PATRICIO BARRETO

Full-Stack Engineer | AI Product Engineer | Founding Engineer | AI Engineer | Software Engineer | AI Agents, LLMs, RAG and Evals in Production | 0 to 1 at Early-Stage Startups

patriciobarretob@gmail.com | +1 617-794-3227

linkedin.com/in/patricio-barreto-bentin | github.com/patriciobarreto | patriciobarreto.com

PROFESSIONAL SUMMARY

As the first and only engineer at Soundboard, a U.S. tax and bookkeeping startup, I built its multi-tenant SaaS platform and 5 production AI agents from scratch, supporting the company's growth from USD 1K to USD 50K MRR and 177 paying clients. I designed LLM systems end to end (agent orchestration, RAG with hybrid search, evals and observability): I raised transaction categorization accuracy from 90% to 99% at 15x lower cost per transaction and built a tax-return agent that matched the CPA on 34 of 35 returns. For more than two years I also served as Product Manager: I prioritized from user conversations, decided what to build and what to decline based on impact and cost, and closed 15 of the first clients through outbound sales. I combine full-stack engineering (TypeScript, Next.js, Node.js, Python and AWS), AI-assisted development (Claude Code and Codex) and product judgment to take AI products from 0 to 1: I compute in code everything that can be calculated and use the model only for judgment calls.

PROFESSIONAL EXPERIENCE

SOFTWARE ENGINEER (FIRST ENGINEERING HIRE) - SOUNDBOARD INC.

2023-07 - Present | United States (Cambridge, MA; remote since 2025-09)

- **USD 1K to USD 50K MRR and 177 paying clients:** built from scratch the multi-tenant bookkeeping and tax SaaS platform (Next.js, React, TypeScript, Node.js, PostgreSQL and Prisma), with 100+ data models, 70+ REST endpoints and OAuth2 integrations with QuickBooks Online, Stripe and HubSpot, including a two-way QuickBooks sync on undocumented endpoints that now connects 173 companies, at a hosting cost of USD 130 per month (Vercel and AWS). Took the first client to production one month after the first commit.
- Designed sb-ai, the platform that runs all 5 production AI agents: TypeScript orchestration (Node.js and Express) and a Python LLM service (FastAPI) on AWS ECS Fargate, containerized with Docker and deployed through CI/CD on GitHub Actions, with multi-provider LLM routing (Claude, GPT and Gemini), structured outputs validated in code (100K+) and RAG with hybrid search on PostgreSQL and Qdrant (94K indexed chunks). Added BugBot, a triage agent that grouped 5,578 production errors into 1,917 clusters with a daily diagnosis (Vercel, Datadog and Claude, USD 0.22 per run) and cut my fix time by about 30%.
- Rebuilt the transaction categorization agent (about 6,300 transactions a month), which frees about 260 bookkeeping hours a month (USD 75K a year): raised accuracy from 90% to 99%, measured against bookkeepers' review of every transaction, cut latency from 20 to 3 seconds and reduced cost per transaction from USD 0.060 to USD 0.004 (15x cheaper) by replacing the OpenAI Assistants API with hybrid search and selecting Claude Haiku after testing it against a golden dataset.
- Built a Form 1120-S tax preparation agent that matched the CPA on 34 of 35 validation returns, computing 50 of 155 lines in code and leaving the LLM only judgment calls, and kept agent quality in check with golden datasets (5,425 CPA-verified return lines, 1,635 Ops Q&A cases and 532 categorized transactions), regression tests before every change, retrieval evaluation and observability with Arize Phoenix over OpenTelemetry.
- Cut prep time for each monthly client call from 30 to 5 minutes (about 75 hours a month) with a summary pipeline that pulls from 7 sources (QuickBooks, HubSpot notes, meetings and emails, platform questions, the onboarding form and the deductions assistant) and refreshes for USD 0.08 instead of USD 1.20, and launched an Ops Q&A agent that drafts about 25 messages a week and speeds up client replies by about 40%. Both came out of monthly calls with bookkeepers and tax preparers in which I served as Product Manager for more than two years, shipping about 20 features a month.

SOFTWARE ENGINEERING CO-OP - NTT DATA

2022-01 - 2022-07 | United States (Boston, MA)

- **20% more tests per day:** proposed replacing step-by-step screenshots with a screen recording of the full flow as test evidence, which sped up defect reporting and raised QA team productivity in Agile sprints.
- Designed test cases from user stories for an energy sector client's software and optimized them to make testing more effective.
- Ran end-to-end manual functional tests as an end user, 1 flow (one feature) per day, and reported each defect with its evidence to a senior tester, as part of Northeastern University's co-op program.

SOFTWARE DEVELOPMENT INTERN - EVERIS

2021-04 - 2021-08 | Peru (Lima, remote)

- **3 client web applications:** developed full-stack features (Java and Spring Boot backend, HTML and CSS frontend) in an Agile team that designed and implemented web applications for clients.
- Implemented features for a banking client's task management tool used to log tasks and track progress.
- Adopted software design best practices working with the team lead, 100% remotely.

EDUCATION

B.S. in Computer Engineering and Computer Science (Double Major), Minor in Mathematics

NORTHEASTERN UNIVERSITY (Boston, MA, United States) | 2018-08 - 2023-05

SKILLS

AI Agents, Agent Orchestration, Tool Calling, Large Language Models (LLMs), Generative AI, Retrieval-Augmented Generation (RAG), Hybrid Search, Re-ranking, Embeddings, Vector Databases (Qdrant), LLM Evaluation (Evals), Golden Datasets, Regression Testing, LLM Observability (Arize Phoenix, OpenTelemetry), Structured Outputs, Prompt Engineering, Multi-provider LLM Routing (Claude, GPT, Gemini), Anthropic API, OpenAI API, Gemini API, LangChain, Document AI and OCR (Mistral OCR), AI-Assisted Development (Claude Code, Codex), Full-Stack Development, TypeScript, JavaScript, Python, SQL, Java, C++, React, Next.js, Tailwind CSS, HTML, CSS, Node.js, Express, FastAPI, Spring Boot, REST APIs, OAuth2, PostgreSQL, Prisma ORM, Multi-tenant SaaS, System Design, Software Architecture, Software Engineering, AWS (ECS, Fargate, S3, CloudFront, IAM), Docker, CI/CD (GitHub Actions), Vercel, Datadog (RUM), Jest, pytest, Playwright, Selenium, QuickBooks API, Stripe, HubSpot, Product Development, Product Discovery, Prioritization, 0 to 1, Fintech, Accounting and Bookkeeping, Tax Technology, Agile.

LANGUAGES

English - Native | Spanish - Native