Trajectory

Started as a quant — built Bloomberg's #1 US corporate-bond pricing model — then spent a decade pivoting into hands-on software engineering. Self-taught on the engineering side, with a Princeton OR/FE PhD and a Stanford math professorship sitting underneath — useful when a model needs to be more than vibes.

What I build

I love building real systems end-to-end — CI, architecture, developer experience, the way teams ship — not just code. Most at home shipping AI infrastructure, agent runtimes, and the platform work around them, with small technical teams that care about craft. Happy to nerd out on DDD, microservices, or the Google SRE canon. Currently reading the AI agent literature widely and prototyping a next-generation agent runtime I think can do quite a bit better than what's out there.

Professional Experience

March 2026 – June 2026 — **Member of Technical Staff**, ***TensorZero***, New York

TensorZero is an open-source LLMOps stack for industrial-grade agentic LLM applications: an LLM gateway unified with observability, evaluation, optimization, and experimentation.

- Worked on the LLM gateway and observability layer: unified inference API across providers with routing, fallbacks, and structured tool-use; inference traces with downstream metrics and natural-language feedback, exported via OpenTelemetry (OTLP) and Prometheus.
- Contributed to evaluation and optimization workflows: curating datasets from production traces and replaying historical inferences against new prompts, models, and strategies.
- Improved the developer experience for LLM engineers: typed schemas for prompts and tool calls, experimentation primitives, and feedback loops that turn production data into model improvements.

October 2024 – January 2026 — **Senior Founding Engineer**, ***PointOne***, New York

PointOne uses AI to passively track time — ending revenue leakage across practice, business, and people.

- Raised release confidence: built the DevOps/test infrastructure that grew coverage 5x, with scenario-based full-pipeline and post-deployment integration tests.
- Collapsed the developer iteration loop from ~15 minutes to near-instant by re-architecting the core to run flows locally.
- Owned cross-platform monitoring end-to-end: redesigned the modular monitoring stack in Go and delivered new ingestion paths in the core product.
- Set the engineering bar for the team: OpenTelemetry-based observability on Kubernetes, LLM recorders for model testing, and mentored juniors into the practice.

July 2023 – Current — **Founder**, ***codefly.ai***

codefly.ai turns developer intent into a working microservice architecture — a DAG of agents owning code, infrastructure, and tooling.

- Agent runtime and CLI in Go with SDKs in Go, Python, TypeScript, and Ruby; toolbox plugin model (gRPC, Nix, Docker, web, Python REPL) so agents call typed tools instead of shelling out.
- 30+ service and infrastructure modules across Postgres, Redis, S3, DynamoDB, Temporal, KrakenD, Envoy, and Pulumi/AWS.
- Currently reading widely in the AI agent literature and prototyping a next-generation agent runtime in stealth, with supporting experiments on LLM tool-calling, retrieval, and agent UX.

October 2022 – June 2023 — **Staff Engineer**, ***Illumio***, Sunnyvale

Joined the Cloud Secure team to architect the next version of the product, addressing structural issues in the existing distributed monolith. The design decisions below are still in use today.

- Shipped the team's first real CI/CD: manual releases to continuous deployment via Go CLI / Jenkins / ArgoCD.
- Re-architected the service layer (gRPC + API Gateway pattern): new endpoints in minutes instead of days.
- Re-modeled Cloud Inventory in Neo4j: dependency graph that let new object types ship in hours instead of weeks.

– New York City
650-575-6287 • *toussaint.antoine@gmail.com* • *antoinetoussaint.com*
Permanent resident – github.com/AntoineToussaint – linkedin.com/in/toussaint-antoine

September 2021 – July 2022 **Staff Engineer**, ***Shopify***, New York
- Joined to lead a new initiative on the company-wide ledger team; reached prototype before the team was redirected to other priorities and the work was wound down.
- Led the Code Yellow existing-merchant MFA workflow: designed a Task Scheduler and drove the implementation across the team.

April 2020 – September 2021 **Lead Engineer**, ***infima.io***, New York

Prediction platform for agency MBS investors.
- From zero to CD as the entire backend team: designed, implemented, and deployed the infima backend (20 KLOCs) on Go/gRPC, gRPC-gateway, Postgres, TimescaleDB, Kubernetes on EKS.
- Built and maintained the infima.io Python libraries, internal and public (`pip install infima-client`).
- Coordinated three frontend consultants across corporate site, platform app, and Excel Add-In.
- A Markov transition classifier POC for unbalanced datasets, speeding up loan-model training by 40%.

2017–2020 **Hands-on Team Lead, ROAR (Go)**, ***JPMorgan Chase***, New York

ROAR: an internal real-time Kaggle — prediction APIs that get better with time, for the business.
- Designed and implemented (15K LOCs) internal ROAR in Go: micro-services on JPMorgan's internal cloud, Kubernetes on EKS.
- Prototyped and oversaw the Web client (ReactJS, OpenID Connect).
- Python client library and ROAR ML library (a time-series prediction framework over sklearn / statsmodels).

2016–2017 **Managing Director**, ***Elefant Markets***, Puerto Rico

FINRA broker-dealer startup in fixed-income algorithmic trading.
- Re-designed and implemented the company's data-flow framework for the trading system: thread-safe / type-safe C++11 library with Lua runtime binding.

2013–2016 **Head Quant-Developer**, ***Bloomberg LP***, New York

BMRK: the world-leading real-time pricing engine for OTC fixed-income products.
- **Created the financial model that became the #1 pricing source for US corporate bonds.**
- Designed the real-time pricing solution (C++) with runtime flexibility, expressing models as abstract syntax trees.
- Built the Python calibration framework (issuer curves, liquidity adjustment as XML AST) and the scenario engine — TDD at the pricing-engine level.

2010–2013 **Associate VP**, ***Benchmark Solutions***, New York

Warburg-Pincus–backed fintech generating market-calibrated prices for US corporate bonds and CDS; IP later acquired by Bloomberg LP.
- Developed (C++) Benchmark Solutions' patented End-of-Day sensitivities solution.
- Conceived and implemented the 'Force', the real-time correction process for the pricing engine (C++).

Side Projects

mind	A coding agent built for collaboration with real engineering teams (in progress).
codefly	Runs your whole microservice backend with one command, from the dependency graph — codefly.ai.
lazybox	Reactive PR inbox in your terminal, in Rust: GitHub events stream to you, and every PR opens into an isolated git worktree with an embedded terminal running Claude Code. Open source — lazybox.ai.
research	384-run study of LLM tool selection: a two-phase design (cheap model picks the tool, smart model fills the arguments) is ~14× cheaper per successful call at ~150-tool scale.
hacker news	pretty-hackernews: a calmer Hacker News reader (Vite, React, Tailwind).

Education/Academia

2007–2010 **Szegö Assistant Professor of Mathematics**, ***Stanford University***

2003–2007 **Ph.D. in Operations Research & Financial Engineering**, ***Princeton University***

Dissertation *Hedging under L^2 convex risk measures* [Advisor: Professor Ronnie Sircar]

Computer skills

Experienced	Golang, Python, C++, AWS, Kubernetes, SQL, OpenTelemetry, gRPC
Intermediate	TypeScript, React, Node, Rails, Neo4j, Pulumi, Temporal

– New York City
650-575-6287 • *toussaint.antoine@gmail.com* • *antoinetoussaint.com*
Permanent resident – github.com/AntoineToussaint – linkedin.com/in/toussaint-antoine