Philip John Basile Resume

PHILIP JOHN BASILE

Principal Agentic AI Engineer | LLM Agents · RAG · MCP · AdTech Automation

New Rochelle, NY | PhilipJohnBasile@gmail.com | 917-650-0552 |
linkedin.com/in/philipjohnbasile | github.com/philipjohnbasile

---

SUMMARY

I've shipped everything from classified Air Force mission planners to Cannes Grand Prix campaigns to production LLM agents, and the common thread is the same: I make ambitious tech actually work.

25 years, 40+ production systems across classified defense, HIPAA-regulated clinical data, a telehealth platform scaling through its NYSE debut, and award-winning global ad campaigns. Most recently: an entire AI function built from scratch — 458 MCP-exposed tools, eval harnesses, governance agents, and the platform architecture powering all client delivery. Production ML and agent design aren't new to me; the LLM era just made the work easier to name.

CORE EXPERTISE

GenAI & Agent Stack: Claude, OpenAI GPT, Gemini, DeepSeek, Llama, Mistral, Jasper; Ollama for local and self-hosted inference; prompt engineering, instruction tuning, context window management, structured outputs; RAG architecture, agentic RAG, Graph RAG, knowledge graphs, vector search (Pinecone, Weaviate, FAISS, Chroma, pgvector); embedding models (text-embedding-3, BGE, Voyage, Cohere embed) and rerankers (Cohere, bge-reranker); NL-to-SQL / Text-to-SQL for agentic data access; compound AI systems, agent orchestration, agent memory architectures (episodic, semantic, procedural); LangChain, LangGraph, CrewAI, LlamaIndex, Haystack, Semantic Kernel, DSPy, Pydantic AI, AutoGen; Anthropic Agent SDK, OpenAI Agents SDK, Google ADK, Computer Use API, MCP server creation, A2A, ACP; Claude skills and plugin creation, multi-agent workflows, tool use orchestration

AI Safety, Governance & Eval: Prompt injection defense, tool sandboxing, governance agents, agent observability, PII/PHI deidentification, OWASP for LLMs, Guardrails AI, Llama Guard, NeMo Guardrails, Bedrock Guardrails, AI red teaming, CI/CD safety

gating, eval-driven quality gates, HITL orchestration, audit logging, clinical decision support systems, model explainability; GDPR, HIPAA, SOC 2, PCI DSS; RAGAS, LangSmith, Promptfoo, Langfuse, Arize Phoenix, Helicone, Portkey; LLM-as-Judge evaluation, SWE-bench, OSWorld; hallucination reduction, retrieval grounding, regression gating, A/B testing for GenAI features, continuous prompt tuning pipelines

AI Infra & MLOps: AWS (Lambda, Bedrock, SageMaker, EC2, S3, CloudWatch, IAM, CloudFormation), Azure (VMs, Functions, Blob, OpenAI, Cognitive Services, ML, Databricks), GCP (Vertex AI, BigQuery, Cloud Functions, Cloud Run); Docker, Kubernetes, Helm, Terraform, GitHub Actions, Jenkins; vLLM, Triton, TensorRT, KServe, BentoML, MLflow, W&B; OpenTelemetry, Prometheus, Grafana, Datadog; PyTorch DDP and FSDP, DeepSpeed ZeRO, mixed precision, gradient checkpointing, CUDA, LoRA, PEFT fine-tuning. Multi-tenant SaaS architecture, per-tenant data isolation, AWS S3-backed content packs. Cost-aware architecture — model routing and fallback, SLM deployment for domain-specific tasks, prompt caching, semantic caching, token efficiency, batching, vLLM inference optimization, serverless vs compute tradeoff analysis, real-time latency budgeting across the full inference stack

AdTech Automation: Google Ads API, Meta Marketing API, LinkedIn Campaign Manager API, DV360, The Trade Desk, Reddit Ads, StackAdapt; budget management (Shape.io, Windsor.ai, Adverity); Google Analytics MCP integration; paid media workflow automation; B2B go-to-market (GTM), ABM/ABX, demand generation and pipeline acceleration; AI-assisted campaign operations, audience workflows, and approval-gated media-spend automation

Full-Stack & Data: Python (uv), TypeScript, Rust, Go, C++, C#, Elixir; FastAPI, Flask + Gunicorn, Node.js, React, Next.js; REST, GraphQL, WebSockets, gRPC, Server-Sent Events (SSE); real-time systems, headless CMS patterns; PyTorch, TensorFlow, scikit-learn, NumPy, pandas, Hugging Face; Spark, Kafka, Airflow, Prefect 3, dbt; Firecrawl for web ingestion, document parsing (Docling, Unstructured.io, PyPDF), Whisper for video transcripts; Snowflake, Databricks, PostgreSQL, Delta Lake, Parquet, feature store patterns

AI-Assisted Development: Claude Code, Cursor, Windsurf, Codex, Devin, Perplexity

Enablement & Productization: Methodology productization — turning domain IP into reusable, packaged AI components (Claude skills, plugins, MCP servers); internal AI training and enablement programs (video, PPTX, interactive web guides); developer education and team onboarding for AI-assisted workflows

---

PROFESSIONAL EXPERIENCE

Principal Agentic AI Engineer · Basilecom

March 2023 – Present — New Rochelle, NY

Founder and principal AI engineer. Primary engagement: ***Transmission Agency****, a global B2B marketing agency, where I lead the full AI engineering function — architecture, model selection, deployment, and governance — across agent platforms, RAG systems, and marketing intelligence tooling. I raise the technical bar, grow the engineers around me, and stay hands-on when the path forward isn't clear.*

Impact

- Delivered AI platforms, agent systems, and GenAI product launches for Transmission Agency and its enterprise B2B client base
- Contributed to $10M+ in client revenue across AI product launches, demand generation, and marketing automation programs
- Defined and shipped Transmission Agency's end-to-end AI stack — MCP orchestration, RAG pipelines, eval harnesses, governance agents — taking the agency from zero AI capability to 458 MCP-exposed tools in under two years
- Drove the AI roadmap that produced the agency's content intelligence platform, aligning product, creative, and engineering around a single architecture now powering all client delivery

Flagship Products

- Built a **multi-tenant natural-language marketing intelligence platform** for Transmission Agency serving 11 client accounts (2.3M rows, 31 Python modules, 80+ REST endpoints, 20+ AI features) with three-tier Qwen 2.5 Coder architecture on local Ollama (7B for SQL + narration, 1.5B for classification) — eliminating per-query LLM cost and keeping data on-prem; engineered **3-layer data isolation** (LLM system prompt + SQL validation + server-side WHERE injection) preventing cross-tenant leakage, with prompt-injection defenses catching jailbreaks and hallucination traps before reaching the LLM
- Delivered the platform's analytics and action layer: NL-to-SQL chat with 5-route intent classifier (DATA/CHAT/WHY/OFFTOPIC/ESCALATION), **Markov-chain attribution modeling** via removal-effect analysis, budget optimizer with marginal efficiency + diminishing returns estimation, scenario simulator for what-if budget reallocation, forecasting, anomaly detection, anonymous cross-account percentile benchmarking, and **agentic auto-actions with confidence-scored approval queues** — plus automated reporting pipelines (monthly HTML, QBRs, weekly digests, gTTS-ready voice briefings, 5-slide video briefings with per-slide narration via MoviePy)
- Shipped a **content intelligence platform** for Transmission Agency using Prefect 3 orchestration with multi-source ingestion (Firecrawl web crawl, PDF parsing, YouTube transcripts); LLM-driven tagging for buyer journey, persona, topics, and quality scores enabled a shift from one-time project engagements to recurring subscription offerings at higher margins; Docker + Terraform infrastructure with GDPR-style data deletion compliance workflows
- Built a **Snowflake MCP server** providing secure, governed warehouse access from Claude Desktop, packaged as a distributable **.mcpb Claude Desktop bundle** with drag-and-drop installation and RSA key pair authentication
- Packaged **23 Claude skills** encoding Transmission Agency's B2B marketing methodologies — audience thresholding, account incrementality calculator, GTM assumption scoper, QBR story coach, BDR account research, estimate builder, weekly intent analyzer, brand voice extractor, data-analysis guardrails, growth intelligence, practice landscape planner, rate card pricer, Bombora audience builder, LinkedIn matched audience, and more — as reusable plugins distributed across agency teams
- Shipped a **bidirectional Claude ↔ Jasper AI MCP integration** (content strategy research + persona/channel content packs) driving a 5-phase content workflow where a single brief spawns 15–25 persona-and-channel variants with human approval gates at quality-critical junctions — reducing content research time 87% and A/B variant creation 98%

- Built multi-tenant knowledge base with per-tenant isolation for account-based marketing — brand guidelines, messaging frameworks, analyst reports, case studies, and structured briefs stored in AWS S3 and synced across AI platforms so models consistently apply per-tenant tone of voice

AdTech & Paid Media Automation

- Published formal API assessments for seven major advertising platforms — **Google Ads, Meta, LinkedIn, DV360, The Trade Desk, Reddit Ads, StackAdapt** — plus MoSCoW analysis of Adverity vs Windsor.ai, defining the technical foundation for Transmission Agency's AI-driven campaign operations at scale
- Built **Google Analytics and Google Ads MCP servers** with team-wide deployment patterns for shared authentication and configuration, making campaign and analytics data a first-class input to AI-driven decisions
- Shipped a campaign-staging pipeline and media-spend guardrail system — daily caps, pacing alerts, approval gates — preventing accidental overspend before automated workflows touch live ad accounts

LLM Agent & MCP Platform

- Developed multi-turn agentic workflows where Claude autonomously researches client sites, generates prototype content, and emits structured JSON briefs and content packs that plug into downstream tools
- Created custom Claude skills and plugins for enterprise content workflows, packaging reusable prompt chains, tool configurations, and domain-specific instructions into deployable components that standardize AI-assisted processes across teams

RAG Systems & Production Infrastructure

- Architected and deployed 5 production RAG systems sustaining ~50 QPS at sub-300ms P95 latency, evolving from static retrieval into agentic RAG patterns where agents dynamically choose retrieval strategies, re-rank results, and trigger iterative retrieval loops, improving retrieval accuracy 40% and cutting inference cost 65%
- Designed serverless RAG architecture on AWS Lambda + Bedrock processing 100K+ queries/month; tradeoff analysis that chose Lambda over EC2 delivered ~70% cost reduction while maintaining latency targets
- Implemented cost-aware model routing and fallback strategies across multiple LLM providers — frontier models for complex reasoning, SLMs for high-volume subtasks, local Ollama for high-frequency analyst queries, prompt caching cutting repetitive query costs up to 90%

Governance, Safety & Engineering Rigor

- Built automated evaluation harness using RAGAS, LangSmith, and Promptfoo with 1,104 passing tests across 29 suites (>80% coverage), modeled on industry benchmarks like SWE-bench and OSWorld, deployed as hard stops in CI/CD to gate hallucination rate, latency, answer quality, and cost per query before anything reaches production
- Implemented prompt injection defenses and tool-use guardrails deployed as always-on governance agents (Llama Guard, NeMo Guardrails) that autonomously monitor agent behavior, flag policy violations, and prevent data exfiltration and unsafe actions across multi-tenant deployments, with full observability tracing agent decision chains, tool call sequences, and reasoning paths via Prometheus and audit logging

- Engineered self-healing agent architectures including circuit breakers, token bucket rate limiting, LRU caching, and priority queues where agents detect failures, reroute workflows, and recover automatically without manual intervention, with iterative self-improvement loops refining agent performance across successive runs
- Enforced strict TypeScript with Zod validation across 77 typed API interfaces, making the full agent surface area auditable and verifiable in CI before deployment

Developer Tooling & Enablement

- Designed CLIs, internal dashboards, and React-based web UIs on top of AI infrastructure so engineers, analysts, and creative teams can explore data, trigger workflows, and understand model behavior without touching raw APIs
- Produced an internal Claude enablement package for Transmission Agency (video sizzle reel, interactive web app, PPTX, PDF) training cross-functional teams on AI-assisted workflows; standardized Claude Code as the agency's agentic coding environment across Python, TypeScript, and other languages

---

Founder & Principal Engineer · Basilecom

April 2001 – March 2023 — New Rochelle, NY

Pre-AI era of the same company. Over 22 years the engagements grew from early web platforms to enterprise search at IBM, classified mission planning for the Air Force, critical infrastructure cybersecurity at Dragos, and global logistics at Atlas Air. Each client brought harder problems, higher stakes, and bigger teams. Engagements ran concurrently with the full-time roles listed below.

- Delivered 40+ production systems with ~90% client retention and 75% repeat business for clients including IBM, U.S. Air Force, Dragos, Phoenix Contact, Publicis Groupe, Atlas Air, Fox Sports, Bayer HealthCare, ADP, IFF, FIS, Shubert Ticketing, and Bremer Bank
- Led IBM global search modernization achieving 3x query performance and 80% infrastructure cost reduction, earning direct commendation from the IBM CTO
- Built green-field mission-planning tool for U.S. Air Force Air Mobility Command (AMC DMR) at Scott AFB, architecting secure micro-frontends and microservices in React, Node.js, and Go within a classified AWS environment, setting technical direction from inception to MVP delivery
- Built the "Hawk" system for Atlas Air, a mission-critical platform centralizing global flight scheduling, cargo logistics, crew management, and compliance with reactive real-time geospatial dashboards using Angular, TypeScript, RxJS, and Azure/.NET
- Integrated AWS SageMaker inference endpoints and Comprehend for predictive insights in fintech workflows, and Azure Cognitive Services, Azure ML, and Azure Databricks for identity management and audience targeting platforms
- Built cybersecurity and data platforms for critical infrastructure at Dragos, combining analytics, dashboards, and workflow automation for ICS and OT security operators
- Built mission-critical Rust and Go platforms achieving 99.9% uptime for defense and logistics clients, including operator CLIs and control panel frontends
- Managed distributed teams of 4 to 20 engineers across 12 time zones; coached 5 engineers to senior roles and mentored developers in Python, TypeScript, and AI tooling
- Entrusted by clients to lead technical interviews and evaluate candidates for full-time and contract engineering hires, shaping team composition across multiple engagements

- 3-year technical advisor to Polywork (Product Hunt Golden Kitty winner, 50K+ users in 48 hours), advising the CEO on product direction, growth, and community

---

Senior Full Stack Engineer & Data Scientist · IntegraMed Fertility

May 2017 – Nov 2018 — Purchase, NY

First dedicated ML role: predictive models on clinical data with real regulatory constraints and real patient outcomes. Built HIPAA, SOC 2, and PCI DSS compliant platform for 50+ clinics processing 40K+ IVF cycles annually with $100M+ in patient financing at 99.9% uptime.

- Developed ensemble ML models (scikit-learn, XGBoost) over 200+ clinical features improving IVF success prediction ~30%, generating $5M+ revenue impact through optimized treatment protocols
- Built real-time dashboards using WebSockets, Redis, and D3 processing 100K+ daily transactions at sub-second latency for 50+ clinics at 99.9% uptime, with zero-downtime integration of 9 acquired clinics and 50K+ patient records
- Migrated 20+ legacy services to Dockerized Node.js/React microservices, cutting infrastructure costs ~40% and compressing deployment cycles from days to hours; built Python ETL pipelines with validation and anomaly detection reducing diagnosis time ~25%

Senior Full Stack Engineer · Teladoc Health

June 2015 – April 2016 — Purchase, NY

Teladoc was sprinting toward its NYSE debut and the platform had to scale as fast as the membership rolls. It did.

- Helped scale telehealth platform through NYSE debut from 12.2M to 15.1M members and 240K+ quarterly visits under real growth pressure
- Led migration from Ruby on Rails to Elixir and Phoenix, leveraging BEAM VM concurrency and fault tolerance for real-time telehealth operations at scale, with Phoenix Channels (WebSockets) for live features
- Built Python ETL pipelines with anomaly detection processing 1M+ patient records/month; developed React/Redux/TypeScript frontend with WebRTC supporting ~50K concurrent sessions at sub-2-second connection time
- Led the API integration of Teladoc's virtual care workflows with CVS Health's MinuteClinic pilot, extending telehealth into retail clinics nationwide; platform achieved the industry's first NCQA telehealth credentialing

---

Senior Full Stack Engineer · BaubleBar

Feb 2013 – Nov 2014 — New York, NY

First technical hire at a venture-backed startup that went from scrappy to $10M+ in revenue on my watch. I built the platform, the celebrity launch infrastructure, and sat on the panel that hired our CTO.

- First technical hire; re-architected legacy e-commerce monolith into a mobile-first platform with Python-driven personalization and Redis caching, driving $10M+ revenue and 30% conversion rate improvement
- Built viral celebrity launch sites (Emma Roberts, Cara Delevingne, Coco Rocha, Ashley Madekwe, Harley Viera-Newton, Nina Garcia, Erin Wasson, and others) with CDN, load balancing, and real-time analytics supporting 100K+ concurrent users
- Served on the executive search panel for the company's new CTO, interviewing and evaluating candidates alongside the board

---

Senior Full Stack Engineer · 360i

Aug 2010 – Feb 2013 — New York, NY

The Dentsu agency was winning Agency of the Year awards and I was the engineer behind the campaigns. Real-time marketing before the term existed, for brands the entire industry was watching.

- Led engineering for Cannes Grand Prix and SABRE Gold winning campaigns during Agency of the Year era: Oreo Daily Twist (including the Super Bowl blackout real-time response), Oscar Mayer Bacon Barter, Coca-Cola Polar Bowl, plus platforms for Marvel, NBC, and National Geographic
- Built real-time marketing platforms with Python/Django, Node.js, Redis, and WebSockets handling millions of concurrent users; analytics and sentiment dashboards enabling sub-5-minute content decisions

---

SELECTED OPEN SOURCE PROJECTS

VecStore | Embeddable Vector Database | github.com/philipjohnbasile/vecstore Lightweight embeddable vector database in Rust with HNSW indexing, hybrid search, and metadata filtering, exposed to Python via bindings for local-first RAG workloads that need fast retrieval without external infrastructure dependencies.

SignalScope | Pharmacovigilance Signal Detection Engine | github.com/philipjohnbasile/signal-scope Offline AI system for drug safety signal analysis, ingesting FDA FAERS case reports and PubMed literature. Combines weakly supervised NLP, disproportionality metrics with Bayesian shrinkage, and local LLM summarization in a fully deterministic, CPU-optimized Rust architecture. Designed for regulated environments where reproducibility and auditability matter more than speed.

PhilJS | Experimental JavaScript Front-End Framework | github.com/philipjohnbasile/philjs Reactive UI and component patterns exploration.

EDUCATION & CERTIFICATIONS

Bachelor of Science in Computer Science, Fordham University, New York, NY — [YEAR] Machine Learning Specialization, Stanford / Coursera — [YEAR] Deep Learning Specialization, DeepLearning.ai — [YEAR]

ADDITIONAL

- Provide fractional CTO services to creative-economy and early-stage companies, contributing to multiple Product Hunt awards
- 15-year volunteer with Civil Air Patrol, U.S. Air Force Auxiliary (security clearance eligible)