Fikresilase Wondmeneh Abebayew

Full-Stack AI Engineer

Addis Ababa, Ethiopia | +251 940 259 534 | shitubilbila@gmail.com | LinkedIn | GitHub | Portfolio

Summary

Full-stack AI engineer with four years shipping and operating production systems, from client-facing product work to enterprise AI at Ethiopia's largest private bank. Sole author of two arXiv preprints and the open-source agent implementing one of them.

Experience

Bank of Abyssinia | AI Engineer *Jul 2025 – Present*

- Rebuilt the internal knowledge assistant from vector RAG into an agent system with hybrid retrieval, reranking, and claim-level verification, after plain retrieval failed on questions spanning multiple documents.
- Architected and built it end to end as the initial sole engineer: ingestion, vector indexing, agent orchestration, FastAPI and Next.js, end-to-end tests, GitHub Actions CI/CD, and Prometheus/Grafana/OpenTelemetry observability on RHEL. Now maintained by a two-engineer team.
- Adopted by 6,790 employees (**nearly 70% of staff**), with monthly actives growing to 6,314, over ∼3,000 documents and 1.2M+ pages.
- Led agent and data orchestration for a four-engineer reporting platform, cutting the 2025/2026 annual report from a four-week, 26-person drafting cycle to a **four-hour run**, with domain experts verifying every generated claim before executive sign-off.
- Implemented evaluation and serving for the voice and text banking agents: nine-category coverage (task success, ASR word error rate, authentication compliance, containment/escalation, latency, hallucination); quantization and multi-token prediction made inference **3.5x faster**.

Defence Construction Enterprise | Lead Full-Stack Developer, Contract *Jan 2025 – Jun 2025*

- Led delivery of document-processing, approval, quality-audit, and employee-management modules for a Next.js, Spring Boot, and PostgreSQL ERP built from scratch in six months within a six-engineer team; now the operational system for ∼2,400 employees.

TAP to Sign | Full-Stack Developer & Customer Success Engineer *Jul 2023 – Jan 2025*

- Migrated a live, single-file HTML frontend to componentized React under integration and end-to-end test coverage, without disrupting 1,200 active users.
- Designed APIs and service boundaries across Go and Node.js backends and maintained them for over a year at **99.99% uptime**; eliminated duplicate processing from client retries with idempotency keys.
- Ran ∼50 client calls a month across onboarding, integrations, and requirements gathering, feeding that into the roadmap.

Renamer AI | Co-Founder & Technical Lead | Remote *Feb 2023 – Jul 2023*

- Built a TypeScript/Electron agent that organizes image libraries from natural-language instruction, months before agent tooling was standard; ∼$10,000 revenue from ∼150 paying designers, IP later sold through Acquire.com.

Minab IT Solutions | Full-Stack Developer & Technical Support Engineer *Feb 2022 – Feb 2023*

- Built the **CareYaya** application end to end on Vue, Nuxt, Hasura GraphQL, and Go; escalation engineer for client projects including **Mastercard** and the **World Bank**.

Products & Open Source

Bestie, AI visual-learning platform | Founder & Sole Builder | 2026 – Present bestie.chrestic.com

- Designed a protocol between the model and a custom browser renderer so lessons are drawn from structured scene descriptions rather than generated frames, at ∼54 seconds and ∼$0.10 per completed lesson.
- Launched in **12 days**: 135 users and 218 lessons in the first five days, 77% returning; 100% lesson completion via automatic recovery from a 6% scene-level retry rate.

Research

The Gatekeeper Knows Enough | Sole author | arXiv:2510.14881 | Oct 2025. Protocol for context-efficient, reliable LLM agents; its open-source reference implementation, Sage (PyPI), reached **73% ± 8% task completion** vs 58% ± 15% for the next-best RAG baseline across three tasks and seven models.

Reduced Bit Median Quantization | Sole author | arXiv:2409.13789 | Sep 2024. Middle-process technique for image compression.

Skills

AI: LLM agents, RAG, hybrid retrieval and reranking, evaluation harnesses, vLLM, quantization, LangChain, OpenAI API, Vertex AI/Gemini, Weaviate, Milvus
Full-Stack: Python, FastAPI, TypeScript, React, Next.js, Node.js, Go, Vue/Nuxt, GraphQL, Spring Boot, PostgreSQL
Production: Docker, GCP, RHEL, GitHub Actions, Prometheus, Grafana, OpenTelemetry, OAuth 2.0/PKCE

Education

BSc, Electrical and Computer Engineering | Addis Ababa Science and Technology University Jun 2025
Top 3% national exit exam | A+ thesis (RBMQ) | 1st place, Ethiopian Digital Health and Stride Ethiopia hackathons