

GEORGE SMITH

Applied AI Solution Architect | Senior Data Scientist | AI Product Leader

Berlin, Germany | +49 1762 8232143 | professorgeo@outlook.com | linkedin.com/in/geo-smith-a296b83 | github.com/GeoTheLeo | geotheleo.com

EXECUTIVE PROFILE

Senior Data Scientist and Applied AI Solution Architect with **14+ years specializing in Machine Learning, Predictive Analytics, and Applied AI** across education, SaaS, and digital-transformation initiatives.

Specialized in translating complex business challenges into scalable AI solutions using **Machine Learning, Predictive Analytics, Natural Language Processing, Retrieval-Augmented Generation (RAG), Agentic AI Workflows, MLOps, Lakehouse Data Engineering, Analytics Engineering, and Executive Business Intelligence.**

A decade-plus track record **leading data science and analytics teams** — across Lead Data Scientist, Data Science & Analytics Lead, and Analytics Program Manager roles — setting technical direction, mentoring team members, and raising the quality bar for team output.

Builder of a growing portfolio of self-directed Applied AI platforms — three **publicly deployed** (EdTech decision intelligence, cross-industry churn prediction, real-time speech ML) plus two **agentic AI systems with human-in-the-loop approval**, one running on a **PySpark / Delta Lake lakehouse** — demonstrating end-to-end design from data engineering and modeling through agent orchestration and deployment.

SELECTED CAREER ACHIEVEMENTS

- Designed, built, and deployed the **NorthStar Enterprise AI Decision Intelligence Platform** end-to-end — executive BI dashboard, predictive at-risk-student classifier, learner segmentation, a **RAG-based knowledge assistant**, and a **model registry/monitoring layer** — containerized with Docker and shipped with a GitHub Actions CI/CD pipeline.
- Built and deployed the **Retention Intelligence Suite**, a cross-industry churn-prediction platform applying one shared ML core across **five industries** (music, education, healthcare, e-commerce, gaming) with zero per-industry code changes.
- Led development of **20+ AI-driven learning products** and analytics initiatives, improving learner retention by **25%** while reducing product iteration cycles by **40%**.
- Mentored **50+ educators and professionals** in analytics, statistical methods, research methodology, and evidence-based decision making, strengthening organizational capability and fostering data-driven cultures.

INDEPENDENT AI / PORTFOLIO PROJECTS

NorthStar — Enterprise AI Decision Intelligence Platform (Self-Directed)

2025 – Present

[Live Demo](#) · [GitHub](#)

- Architected and built a modular platform end-to-end — executive BI dashboard, a **Random Forest** at-risk-student classifier, **KMeans** learner segmentation, a **RAG-based knowledge assistant**, and a model-registry/monitoring layer — sharing a common data contract across services.
- Hardened the RAG assistant: **Qdrant** vector search, three **hallucination guardrails** (relevance gate, cited answers with explicit refusal, LLM grounding check), and **OpenTelemetry** tracing; golden-set evals show **100% correctness and 0 hallucinations**.
- Containerized with **Docker** and deployed via a **GitHub Actions CI/CD** pipeline that builds and tests on every push; live, publicly deployed **Streamlit** executive dashboard.

NorthStar Agentic Intervention Copilot — Agentic AI on Next.js/TypeScript

[GitHub](#)

- Two-phase Claude tool-calling agent (explicit **planning phase**, then a streamed **tool-use loop**) whose tools call NorthStar's real trained models via **FastAPI**. Its only side-effecting tool queues a proposal; anything real happens in a Server Action triggered by human approval, outside the LLM's control loop. **Next.js, React, TypeScript**, Playwright end-to-end tests.

Music Man — Music-Streaming Royalty-Fraud Lakehouse with Agentic Investigation

[GitHub](#)

- Engineered a **medallion lakehouse (bronze → silver → gold)** on **PySpark and Delta Lake** — **89,740 real tracks / 31,428 artists** seeded from a public Spotify dataset, plus **2.1M synthetic play events** with four injected bot-farm fraud signatures — cleaned, deduplicated, sessionized, and aggregated into per-artist daily fraud-signal features (play velocity, device/IP concentration, geo-impossible plays).
- Built a **Neo4j fraud-ring graph** (9,122 accounts, 14,002 devices; **Graph Data Science** connected components, **Cypher**) that exposed a labelling bug: every apparent false positive was day two of a real bot attack. With corrected labels, the **MLflow**-tracked **Gradient Boosting** classifier scores **precision 1.0, recall 1.0** (5-fold CV 1.0 / 0.98); the agent gained a `find_fraud_ring` tool. Portable to **Databricks**.
- A **Claude plan-then-execute agent** investigates flagged artists and can only propose royalty holds; nothing is withheld until a human approves in a **Streamlit** review app. Hardened with a compare-and-set state machine, rollback path, idempotent proposals, and dry-run mode. An **agentic eval suite** (LLM-as-judge, golden transcripts, red-team prompt injection) passed **20/20 runs**, and measured **model routing** cut cost per run by **25%** with no quality loss.
- Moved every guardrail into one shared operations layer, then exposed it three ways: an **MCP server** (any MCP client inherits the guardrails), a **LangGraph** agent (interrupt-based human review, SQLite checkpointing with resume), and an HMAC-signed, idempotent **webhook** trigger. The LangGraph agent matched the hand-written loop in evals (8/8 passes, \$0.072 vs. \$0.074 per run).

Retention Intelligence Suite — Cross-Industry Churn-Prediction Platform

[Live Demo](#) · [GitHub](#)

- One shared ML core (**XGBoost, KMeans, SHAP**) applied across **five industries** via adapters with zero per-industry code changes (**AUC 0.63–0.69** on all five), plus a live **Streamlit** dashboard whose explainable-AI advisor answers executive questions from real model output (per-segment stats, SHAP drivers).

AI Pronunciation Coach — Real-Time Multilingual Speech-Coaching Tool

[Live Demo](#) · [GitHub](#)

- Built a real-time, multilingual pronunciation-coaching tool using **OpenAI Whisper** for speech recognition, text-similarity scoring against target phrases, and **GPT-4o-mini**-generated coaching feedback, delivering live audio feedback rather than a static score — publicly deployed with a FastAPI backend and an interactive live demo.

CORE EXPERTISE

Applied AI: Machine Learning • Predictive Analytics • LLMs • RAG • Agentic AI (Tool-Calling, Human-in-the-Loop, LangGraph, MCP) • Vector Search (Qdrant) • Graph Analytics (Neo4j, GDS, Cypher) • AI Evals & Hallucination Guardrails • AI Observability (OpenTelemetry) • NLP • Intelligent Automation

AI Architecture & Leadership: AI Solution Architecture • Modular Design • MLOps • API Integration • Executive BI • Product Strategy • Team Leadership & Mentoring • Stakeholder Leadership

Data Science: Classification • Regression • Clustering • Recommendation Systems • Feature Engineering • Experimentation & A/B Testing

Data Engineering: Lakehouse / Medallion Architecture (Bronze-Silver-Gold) • PySpark • Delta Lake • Batch ETL • MLflow Experiment Tracking & Model Registry

SaaS & Cloud Platforms: Multi-Tenant Systems Architecture (AWS-trained) • Cloud Authentication • LMS/SaaS Platform Administration • SaaS Adoption & Customer Success • Data Privacy Compliance

HR & People Systems: PeopleSoft • Oracle HCM • LMS/HR Modules in Education Platforms • Custom-Built Internal HR & CRM-Style Tools • Staff & Contract Tracking • Scheduling • Performance Reviews • Recruitment Pipelines

Platforms & Tools: Python, TypeScript, Next.js, React, R, SQL, NoSQL, PySpark, Delta Lake, MLflow, Scikit-Learn, XGBoost, TensorFlow, PyTorch, SHAP, FastAPI, Streamlit, Power BI, Tableau, Microsoft Excel (expert) & Office Suite, Docker, Kubernetes, AWS (SageMaker), GCP, Azure, GitHub Actions (CI/CD), OpenAI API (GPT-4o-mini, Whisper), Anthropic API (Claude), LangGraph, MCP (Model Context Protocol), Qdrant, Neo4j (Graph Data Science, Cypher), OpenTelemetry, ColdFusion (early career)

PROFESSIONAL EXPERIENCE

Data Science Consultant, Shixi International School / Suzhou International School

2022 – 2024

- Designed and deployed **Transformer-based NLP pipelines** using Python and modern ML frameworks, processing **5,000+ learner interactions/month** while reducing manual analysis by **20+ hours/week**.
- With school leadership, built **RAG** and **LLM** workflows on Baidu and Alibaba LLMs and a self-hosted **ChatGLM-6B**, owned **end to end in production** (live, used by staff, maintained by me); when staff flagged wrong answers, **constrained them to answer only from retrieved documents**.
- Built classification, recommendation, and predictive analytics models using **Python, SQL, Scikit-Learn, and XGBoost**, supporting data-driven operational decisions.
- Established reproducible **MLOps-ready ML workflows** and administered cloud-based learning platforms (LMS), incorporating feature engineering, model evaluation, SaaS subscription management, and deployment best practices.
- Managed team HR workflows in **enterprise HR systems and LMS/HR modules** (staff and contract tracking, scheduling, performance reviews, and recruitment pipelines) for teams of **7–25** within a larger organization.

Data Science & Analytics Lead, Wall Street English & Meten China – Shanghai, Beijing China

2018 – 2022

- Developed **predictive machine learning models** using Python, Scikit-Learn, and XGBoost, deployed via **AWS SageMaker** within China's regulated cloud environment, to identify at-risk learners and enable earlier intervention.
- Designed a multilingual **200,000+ sentence NLP dataset** (85%→99% annotation consistency) and built **behavioral segmentation and recommendation models** that increased recommendation accuracy by **12%**.
- Automated analytics pipelines and executive reporting with Python, SQL, and Power BI, hosted on **Microsoft Azure (operated via 21Vianet in China)**, improving reporting consistency and reducing manual preparation.
- Managed cloud-based LMS platforms and **SaaS subscription adoption** across a multi-site organization, ensuring data privacy compliance and secure cloud user access at enterprise scale.
- Ran team people operations for teams of **7–25** in **enterprise HR systems** and LMS/HR modules (contracts, scheduling, performance reviews, recruitment pipelines) within the wider multi-site organization.

Lead Data Scientist & Analytics Program Manager, The English Center / Argosy University

2012 – 2018

- Developed predictive models that contributed to a **7% reduction in failing grades**, enabling earlier evidence-based intervention.
- Built and scaled classification and evaluation models — orchestrated with **Kubernetes** — achieving **92% agreement** with stakeholder assessments, improving learner performance evaluation.
- Led adoption of **SaaS-based learning management systems** and digital training platforms, driving curriculum development and instructional design within cloud environments.
- Managed staff, contract, scheduling, and performance-review workflows across **7 sites statewide (California)**.

Senior Data Product Manager & Analytics Lead, Concurrent Consulting Engagements

2012 – 2016

- Designed experimentation frameworks and performance analytics on **Google Cloud Platform (GCP)** that improved learner retention by **25%** while reducing product iteration cycles by **40%**.
- Applied behavioral analytics, statistical analysis, and customer insights — including **SaaS subscription adoption and customer-success support** for institutional clients — to prioritize product enhancements and guide evidence-based product strategy.

EDUCATION

Data Science & Artificial Intelligence Program, Ironhack — Graduated May 2026 (1,600+ hour, year-long program), Final Assessment: Exceeding Expectations.

Full data science/AI lifecycle coursework across **Python, R, SQL, NoSQL, Pandas, NumPy, TensorFlow, PyTorch, Scikit-learn, Power BI, Tableau, Docker, Git**, and major cloud platforms — including **large language models, speech models, and computer vision**, plus cloud-based MLOps — culminating in a comprehensive capstone project.

California State University, East Bay — M.A. TESOL (2014); B.A. English & Linguistics, Minor in Creative Writing (2012)