Michael Baylard

AI & ML Engineer

baylardmichael@gmail.com · linkedin.com/in/michaelbaylard · michael-baylard.dev · speaksense.io · klarix.ai · Chicago, IL (Remote)

SUMMARY

Senior AI & ML Engineer with 3+ years shipping production AI systems at scale. Built production RAG pipelines with vector search (pgvector), multi-agent orchestration with parallel tool-calling, and real-time classification pipelines processing 38M+ images on GPU clusters (A10). Founded two AI products: a multi-tenant LLM co-agent with streaming inference and a config-driven agentic pipeline orchestrating 6 LLM providers with cost-optimized model allocation and multi-tier failover. Engineered production data pipelines across billion-row datasets and forecast models achieving company-best MAPE. $82M+ measurable business impact.

TECHNICAL SKILLS

AI / LLM: RAG (pgvector, embeddings, vector search), LLM integration (OpenAI GPT-4, Anthropic Claude, Vercel AI SDK v5), agentic orchestration (MCP servers, LangChain, parallel tool-calling), prompt engineering, LLM evaluation and quality gates, English-to-SQL, streaming inference

Multi-Agent Systems: Multi-model pipelines (MiniMax M2.7 Highspeed, Gemini 3.1 Pro, Claude 4.5 Opus, Groq Compound + Llama 70B, Tavily), cost-optimized model allocation, 3-tier failover, concurrent processing with config-driven routing, JSONL checkpointing, function/tool-calling, 8-tool agent registry

ML / Computer Vision: PyTorch, TensorFlow/Keras, scikit-learn, PyCaret, OpenCV, MobileNet, EfficientNet, SwinTransformer, DINOv2, model compression, human-in-the-loop labeling, FiftyOne (Voxel51), MLflow

Data & Infrastructure: GPU inference (A10/g5 clusters), ETL architecture, Delta Lake, schema governance, Docker, CI/CD (GitHub Actions), AWS (Databricks), Vercel, REST API design (FastAPI, Next.js API routes)

Databases: PostgreSQL (Neon — pgvector, RLS, triggers, JSONB), Databricks (Delta Lake, Unity Catalog), schema design, migrations, query optimization

Languages: Python, TypeScript/JavaScript, SQL, Bash

PROFESSIONAL EXPERIENCE

Speaksense — Founder & AI Engineer

Chicago, IL | 2024 – Present

- Architected production RAG system with English-to-SQL co-agent — parallel tool-calling synthesizes multi-channel analytics in real time via streaming inference (Vercel AI SDK v5) with persistent chat memory
- Built multi-tenant AI platform: 35+ PostgreSQL tables, agency-scoped RBAC (JSONB permissions), pgvector embeddings for semantic search, Stripe subscription billing with webhook-driven plan management
- Registered 8 AI tools in agentic orchestration layer: channel analysis, natural language SQL generation, thumbnail evaluation, competitor intelligence, image generation, community trend analysis
- Integrated 6 MCP servers for automated documentation, database queries, and git workflows — achieved 10x development velocity across planning, implementation, and deployment cycles
- **Stack:** Next.js 15, React 19, TypeScript, Neon PostgreSQL (pgvector), OpenAI GPT-4, Vercel AI SDK v5, Stripe, Docker

Klarix — Founder & AI Pipeline Engineer

Chicago, IL | 2026 – Present

- Architected 4-phase DAG pipeline (Score → Research → Generate → Deliver) orchestrating 14+ Python scripts with 3 parallel execution tracks — config-driven multi-client architecture via context.json (ICP, personas, services, wedges) and YAML configs (trial selection, 8 custom research queries per company)
- Integrated 7 API providers with cost-optimized model allocation: MiniMax M2.7 Highspeed (ICP scoring, contact ranking, quality review), Gemini 3.1 Pro (research synthesis, one-pagers, SWOT, battle cards, event triggers), Claude 4.5 Opus (outreach generation), Groq Compound + Llama 70B (dossier research), Tavily (deep web research), Apollo (org + people extraction), OpenRouter (failover routing)
- Enforced structured JSON outputs across all scoring and extraction stages — schema-driven prompts (15+ field business schema, 10+ field contact schema) with multi-tier parse fallback: direct parse → code block extraction → trailing comma repair → regex object extraction; Gemini native JSON mode for guaranteed structured output
- Engineered multi-tier failover (MiniMax → OpenRouter → Groq), per-API rate limiting (450–950 RPM), ThreadPoolExecutor parallelism (5–50 workers per stage), and crash-safe incremental JSONL processing with thread-safe append-under-lock
- Implemented incremental result caching and lead deduplication (domain normalization + email dedup) — JSONL-based scored-record deduplication skips already-processed records on re-runs, reducing API costs 60–80% for repeat client pipelines
- Generated 7 AI-synthesized deliverable types per prospect: dossiers, SWOT analyses, battle cards, one-pagers, personalized outreach sequences, market overviews, and trade show/event triggers — branded PDF bundling (Playwright) + merge (PyMuPDF)
- Implemented AI quality gate with automated review scoring and flagging across all deliverables before client handoff — scored 17,700+ companies and 6,300+ contacts; 78% verified email rate; full pipeline in under 3 hours (was weeks)

John Deere — AI Engineer / ML Engineer / Data Scientist (FurrowVision)

Chicago, IL | 2023 – Present | Precision Agriculture / Industrial IoT

Production Model Deployment

- Deployed real-time classification pipeline on A10 GPU clusters processing **38M+ sensor images** — F1 improved from 0.44 to **0.92** through 8 months of iterative dataset curation, model compression (SwinTransformer → MobileNetV2), and DINOv2 embedding-based expansion
- Curated 15,000-image training dataset from 38M+ corpus using embedding distance — engineered self-serve workflow enabling **4,000+ images** added post-handoff without data science involvement
- Integrated international (Brazil) sensor data without re-engineering — validated pipeline scalability across geographies

ML Platform & Self-Serve Workflows

- Engineered self-serve dataset curation workflow — labeling throughput improved **3–5x** (SuperAnnotate → FiftyOne migration); engineering team operates independently post-handoff
- Delivered zero-follow-up platform handoff — **13 documentation files** enabling full CVML self-service; praised for "extensive documentation"

Data Pipeline & Backend Architecture

- Built **7-table real-time analytics backend** (raw telemetry → 1Hz aggregation → dashboard-ready) powering field operations across 300+ acre deployments
- Documented **76+ table schemas** across 5 databases — Unity Catalog generators, partition strategies, data lineage; identified **1B-row table with zero partitioning** and flagged 130M-row and 42M-row tables for remediation

Forecasting & Business Impact

- Built used-equipment forecast models achieving **company-best MAPE** across **7 product lines** (PyCaret, Auto ARIMA) — informed pricing and stocking decisions
- Analytics influenced **$82M+ revenue** across FY23–24; defined **17 system metrics** with Systems Engineering consumed by all downstream teams

SELECTED PROJECTS

Production RAG Co-Agent — Speaksense

Next.js 15, TypeScript, Vercel AI SDK v5, PostgreSQL/pgvector, OpenAI

English-to-SQL with parallel tool-calling; streaming chat UX with persistent memory; agency-scoped multi-tenant RBAC isolation

Multi-Model Agentic Pipeline — Klarix

Python, MiniMax M2.7, Gemini 3.1 Pro, Claude 4.5 Opus, Groq Compound, Tavily, Apollo API, OpenRouter, Playwright, PyMuPDF

4-phase DAG with 7 API integrations, 3 parallel execution tracks, 14+ orchestrated scripts; 7 AI-synthesized deliverable types per prospect; 17,700+ companies scored; same-day delivery (was weeks)

Production Computer Vision System — John Deere

Python, PyTorch, TensorFlow/Keras, MobileNet, DINOv2, FiftyOne, Databricks (AWS)

38M+ images classified and segmented; F1 = 0.92 in production; self-serve ML platform for engineering teams; model compression for constrained deployment

Financial Forecasting System — John Deere

Python, PyCaret, Auto ARIMA, scikit-learn

Company-best MAPE across 7 product lines; engineered lag features and external economic signal correlation

CERTIFICATIONS

- **Microsoft Certified: Power BI Data Analyst Associate** — Microsoft