Aarya Gadekar

aaryag@berkeley.edu | github.com/gd3kr | twitter.com/gd3kr

Education

University of Wisconsin-Madison — Madison, WI, USA
Bachelor of Science in Computer Science — *Aug. 2023 – Dec. 2024*

University of California, Berkeley — Berkeley, CA, USA
Concurrent Enrollment — *Jan. 2025 – Ongoing*

The International School Bangalore — Bangalore, KA, IN
IB Diploma — *Aug. 2020 – May 2022*

- Physics - Higher Level: 7/7
- Computer Science - Higher Level: 7/7
- Mathematics (Analysis and Approaches) - Higher Level: 6/7

Work Experience

ML Research and Engineering Intern — June 2024 – August 2024
Truffle - Deepshard Inc. — *Los Angeles, CA*

- Led backend development for *Look Mom No Cloud*, a macOS app enabling one-click deployment of 3000+ open source LLMs, managing model quantization and database infrastructure.
- Architected distributed training API with job orchestration across 4 A100 GPUs, supporting hyperparameter sweeps and LoRA fine-tuning for 7B-70B models.
- Built distributed inference pipeline with load balancing and failover protection, enabling parallel model serving across GPU clusters.
- Built automated news pipeline using Twitter API and LLMs for content scraping, processing, and database integration.
- Assisted in conducting large-scale quantization analysis across 3000+ models, measuring performance degradation on benchmarks like MMLU and HumanEval.

Large Language Models and Software Engineering — January 2023 – May 2023
Caesar Labs (Now Julius.ai) — *Remote - Fulltime*

- Optimized and deployed OpenAI's GPT-3 and Meta's LLaMa using Parameter Efficient fine-tuning (PEFT) on user-generated data, achieving greater token-efficiency and reducing operational costs.
- Engineered *prompt chains* into OpenAI's GPT-3, augmenting its functionality to incorporate external knowledge and enable function invocation.
- Developed a full stack React and Node.js-based web interface to integrate LLMs into *Zapier-esque* productivity suites using drag-and-drop components.

Selected Projects

BlenderGPT | *Python, Node.js* — March 2023 – Sept 2023

- Developed and open-sourced BlenderGPT, an extension that generates Blender-compatible Python code from natural language via a Language Model.
- Curated custom dataset and fine-tuned OpenAI's GPT-3 LLM to generate Blender compatible code with high proficiency at a lower cost per token.
- Developed robust backend infrastructure with key provisioning and rate limiting systems; licensed 200 copies and garnered 3.9K+ stars on GitHub.

BlenderGPT.org | *Python, PyTorch, Node.js* — Oct 2023 – *present*

- Implemented novel text-to-3D diffusion model combining techniques from Unique3D and Flow Matching Generative Models to achieve state-of-the-art results.
- Optimized inference through quantization and architecture improvements, reducing generation time from 120+ to 30 seconds while maintaining geometric consistency.
- Achieved $1K Monthly Recurring Revenue within first week of launching paid tier on blendergpt.org.

Superdub.co | *Next.js, Node.js, Python, PyTorch* — April 2023

- Engineered Superdub.co, a web application that allowed musicians to *dub* their vocals in the manner of popular artists using the *so-vits-svc* vocal synthesis framework.
- Curated audio datasets, conducted training, and optimised models for realistic wav-to-wav audio conversion.
- Implemented dynamic queuing and allocation of inference jobs on GPU-enabled servers on AWS and Lambda Cloud, auto-scaled by application usage.
- Forced to shut down after legal compliance notice concerning fair use of AI from Universal Music, Los Angeles.

AutoRegex.xyz | *Node.js, React.js, PostgreSQL, GPT-3* — July 2022 – *present*

- Created AutoRegex, a full stack web application utilizing OpenAI's GPT-3 for seamless English-to-Regex translation.
- Implemented advanced vertical and horizontal server scaling techniques to support a rapid user growth to 150,000 DAUs within one month.
- Incorporated Stripe for revenue streams via a subscription-based model, achieving $200 in Monthly Recurring Revenue (later made completely free).