# Alex Woodka

College Park, MD, 20740
240-671-5628 | alexwoodka@gmail.com | linkedin.com/in/alex-woodka | github.com/awoodka | alexwoodka.com

## EDUCATION

**University of Maryland** College Park, MD
*B.S. Computer Science, Minor in General Business | GPA 3.88* *Expected Dec 2026*
**Relevant Coursework:** Data Structures, Algorithms, Computer Systems, Web Application Development

## PROJECTS

**Timbre** | *Python, JavaScript, FastAPI, React/Next.js, PostgreSQL, Docker* May 2026 – Sep 2026
*timbre.alexwoodka.com*
- Built a full-stack cross-media recommender to test whether how a work feels could drive recommendations, using Gemini to score 500 books, films, shows, anime, manga, and games on 31 emotional dimensions.
- Implemented nearest-neighbor search in PostgreSQL with pgvector, and verified retrieval by checking that synthetic items built from shared components land as neighbors.
- Split each user's ratings into distinct taste modes, and shipped mood search, LLM-written explanations, and a 3D collection map as a Docker Compose stack behind a Cloudflare Tunnel.

**Debrief** | *Python, pytest, Docker, systemd, Claude Code* Sep 2025 – Present
*debrief.alexwoodka.com*
- Built a personal daily AI/ML briefing to stay informed without having to read dozens of sources, turning each day's news, papers, and discussion threads into one short article on what happened and why it matters.
- Designed it as a fixed Python collector plus two AI steps: one structured Claude call triages stories (matching 119 of 120 hand-labeled examples) and a Claude Code skill writes the article, checked so every quote is grounded.
- Simplified an earlier five-subagent design into the current version, which runs nightly on a home server with systemd and locked-down Docker containers and is tested with pytest.

**Local Inference** | *Python, vLLM, llama.cpp, CUDA, Proxmox VE* Sep 2026 – Present
*localinference.alexwoodka.com*
- Serve Qwen3.8 27B on one RTX 3090 (250 W cap) with an open-source vLLM stack using 4-bit quantization and speculative decoding to achieve 146 tok/s for single-user decoding vs. 31 tok/s on llama.cpp.
- Built a reproducible benchmark and eval harness (hash-locked configs, sandboxed AIME 2025, LiveCodeBench, GPQA Diamond, and BFCL runs) that publishes every run with its settings.
- Found the model often ran out of time mid-reasoning on evals, then forked the serving stack to add two opt-in controls that shorten its thinking without slowing decoding, cutting AIME 2025 reasoning tokens about 10% at the same accuracy in a pilot.
- Built the host on a Proxmox server behind a building-managed network with no port forwarding, running the model in unprivileged LXC containers with GPU passthrough and reaching it only through Cloudflare Tunnel and Tailscale.

## EXPERIENCE

**Fellow, Startup Shell** | *UMD student startup incubator, College Park, MD* Jan 2024 – Present
- Work out of the incubator daily alongside student founders from different disciplines, trading feedback on each other's prototypes.
- Won an honorable mention and cash prize at the UMD x Ironsite Hackathon (Feb 2026) for Experience Engine, a construction coaching tool, building error detection and tutorial video generation (TypeScript, Gemini, Claude).

**Head Bartender, Cornerstone Grill & Loft** | *College Park, MD* Aug 2024 – Present
- Responsible for the quality of service at the bar, communicating clearly with guests and staff to keep orders accurate and service friendly and fast.
- Balance full-time hours behind the bar with a full-time CS course load, staying composed through high-stress, high-volume late-night shifts.

## SKILLS

**Languages:** Python, Java, JavaScript, TypeScript, SQL, C, C++, Rust, Bash
**Frameworks & Tools:** FastAPI, REST APIs, React, Next.js, PostgreSQL, Docker, Linux, Git, pytest, systemd, Proxmox, LXC, Cloudflare, Tailscale
**AI / ML:** LLM inference and evaluation, vLLM, llama.cpp, CUDA, PyTorch, scikit-learn, quantization, speculative decoding, pgvector, Gemini API, Claude Code