Prasanth Janardhanan

AI Developer & LLM Systems Architect · ML Researcher

prasanthmj@gmail.com | +91 9845422369 | prasanthmj.github.io | LinkedIn | GitHub

Based in India · Open to relocation to Europe · EU Blue Card eligible

SUMMARY

AI Developer, LLM Systems Architect, and ML Researcher with 15+ years of engineering experience. Specialise in production LLM systems, agentic RAG pipelines, and AI agent infrastructure — including one of the early real-world LLM deployments in financial services.

Creator of **gomcpgo** (open-source Go MCP framework, 12 servers) and the **first ML model for Grade 2 contracted Braille OCR** — a ByT5 seq2seq model achieving 92.9% exact match on real-world Braille, 9x better than any existing baseline, published on Hugging Face. Also built the first open-source EU AI Act compliance navigator with a 5-stage agentic RAG pipeline — directly relevant to operating under EU regulation.

CORE SKILLS

AI / LLM	Large Language Models · RAG Pipelines · Agentic AI · MCP · AI Agent Design · Prompt Engineering · NLP · Financial Document AI · LLM Fine-tuning (LoRA/QLoRA) · Vector Databases (Qdrant)
AI Governance	EU AI Act · Responsible AI Design · AI Lifecycle Governance · Explainable AI · SOC2 · Regulatory Compliance
ML & Frameworks	Python · PyTorch · Hugging Face Transformers · Seq2Seq Fine-tuning · ByT5 · YOLOv8 · CUDA / A100 GPU Training · Temporal
Languages	Go / Golang (Expert) · Python · TypeScript
Infrastructure	AWS · GCP · Docker · Kubernetes · Terraform

PROFESSIONAL EXPERIENCE

AI Developer / Senior Backend Engineer · CISCO DevNet (Contract)

Jul 2025 – Present

Contributed to LLM and RAG integrations for internal tooling, including fine-tuning models to generate accurate sample code for Cisco integration products, and connecting documentation systems to LLM pipelines. Built MCP servers to expose internal tools and documentation as structured context sources for LLM-powered workflows. Developed AI-assisted documentation quality tooling that scores API docs for technical accuracy, developer friendliness, and semantic search optimisation. Built Go-based API specification analysis tooling using libopenapi for schema validation, semantic diff detection, and policy-based scoring of OpenAPI specs. Completed formal Cisco AI safety training covering safe AI practices and data governance policies.

Technologies: Go · Python · LLM Fine-tuning · RAG · MCP · OpenAPI / libopenapi · Kubernetes · Jenkins CI/CD

Senior Backend Engineer · Dome Global Inc (Remote)

Dec 2022 – Jun 2025

Built the core backend for a PaaS platform that automated multi-tenant SaaS deployment on Kubernetes. Designed an intent-driven provisioning system where users describe their application and the platform generates a complete, ready-to-deploy stack. Architected the automated deployment orchestration layer using client-go and Tekton pipelines, reducing deployment time by **60%**. Led the full technical revamp that delivered the MVP launch and first paying customers.

Technologies: Go · Kubernetes · GCP · AWS · Tekton · client-go

Senior Staff Engineer — AI & Backend · CredCore Inc (Remote)

Sep 2021 – Nov 2022

Architected and led engineering for an AI-powered capital structure management platform for enterprise financial institutions — one of the early production LLM deployments in fintech document intelligence. Designed and deployed an LLM-powered financial document analysis pipeline achieving **90%+ accuracy** in automated extraction of covenant clauses. Engineered AI workflow orchestration using Temporal to reliably process **10,000+ daily AI jobs**. Led SOC2 Type II compliance implementation and built multi-cloud infrastructure across AWS and GCP. Implemented Stripe billing processing **$1M+ ARR**.

Technologies: Go · Python · TypeScript · PostgreSQL · Redis · Temporal · Kafka · AWS · GCP · Kubernetes

Backend Engineer · TPS Inc

Jan 2015 – May 2021

Delivered two large-scale data engineering platforms: decomposed a monolithic transportation intelligence system into event-driven microservices (70% response time improvement, 100,000+ daily events via Kafka), and architected a real-time telecom tower monitoring pipeline processing 10TB+ monthly across 1,000+ concurrent streams on bare-metal Kubernetes with 99.9% uptime.

Technologies: Go · C · Node.js · Kafka · Kubernetes · InfluxDB

PROJECTS

EU AI Act Compliance Navigator

github.com/prasanthmj/eu-ai-act-rag · Go · Qdrant · RAG · Agentic AI · 2026

The first open-source developer tool for EU AI Act (Regulation EU 2024/1689) risk-tier classification with citation-backed output. Given a plain-English description of an AI system, it classifies the risk tier, identifies applicable Articles, Recitals, and Annexes, generates a prioritised compliance checklist with direct citations, and flags ambiguous scenarios.

Core is a 5-stage agentic RAG pipeline (single Go binary, 100K+ words of structured legal text): classifier agent, 3-hop multi-hop retriever following legal cross-references, obligation mapper, confidence scorer with per-claim citation verification, and checklist generator. Hybrid dense + sparse search (BM25 + RRF fusion via Qdrant). Evaluated using RAGAS framework (faithfulness ≥0.85, answer relevancy ≥0.80). Exposed as an MCP server via gomcpgo. React + TypeScript frontend on Vercel.

Stack: Go · Qdrant · OpenAI · gomcpgo · React · TypeScript · Python (ingestion + RAGAS evaluation)

Braille OCR Transcriber · First working Grade 2 model

github.com/braille-reader/braille-transcriber · model on Hugging Face · Python · ByT5 · 2026

A practical accessibility tool: teachers of visually impaired children often cannot read Braille, making it impossible to evaluate written exams. This system lets a teacher photograph a Braille answer sheet and get readable English text.

90–95% of real Braille uses Grade 2 contractions — no existing OCR supports this. Two-stage pipeline: YOLOv8 cell detection (98–99% accuracy), then a **ByT5-small** (300M param) seq2seq model fine-tuned on 25,138 synthetic Braille–English pairs. ByT5's byte-level architecture natively handles Unicode Braille (T5 maps all Braille to <unk>). Achieves **92.9% exact match** on real-world human-transcribed Braille (CER 0.004) — 9x better than the liblouis baseline. Zero hallucinations. Published on Hugging Face.

Stack: Python · PyTorch · ByT5 · Hugging Face Seq2SeqTrainer · Liblouis · YOLOv8 · CUDA / A100 GPU

gomcpgo — Go MCP Framework & Server Ecosystem

github.com/orgs/gomcpgo · 2025 – Present

A production-grade Go framework and ecosystem of 12 open-source MCP servers, enabling LLMs to securely interact with external tools, APIs, and data sources. MCP — donated by Anthropic to the Linux Foundation — is the open standard for AI agent tool integration. Fills a genuine gap: a Go-native alternative to Python and TypeScript SDKs, deployable as a single binary with zero runtime dependencies. Servers cover: universal REST API wrapping, real-time AI research, secure filesystem access, speech-to-text, image/video generation, email, web content fetching, and document generation.

EDUCATION & CERTIFICATIONS

- **Machine Learning Specialization** (Andrew Ng) — Coursera / Stanford Online, 2025
- **Certified Kubernetes Administrator (CKA)** — The Linux Foundation, 2019
- **Certified Kubernetes Application Developer (CKAD)** — The Linux Foundation, 2019
- **Bachelor of Science in Computer Science** — Mahatma Gandhi University