Arnol P S

Senior Data Science Specialist | AI/ML Researcher

Kerala, India | arnolps172@gmail.com | +91 90483 57172 | arnol.in | linkedin.com/in/arnol-ps

Summary

Senior data scientist and AI/ML engineer with 5+ years across data science, NLP, and computer vision, specialized over the past two years in generative AI, multi-agent systems, and retrieval-augmented LLM platforms for enterprise and government clients. Strong in LLM orchestration, hybrid and graph RAG, document AI, and self-hosted model serving, with AI safety and observability built in. Comfortable leading small teams and taking systems from research prototype to production across cloud and air-gapped, on-premise environments, and author of patent-pending computer vision research on open-set biometric identification validated across 197 identities.

Skills

Focus Areas: Agentic AI / Multi-Agent Systems, GraphRAG & Hybrid RAG, Document AI (IDP), Vision-Language Models, Knowledge Graphs

AI/ML: PyTorch, Transformers, Scikit-learn, ONNX, LoRA/QLoRA, vLLM, DINOv2, YOLO, OpenCV, Docling, Tesseract

LLMs & RAG: Claude, Gemini, Qwen, LangChain, LangGraph, Model Context Protocol (MCP), OpenRouter, Cross-Encoders, Reciprocal Rank Fusion (RRF)

AI Safety & Eval: Llama Guard, Presidio, Langfuse, RAGAS, PII Detection, Prompt Injection Prevention

Databases: Qdrant, Neo4j, Elasticsearch, PostgreSQL, Redis

Backend: FastAPI, Pydantic, AsyncIO, WebSockets, Celery, SQLAlchemy

Infra: Docker, CUDA, Keycloak, Prometheus/Grafana, Next.js, React, TypeScript

Languages: Python, TypeScript, JavaScript, SQL

Experience

Senior Engineer - Data Science, Reflections Info Systems Pvt. Ltd. – Kerala, India — Aug 2025 – present

- Built a **hybrid RAG platform** fusing Qdrant vector search, Neo4j graph traversal, and Reciprocal Rank Fusion with multi-path query routing, serving enterprise sales intelligence behind a layered security pipeline (rate limiting, input validation, LlamaGuard, RAG guardrails, PII sanitization)
- Architected **multi-agent log analytics**, an event-driven pipeline with semantic error grouping, anomaly detection, and decision-tree workflow routing
- Implemented **real-time WebSocket streaming** with a buffer-then-sanitize pattern and **Langfuse observability** for end-to-end LLM cost tracking
- Built **VLM document extraction** across diverse document formats with multi-provider failover and automatic VLM escalation below the OCR threshold
- Deployed **compliance verification** automating regulatory-criteria checks via multi-agent LangGraph orchestration with bilingual (EN/AR) support
- Built an **agentic outreach platform** with a custom MCP server and an LLM workflow spanning discovery, enrichment, scoring, document generation, and outreach, with a template-driven PPTX/DOCX/XLSX pipeline
- Contributed to the **ISO 42001:2023** AI Management System surveillance audit; received a Shout-Out award for documentation quality and audit preparedness

Senior Data Scientist - Consultant, Digital University Kerala – Kerala, India — Dec 2024 – July 2025

- Led **patent-pending** cattle biometric ID: YOLO26m muzzle detection → DINOv2 1024-dim embeddings → Qdrant search with open-set margin decision; validated across **197 animals (87.5% top-1, 93.5% top-5 precision)**

Senior Software Engineer - AI/ML, Techversant Infotech – Kerala, India — June 2024 – Nov 2024

- Built **RAG with conversational memory** for multi-turn enterprise interactions; developed face recognition and **YOLO-based proctoring** systems

Senior Engineer - Data Science, Digital University Kerala – Kerala, India — Sept 2023 – June 2024

- Led a 3-person team building **ML search infrastructure** with Elasticsearch ETL; built the AWS backend for "Fun With AI" at Global Science Fest Kerala

Data Analyst, Digital University Kerala – Kerala, India — June 2021 – Aug 2023

- Built automated data pipelines and visualizations; optimized DB performance via indexing strategies

Research Fellow, ICFOSS – Kerala, India — Sept 2019 – Sept 2020

- Developed a Malayalam Morphological Analyzer and sentiment analysis systems; managed large-scale NLP datasets

Projects

PRISM - On-Prem Enterprise AI Platform

Self-hostable enterprise AI platform that orchestrates LLM tool-calling across a federation of MCP tool servers, backed by a fully self-hosted inference stack

- Built an **MCP-federated orchestration** backend: a custom multi-server MCP client with startup tool discovery, per-tool timeout budgets, and trace-ID and RBAC propagation, fronted by resilient SSE streaming with Redis-backed stream resume
- Built the **document RAG service**: Docling chunking, bge-m3 embeddings in per-tenant **Qdrant**, two-stage dense retrieval with cross-encoder reranking, and an adaptive map-reduce summarizer
- Stood up a **self-hosted, OpenAI-compatible inference stack** (vLLM with FP8 quantization, plus self-hosted ASR and TTS) behind a pluggable engine abstraction for air-gapped deployment

SalesBot - Hybrid RAG Proposal Intelligence

Multi-tenant hybrid RAG assistant over a sales-proposal knowledge base, with classification-driven retrieval fusion, citation and groundedness verification, and a layered security path

- Built a **multi-path RAG pipeline** that LLM-classifies and routes each query across graph, dense-vector, and keyword retrieval, fused with classification-aware **weighted Reciprocal Rank Fusion** and a cross-encoder reranker
- Added **per-sentence citation attribution** and NLI-based groundedness scoring, plus a self-correcting **natural-language-to-Cypher** engine with live schema introspection and query-safety validation
- Engineered a resilient **multi-provider LLM layer** with per-provider circuit breakers and fallback, behind a security path of rate limiting, input validation, LlamaGuard, and PII/secret redaction

Start2Scale - Agentic M&A Outreach Automation

Agentic M&A target-discovery and deliverable-generation system built on the Claude Code harness, orchestrating skills, subagents, and a custom Apollo.io MCP server

- Built a **custom Apollo.io MCP server** (FastMCP, async httpx) exposing ~11 discovery and enrichment tools, with response trimming, structured error feedback for agent self-correction, and fallback retries (69 mocked tests)
- Decomposed the workflow into **5 slash-command skills and 2 subagents** that fan out research and draft publication-ready teaser content with self-review, scored against Pydantic-validated buyer mandates
- Designed a **template-driven document layer** (python-pptx, python-docx, openpyxl) that populates branded PowerPoint, Word, and Excel deliverables via cross-run placeholder replacement

Cattle Muzzle Biometric Identification

Patent-pending open-set livestock identification from muzzle prints using zero-shot vision embeddings and vector search, validated across 197 animal identities

- Built a patent-pending **open-set identification** system reaching **87.5% top-1 / 93.5% top-5 precision** across 197 identities, using zero-shot **DINOv2** (ViT-L/14, 1024-dim) embeddings over **Qdrant**
- Designed a **margin-based open-set decision rule** (score floor plus runner-up gap) that cuts false accepts versus a single-threshold baseline on held-out unknown identities
- Engineered a **3-stage vision pipeline** (YOLO26m detection, wavelet ridge enhancement, DINOv2 embedding) behind a FastAPI service over Postgres, Qdrant, and Neo4j

Research

- **Handwritten Text Recognition (IJDAR)** – co-authored a multi-line HTR pipeline integrating CRAFT line segmentation, TrOCR transformer recognition, and a Mixtral LLM post-correction stage
- **Robotics in Agriculture (Scientometric Review)** – citation main-path and network-cluster analysis of ~4,000 Web of Science records (2015-2025) using Pajek and NLP topic profiling

Proofs of Concept

- **Invoice & Document Extraction** – multi-format document-to-JSON service combining Docling/Tesseract OCR, page-level routing to multi-provider vision-LLMs, and local DeepSeek-OCR inference (FastAPI, Next.js)
- **QFMA Compliance Verification** – RegTech PoC verifying governance reports against regulatory articles via hybrid Qdrant RAG with selective LLM validation in a LangGraph pipeline, with bilingual (EN/AR) reporting

Education

IIITM-K, M.Sc. in Computer Science (Data Analytics), First Class – Kerala, India — June 2017 – June 2020

Kristu Jayanti College, B.Sc. in Computer Science, Mathematics & Statistics, First Class – Bengaluru — June 2014 – June 2017

Certifications

Google Data Analytics Professional (Google) · Real-Time Video AI (NVIDIA DLI) · Deep Learning (NVIDIA DLI)