant, Amrita Vishwa Vidyapeetham Aug 2024 – Dec 2024

- Assisted faculty in lab sessions covering probability and supervised learning, mentored junior students, and evaluated coursework in AI-related subjects.

Technical Projects

Embodied Multi-Agent Robotic Assistant using GPT-4o and ROS2 | GPT-4o, ROS2, LangGraph, VLMs

- Architected an embodied AI robotic system integrating GPT-4o, ROS2, and Vision-Language Models (VLMs) for language-driven manipulation tasks. Implemented an A2A multi-agent architecture with inter-agent communication, tool-calling, Grounding DINO-based object grounding, and memory-aware contextual reasoning, achieving a 96.7% task success rate, 100% memory retention, and an 83.35% error recovery rate.
- Built scalable ROS2 modular pipelines using Python, LangChain, and LangGraph to enable real-time simulated robotic perception, memory integration, and autonomous manipulation in dynamic environments.

Deep Learning and Statistical Testing for Cyberattack Classification | PyTorch, LSTM, Bayesian Neural Networks, Naive Bayes

- Implemented Bayesian Neural Networks (BNN), LSTM, and Naive Bayes classifiers for multiclass cyberattack detection using the CSE-CIC-IDS2018 network intrusion dataset. Addressed class imbalance through stratified uniform sampling and applied Chi-square feature selection to identify discriminative network traffic features.
- Conducted Kolmogorov–Smirnov statistical testing to analyze feature distribution differences across attack classes; evaluated model performance over 10 iterations, achieving 87.2% mean LSTM accuracy and 84.4% mean BNN accuracy.

ASTROF — Autonomous Telescope Scheduler | Reinforcement Learning, LLM Agents, OpenEnv, FastAPI, Docker

- Built an OpenEnv environment for autonomous exoplanet scheduling, enabling LLM agents to select among 20 NASA Exoplanet Archive targets while reasoning over scientific priority, atmospheric conditions, stochastic weather, and time-sensitive deadlines.
- Engineered a 21-action, 15-minute decision framework and multi-objective reward system across three increasingly difficult scenarios (44/32/18 timesteps), supporting reproducible agent evaluation through FastAPI and Docker.
- Evaluated Qwen3-8B agents across scheduling tasks, achieving 0.75, 0.64, and 0.58 scores on easy, medium, and hard scenarios respectively, with server-side grading for priority and deadline fulfillment.

Extract Relevant Clip in the Video Based on Text Query | PyTorch, Transformers, Multimodal Retrieval

- Engineered a multimodal retrieval system combining ASR, NLP, and gated fusion of visual, audio, and text features, enabling precise clip extraction from long-form educational videos.
- Designed and trained a transformer-based architecture with contrastive learning, reducing validation loss from 1.9 $\rightarrow$ 1.15 and achieving Recall@10 of 85% on the HowTo100M dataset (200,000+ clips).
- Optimized retrieval performance with attention-driven fusion, delivering a 20%+ boost in recall over baseline keyword-based systems, with applications in e-learning and automated summarization.

Custom Resource-Aware Scheduling & Dynamic Scaling | Kubernetes, Docker, Distributed Systems

- Built a custom resource-aware scheduling framework using Kubernetes node affinity/anti-affinity to route CPU- and memory-intensive workloads to suitable nodes.
- Implemented Simulated Annealing to optimize container CPU/memory configurations, reducing fitness from 395.24 $\rightarrow$ 52.68 while lowering memory allocation from 512 MB to 191 MB.
- Developed Binary Search Decision Strategy (BSDS) for task-to-container allocation across 100 Docker nodes, reducing search overhead versus linear allocation and cutting worst-case search time from 12+ seconds to a fraction of that.

Achievements

Silver Medalist - Amrita Vishwa Vidyapeetham	August 2026
Thank Award - Embrace Curiosity, SAP Labs	Jun 2026
Thank Award - We Take Charge, SAP Labs	Jun 2026
LeetCode — 150+ problems (Data Structures & Algorithms)	Present
Academic Excellence Award, Amrita	May 2026 Apr 2025
Meta OpenEnv Hackathon - Finalist	May 2026
UX Competition, GDG, Bengaluru – Winner	Jan 2025
Best Research Paper Award, ITI Conference	Oct 2024
AI Arena, Bengaluru – Winner	Apr 2024

Publications

ProWGAN a hybrid generative adversarial network for automated landscape generation in media and video games	Discover Artificial Intelligene - Q1 Journal
A Comprehensive Review of Deepfake Detection Techniques Utilizing Remote Photoplethysmography	IEEE Access - Q1 Journal
Hybrid Metaheuristic Optimisation for Lung Cancer Image Classification: Leveraging MOEA, PSO, and ACO Algorithms	ICMLDE - Conference