Fernando A. Flores

alex@alexetc.me · Texas, USA

github.com/Anodev-sudo · huggingface.co/peering-latent

Summary

Independent iOS dev and ML student. I've shipped two App Store apps. FirstPass, which I use as a platform for much of my ML tinkering, runs a language model I fine-tuned and shrank to a user-tolerable disk size under 100MB. I've followed AI safety since 2018, with more of a technical interest since early 2025. My greatest concerns are power concentration and gradual disempowerment. Now working through the ARENA interpretability curriculum.

Education

Associate's degree in Computer Science · Del Mar College, Texas · 2024

Summer Bridge Program · Frank H. Dotterweich College of Engineering, Texas A&M University–Kingsville · July 2022 · NSF-funded engineering program

AI safety

- **ARENA curriculum** (self-study, in progress): mechanistic interpretability, on a weekly schedule through January 2027.
- **Reading:** LessWrong since about 2018; technical safety research since early 2025, including emergent misalignment, the persona selection model, and Theia Vogel's work on introspection in small models.

AI and ML projects

On-device summarization model (powers FirstPass Briefs) · model on Hugging Face · code on GitHub

- Fine-tuned a 270M-parameter Gemma 3 model to write one-sentence summaries of arXiv papers, distilled from outputs of larger open models (DeepSeek, Kimi).

- Pruned its vocabulary to the tokens arXiv uses and quantized it to 5-bit: 97 MB, a quarter the size of the 374 MB Granite 4.0 350M model it replaced, with no significant difference in faithfulness. Runs in about 0.2s per paper on iPhone.
- Trained entirely on a MacBook in Python; verified in MLX. Code is public on GitHub under the MIT license; weights are on Hugging Face under the Gemma license.

Paper retrieval and ranking (FirstPass backend)

- Embedded about 3M arXiv papers with Snowflake Arctic Embed Small; SPECTER2 for paper-level ranking and a related-paper graph.
- Rank each user's feed on-device from their likes, alongside a citation network.

Guided Skim (FirstPass, on-device)

- Highlights the sentences most worth reading, in place in the paper.
- A static embedder plus a logistic-regression re-ranker: 8.5 MB, no network, no transformer at inference. Beats an abstract-similarity baseline on 52 hand-labeled papers (leave-papers-out cross-validation).

Related passages (Bifolium, on-device)

- Ships MongoDB's LEAF embedding model as a 44 MB Core ML package and indexes passages across a reader's own files. No document text leaves the device.
- Evaluated the ranking on a 418-document Gwern.net snapshot with blinded LLM-judged labels, a frozen held-out split, and preregistered release gates. The holdout showed keyword scoring hurt ranking (nDCG@4 0.92 vs 0.85), so keywords now only widen the candidate pool.

Shipped software

FirstPass, arXiv reader for iPhone and iPad · App Store · shipped June 2026, nine releases through 1.3 in November 2026

- Freemium app with a Plus subscription; largest audiences in the US, China, India, Canada, and Japan.
- Personalized feed, search, topic monitors, author follows, guided skimming, highlights, and on-device Briefs.
- Ask (Apple Private Cloud Compute): fits each prompt to a token budget, sending the whole paper or sections ranked on-device against the question, plus the reader's clips, figures, and passages from cited papers. Answers cite the paragraphs they use.

- Reading guide (Claude API): a streamed card of each paper's main claims, each linked to the passage it comes from, and which parts to read first. Prompt chosen through rounds of harness evals.
- Run its backend myself: a server that builds daily feeds and pushes them to the app.

Bifolium, Markdown reader for iPhone, iPad, and Mac · App Store · shipped September 2026

- Paid app with a native macOS build.

Experience

Independent iOS developer · 2026 to present

- Design, build, and run two App Store apps solo, from model training to App Review.

Delivery driver, DoorDash · current

Skills

- **ML:** Python, MLX, Hugging Face, fine-tuning, distillation, quantization, embeddings and retrieval, TransformerLens, SAELens
- **Apple platforms:** Foundation Models, Private Cloud Compute, App Intents, StoreKit
- **Other:** Claude API, Claude Code and custom agent tooling, Linux server operations