



## Eric Auld

602-469-6718 [ericauld@gmail.com](mailto:ericauld@gmail.com) [ericauld.github.io](https://github.com/ericauld) [linkedin.com/in/eric-auld/](https://www.linkedin.com/in/eric-auld/)

### Education

M.A., Mathematics from UCLA, graduated 2018 (2 courses shy of completing all PhD coursework, no published research)

M.S., Mathematics from New York University, graduated 2014

B.S., Mathematics from Arizona State University, graduated 2012.

### Tech Experience

Together AI (Aug 2024-current)

Systems Research Engineer, GPU Programming

### Professional Interests

Systems programming for ML, hardware-aware algorithms

CUDA C++, Triton, nsys, ncu, PyTorch Profiler. Descending to PTX for reasoning about instruction descriptors and `mbarr ier`'s. Reducing synchronization costs by close reasoning about PTX memory consistency model.

Rapid, robust kernel generation. I'm bullish on CuTeDSL here, its low compile time shortening agent iteration. I've been playing with the right context repos to boost CuTeDSL generation

Weekly audits of my agent-based workflows as capabilities explode. Currently:

- Codex Desktop threads editing multiple local git worktrees -> mutagen auto-sync each worktree to all hardware-compatible dev machines -> mount each worktree in appropriate container -> local commands to run arbitrary commands in a remote, auto-restricting to idle GPUs. So the agent can run any individual experiment anywhere the right kind of GPU is free, and the code stays current
- I'm a fan of Karpathy's paradigm of LLM-maintained markdown per-project wikis and custom Marp slideshows - to streamline my own learning, and to accrete reusable domain context for myself and my teammates

Integrating best of open-source engines/kernels with custom mods (vLLM, SGLang, TensorRT-LLM)

Automated compilation / kernel fusion with torch compile and custom tools, megakernels

Math for hardware-aware algorithms, e.g. digital signal processing's usage in Tri Dao's work

### Media

[Talk](#) on CUTLASS 3.0 for the "CUDA Mode" Discord server, where I served as a moderator.

Several [blog posts](#) featuring my technical writing. Recents: the Central Limit Theorem, martingales

### Miscellaneous

TA for twenty-five classes at UCLA, received effusive reviews from students. <https://ericauld.github.io/2023/04/01/teaching.html>

Film and music nerd - favs Wayne Shorter, My Bloody Valentine, Brian Eno, McCoy Tyner, Sidney Lumet, William Friedkin