

Tommaso Galeazzo

Los Angeles, CA | tommaso.galeazzo@gmail.com | +19493378083 | [linkedin.com/in/tommaso-galeazzo](https://www.linkedin.com/in/tommaso-galeazzo)
US permanent resident (no visa sponsorship required)

PROFESSIONAL SUMMARY

Senior AI Engineer building and deploying production LLM agents and RAG systems inside a regulated pharmaceutical enterprise, on top of a decade of applied machine learning in science. Owns use cases end to end: translating ambiguous business problems into agentic workflows with clear success criteria, building the evaluation frameworks that prove they work, and shipping them as reliable, observable, multi-user platforms that are adopted daily. Sole author of a production async agentic library on the OpenAI and Anthropic SDKs, and technical lead of an AI co-scientist benchmarked at ~86% on BixBench. Comfortable flexing across the full stack, including frontend, and working directly with business stakeholders, scientists, and executives.

CORE SKILLS

Agentic AI: multi-agent workflows, plan-and-execute orchestration, parallel tool calling, Model Context Protocol (MCP), agent skills, structured outputs, sandboxed code execution, Claude Agent SDK, OpenAI SDK, Anthropic SDK

Retrieval: retrieval-augmented generation (RAG), GraphRAG, natural language to Cypher, Elasticsearch, Neo4j, PubMed retrieval tooling

LLM evaluation: benchmark harnesses, containerized evaluation environments, scorer design, LLM-as-judge, rubric-based trace evaluation, failure analysis, cost and latency control, vendor evaluation

Backend and infrastructure: Python, asyncio, FastAPI, Celery, Redis, relational database schema design, S3, Docker, Kubernetes (EKS), GitHub Actions, ArgoCD, GitOps CI/CD, AWS

Frontend: React, TypeScript

Machine learning: graph neural networks, XGBoost, nested cross-validation, NLP text classification and NER

Working style: customer discovery, success-criteria definition, stakeholder management, executive communication, technical leadership, mentoring

PROFESSIONAL EXPERIENCE

AbbVie | Senior AI Engineer, Applied AI | Aug 2023 - Present

- Built and deployed, as sole engineer, an AI screening platform for biotech firms screening at a top-3 pharma's venture arm: cut screening time 98% (4 hrs → 5 min), freed ~1 FTE of analyst capacity for due diligence, and saved ~\$190K/yr with a 3-month payback and 75% team adoption.
- Ran discovery with sponsoring partners, defined success criteria and evaluation methodology, and shipped a 3-week MVP that secured funding; productionized it, onboarded users, and turned post-launch feedback into a funded 2.0 roadmap.
- Architected multimodal document extraction across 3 parallel streams, LLM-driven entity resolution at ingestion, parallel market/IP/funding analyses synthesized into one recommendation, and a per-user ReAct agent with streaming retrieval over internal data, literature, and sources.
- Raised extraction accuracy from 70% to 92% by encoding SME definitions (biology & finance) into structured-output schemas, backed by a trace-based eval loop scoring persisted extractions; shipped on FastAPI, React/TS, Celery/Redis, JWT auth, and GitOps CI/CD (GitHub Actions, ArgoCD) to EKS.
- Designed and built, as sole author, a production async agentic AI library over the OpenAI and Anthropic SDKs (conversation management, parallel tool execution, streaming, MCP, skills, multimodal input, sandboxed Python execution) that became the foundation of multiple production services serving ~50 users; team's point of reference for generative AI and LLMops for three years.
- Led three generations of an oncology AI co-scientist: an internal reimplement of Google's co-scientist (hypothesis generation, Elo-rated tournaments, LLM-as-judge), a custom orchestrator/planner/executor multi-agent pipeline, and a Claude Agent SDK system with skills, MCP, literature deep research, and GraphRAG over an ontology-anchored Neo4j graph of TCGA (NL-to-Cypher with query verification).
- Built the benchmark infrastructure (per-task Docker sandboxes, reconstructed environments, automated scorers) and closed a domain-reasoning gap found through failure analysis, reaching 86% on BixBench-Verified-50, within 3 points of a leading commercial platform (88.7%); presented as a poster at a Microsoft Research AI4Science workshop (Aug 2026).

- Reused this infrastructure to co-lead benchmarking of external AI-for-science vendors, informing a contract extension with one vendor and the exit of another; now advising executives as the engineering-side technical expert on evaluation design in a joint pilot with a major AI provider.
- Built and operate two production agentic API services on EKS consumed by multiple internal applications: a multimodal document-extraction service (~1 minute per 60-page document) and a literature-research service with semaphore-bounded LLM relevance filtering and quota-based early stopping for cost and latency control (research reports in 2-3 minutes versus hours manually).
- Building an AI assistant for a corporate governance team that guides employees through governance requests, combining Elasticsearch RAG over policy documents with two MCP-based data-retrieval integrations; in active development toward MVP.
- Managed a summer intern who converted to a three-month contractor, scoping and directing their work on the co-scientist.

University of California, Irvine | Research Scientist, Postdoc Scholar | Sep 2019 - Aug 2023

- Applied machine learning to molecular property prediction in atmospheric chemistry: graph convolutional neural networks, neural networks, and XGBoost on mol2vec embeddings, with nested cross-validation for small datasets.
- Built and harmonized training datasets through web scraping of heterogeneous sources at scale.
- Deployed a published model as a public prediction service with FastAPI on AWS ECS.
- Continued numerical modeling in FORTRAN for chemical kinetics during the first phase of the appointment.

With Intelligence (now S&P Global) | Machine Learning Engineer (contract) | Aug 2020 - Sep 2020

- Financial text classification and named-entity recognition with multi-channel CNNs and RNNs on word embeddings, in a team of four.

EDUCATION

- PhD, Physical Chemistry, Sorbonne University, Paris. Wrote numerical ODE solvers in FORTRAN (Runge-Kutta methods) for chemical kinetics simulation.
- MSc, Physical Chemistry, University of Copenhagen
- BSc, Chemistry, University of Padova

PUBLICATIONS AND PRESENTATIONS

- "CURIE: Evidence grounded AI Co-Scientist for oncology"; Poster, AI Scientist Summer Workshop, Microsoft Research, Boston, Aug 2026.
- Krüger, M.*; Galeazzo, T.*; Eremets, I.; Schmidt, B.; Pöschl, U.; Shiraiwa, M.; Berkemeier, T. Improved vapor pressure predictions using group contribution-assisted graph convolutional neural networks (GC²NN). *EGUsphere* 2025, 18, 7357-7371 (2025).
- Galeazzo, T. & Shiraiwa, M. Predicting glass transition temperature and melting point of organic compounds via machine learning and molecular embeddings. *Environ. Sci.: Atmos.* 2, 362–374 (2022).