FX REPLAY

Junior AI Engineer

About the Role

FX Replay is building an ambitious AI product, and we’re assembling a small, tight-knit team to climb this mountain together. We’re looking for an early-career **AI Engineer** with **1–2+ years** of hands-on experience integrating **LLM APIs** and a solid foundation in **Python**, plus willingness (or some experience) to contribute on the **front end** (preferably **Angular/TypeScript**).
If you love figuring things out in uncharted territory, moving fast, and owning outcomes, you’ll fit right in.

What You’ll Work On

- Build AI features that integrate with **OpenAI / Gemini / Anthropic** APIs (tool use, chat flows, function calling, etc.).
- Prototype, test, and iterate rapidly on LLM prompts, chains, and evaluation harnesses.
- Implement backend services in **Python**; collaborate closely with front end to ship cohesive user experiences.
- Contribute to the **front end** (Angular/TypeScript) as needed to wire AI features end-to-end.
- Instrument and monitor LLM performance (latency, cost, quality); help optimize tokens and guardrails.
- Work closely with a small product/engineering crew: planning, code reviews, and hands-on problem solving.

Our Tech Stack (you don’t need all of it—come ready to learn)

- **AI/LLM:** OpenAI, Gemini, Anthropic; embeddings, vector search, prompt tooling & eval
- **Backend:** Python (FastAPI/Flask or similar)
- **Frontend:** Angular, TypeScript (willingness to learn is fine)
- **Infra & Tools:** Git/GitHub, CI/CD, AWS (nice to have)

What We’re Looking For

- **1–2+ years** building with **LLMs in production or serious prototypes** (API integration vs. just using a chat UI).
- Strong **Python** fundamentals and ability to ship clean, maintainable code.
- Some front-end experience (**Angular/TypeScript**) **or** a clear willingness to learn it quickly.
- Comfort with ambiguity, rapid iteration, and owning problems end-to-end.

- Collaborative, low-ego, “catch each other’s rope” team mindset.

Nice to Have

- Experience with **browser extensions** (Chrome/Edge) or other client-side integrations.
- Familiarity with **RAG**, embeddings & tokenization basics, and LLM evaluation methods.
- Exposure to Node.js/NestJS, or cloud services (AWS Lambda, S3, CloudFront).

Location & Time Zone

- **Remote, Americas time zones preferred**

What We Offer

- Flexible, remote-friendly work environment.
- A supportive, collaborative team building something genuinely new.
- Meaningful ownership and growth opportunities in a fast-moving product.
- Modern stack and real impact on thousands of users.