DAVID LUKIC

Senior AI Engineer | LLM & Multi-Agent Systems | Production RAG & MLOps

ldavid797@gmail.com • +381 65 616 0655 • linkedin.com/in/davidlukic99 • davidlukic99.github.io

PROFESSIONAL SUMMARY

Senior AI Engineer and Senior Data Engineer with 7+ years shipping production LLM and autonomous agent systems alongside large-scale data platforms. Top Rated Upwork freelancer, 100% Job Success Score. Own the full lifecycle — prototype through scaled, monitored production — including evaluation, observability, and guardrails. Recent work: a multi-agent customer support system serving ~500K users on AWS Bedrock and ECS; an internal hybrid RAG + MCP data platform unifying MySQL, S3, Google Drive, and Google Sheets behind a natural-language interface; a CrewAI-based multi-agent SEO content pipeline; and a compliance-report AI application that grades client materials against regulatory requirements. Fluent across major model providers (OpenAI, Anthropic, Gemini, AWS Bedrock), async Python, FastAPI, and containerized infrastructure (Docker, Kubernetes, ECS).

Core strengths: Multi-agent system design • Production RAG & MCP servers • LLM evaluation, observability & guardrails • Cost & latency optimization • End-to-end ownership (prototype → production) • Senior data engineering (Airflow, dbt, Spark, Kafka, Snowflake)

PROFESSIONAL EXPERIENCE

Senior AI Engineer | *BetterCollective* Nov 2024 – Present

- Built an internal Model Context Protocol (MCP) server unifying MySQL, AWS S3, Google Drive, and Google Sheets behind a hybrid RAG + MCP interface. Analysts ask questions in natural language and get immediate SQL execution, default visualizations, and a custom Power BI–style dashboard builder — replacing a workflow that previously required a developer to write queries and build reports manually.
- Shipped the MCP platform as a Django application with a custom UI; stack: FastMCP, FastAPI, vector DB for hybrid retrieval.
- Designed and deployed a multi-agent SEO content system — ~8 autonomous CrewAI agents covering research, analysis, scraping, and writing — running the full content pipeline end-to-end (CrewAI, FastAPI, Docker, Pydantic).
- Built three production RAG systems: (1) a cross-source knowledge RAG spanning Google Drive, Slack, internal docs, and external conference and article content, with full source attribution for every answer; (2) a competitor content intelligence RAG for angle and gap analysis against competitor articles and site structures; (3) a historical content performance recommender indexing past articles against their metrics so writers can surface what worked for a given topic with evidence.
- Established evaluation and observability for the RAG and agent systems using LangFuse and LangSmith — tracking groundedness, answer quality, latency, and cost per query across OpenAI, Anthropic, and Gemini.
- Own architecture decisions, model selection, prompt design, guardrails, and cost/latency tuning.
- Partner with product, marketing, and data teams to turn vague problems into shipped AI features.

Senior AI Engineer (Part-time) | *ComplianceLabX* Jan 2025 – Present

- Build and ship a production AI application that generates compliance reports — ingests client materials, evaluates them against regulatory requirements, and returns a structured report flagging compliant and non-compliant areas with evidence.
- Designed the end-to-end pipeline from ingestion and retrieval through inference, evaluation, and monitoring — with a focus on reliability, low-latency inference, and auditable outputs.
- Implemented guardrails and evaluation harnesses for groundedness, citation accuracy, and hallucination rate, so every report claim is traceable to the source regulation.
- Optimize for cost and performance through batching, caching, vector search tuning, and async processing; established LLM observability across requests, tokens, and latency.
- Own architecture decisions, production debugging, and post-mortems across the stack.

Senior AI Engineer (Contract) | *OnTheGoSystems* Oct 2025 – Apr 2026

- Built and shipped a multi-agent AI customer support system serving ~500,000 users in production. Roughly 10 autonomous agents handle refunds, how-to support, ticket triage, bug reports, and credit-related queries.
- Designed across multiple repositories (agent service, support backend, infrastructure) using FastAPI, Docker, and AWS ECS.
- Integrated multiple model providers — AWS Bedrock, OpenAI, Anthropic, Gemini — with routing logic balancing cost and latency per agent role.
- Owned the full pipeline: ingestion, retrieval, inference, tool use, guardrails, and LLM observability (tracing, latency, cost, quality metrics).

- Built evaluation harnesses over real production conversation data to catch regressions in agent behavior; iterated prompts and routing with product and backend teams.

Senior Data Engineer | *index.dev* 2024 – 2025

- Architected scalable ETL pipelines with Pydantic, dbt, and Airflow — 30% faster processing and 20% operational efficiency gain for enterprise clients.
- Tuned Postgres and DuckDB analytical workloads through query optimization and indexing strategies, accelerating decision-making by 20%.
- Enhanced Snowflake warehouse performance and reduced query costs on high-volume analytics.
- Built 10+ reproducible containerized development environments (Docker, Poetry), cutting onboarding time by 40%.
- Delivered interactive Power BI and Plotly Dash dashboards with cross-filtering, callbacks, and drill-downs.
- Enforced testing (pytest), type safety (mypy/pyright), and code quality (ruff/black) across production codebases.

Data Engineer / Full-Stack Developer | *BetterCollective* 2022 – Nov 2024

Promoted to Senior AI Engineer in Nov 2024

- Shipped 14 production Django + Plotly Dash applications for 50+ internal users, eliminating ~80% of manual reporting workflows.
- Designed LogAna, an automated log analysis platform processing 10M+ daily log entries — cut manual analysis from 8 hours to 30 minutes (400% improvement in processing speed).
- Engineered high-throughput data pipelines (pandas, Polars) processing millions of daily records with 99.9% uptime.
- Built ETL workflows across REST APIs, AWS S3, PostgreSQL, and DuckDB with sub-second query performance.
- Implemented asynchronous web scraping services with proxy rotation, rate limiting, and robust error handling.
- Managed AWS-hosted PostgreSQL with Redis caching and strategic indexing for fast, reliable data access.

AI & Data Engineering Consultant | *Self-Employed — Upwork* 2020 – Present

Top Rated Upwork Freelancer • 100% Job Success Score

- Architect and deploy production LLM applications — LangChain, LlamaIndex, RAG with Pinecone and PGVector — delivering document Q&A and semantic search for global clients.
- Build custom AI automation workflows integrating GPT-4, Claude, and open-source models; clients report ~60% operational efficiency gains.
- Design end-to-end ETL pipelines for lead generation and marketing analytics — 60% accuracy improvement and 70% processing time reduction.
- Founded knowtheprice.com.au — full-stack real-time price comparison platform (Next.js, Supabase, Vercel).

TECHNICAL SKILLS

AI & LLM Engineering: LangChain, LangGraph, LlamaIndex, CrewAI, DSPy, MCP (Model Context Protocol), FastMCP, RAG, Multi-Agent & Autonomous Agent Systems, Prompt Engineering, LLM Evaluation, Observability & Guardrails, LangFuse, LangSmith, Pinecone, Weaviate, PGVector, Hybrid/Semantic Search
LLM Providers: OpenAI API, Anthropic Claude, Google Gemini, AWS Bedrock, Open-source LLMs
Programming & Backend: Python, SQL, Bash, FastAPI, Django, Flask, Pydantic, Async Python
Data Engineering: Apache Spark (PySpark), Airflow, Dagster, dbt, Kafka, Delta Lake, ETL/ELT, Data Modeling, Data Warehousing, pandas, Polars, NumPy, Real-time Pipelines
Cloud & Infrastructure: AWS (Bedrock, SageMaker, Lambda, ECS, S3, Redshift), Azure, GCP, Databricks, Snowflake, Vercel, Supabase
MLOps & DevOps: Docker, Kubernetes, MLflow, CI/CD (GitHub Actions), Poetry, pytest, mypy, ruff/black, Git, Linux/Unix, Web Scraping, Celery
Databases: PostgreSQL, DuckDB, MongoDB, MariaDB, Redis, S3, Delta Lake
Visualization & BI: Power BI, Plotly Dash, Streamlit, Matplotlib, Seaborn

KEY PROJECTS

- **Multi-Agent Production Support System (OnTheGoSystems) —** ~10 autonomous agents in production for ~500K users. AWS Bedrock, CrewAI, FastAPI, Docker, ECS; multi-provider routing (OpenAI, Anthropic, Gemini).
- **Hybrid RAG + MCP Data Platform (BetterCollective) —** Unified MySQL, S3, Google Drive, and Sheets behind a natural-language interface with SQL execution and custom dashboard builder. FastMCP, FastAPI, vector DB, Django.
- **Multi-Agent SEO Content Pipeline (BetterCollective) —** ~8 CrewAI agents automating research, analysis, scraping, and writing.
- **Compliance Report AI (ComplianceLabX) —** Production AI app grading client materials against regulatory requirements with source-attributed findings; guardrails and evaluation for groundedness and hallucination rate.