

Maxwell Grody

Rockville, MD | 301-706-1521 | mgrody1@umbc.edu | github.com/mgrody1 | huggingface.co/StargazerLabs

Machine learning engineer with six years of production data work and a public portfolio of shipped LLM systems: open-weight model releases with ~3k downloads/month, an MLX LoRA post-training toolkit, a hybrid-retrieval RAG system exposed as an MCP server for Claude agents, and a token-level decoding instrument. Expert in Python, SQL, and Spark. Owns systems end to end, from data pipeline to deployed inference service; two of the projects below are live at public URLs.

PROFESSIONAL EXPERIENCE

Data Engineer | Mathematica, Remote

July 2023 – Present

- Lead ETL architect on cross-functional Medicaid analytics teams. Brought onto Utah's healthcare cost benchmark program for healthcare data expertise; that engagement produced a \$560,000 contract renewal.
- Designed the PySpark/Spark SQL pipelines in Databricks that process millions of Medicaid claims. Cut weekly runtime by 75% during a platform migration without missing a delivery.
- Built the Physical and Behavioral Health Integration module from scratch. It identifies co-occurring conditions across millions of records and became a new analytics product delivered to CMS.
- Onboarded new developers onto existing pipelines with live walkthroughs, written guides, and documentation.

Data Scientist Lead | Automated Export Processing, Remote

May 2020 – May 2023

- Reduced manual product labeling by 60% with a deep learning taxonomy classifier for unstructured e-commerce data, which freed capacity to take on new clients.
- Architected the company's analytics platform (Python, SQL, Tableau/Power BI) for customer profiling, market analysis, and forecasting. It became a core product feature.
- Built automated QA pipelines that found systemic data discrepancies, then worked with the U.S. Census Bureau to correct them.

OPEN SOURCE & INDEPENDENT RESEARCH

StargazerLabs: Open-Weight Model Releases | [Hugging Face](https://huggingface.co)

2026

- Published 7 open-weight MLX models (~3k downloads/month) in bf16 and 3/4/8-bit, including Qwen3.8-32B-Jumbo, a 76-layer model, and Mini-Me, a sheared 23B.
- Jumbo restores knowledge lost during continued training by transplanting the most-drifted layer blocks from a sibling checkpoint, chosen by a per-tensor cosine alignment census. It stays compatible with the family's MTP drafter (82–89% acceptance) for 1.6–2.2× speculative decoding.
- Wrote an MLX LoRA post-training toolkit for the same family: knowledge injection with sparse top-k logit distillation against a teacher bank, layer-gated adapters, region-weighted losses, PEFT-format adapter export, and built-in self-tests.
- Every training run measures catastrophic forgetting on an out-of-domain regression set (foreign-language text, contracts, recipes, dialogue). Held-out evaluation gated release, and only the variants that passed were shipped.

Heart of Gold: Serendipity-Native RAG System and MCP Server

2026

- Built a local RAG system exposed as an MCP server (7 tools over stdio) for Claude agents, serving the Stanford Encyclopedia of Philosophy (1,795 entries, 85k chunks) and a 436-episode podcast corpus. Includes background model warm-up, explicit readiness states, latency budgets with labeled fallbacks, and audit logs.
- Hybrid retrieval (dense embeddings, cross-encoder reranking, lexical channel) with a guaranteed relevance floor and one statistically certified non-obvious passage per query, applying beyond-accuracy metrics from recommender systems to RAG.
- Trained a 17k-feature TopK sparse autoencoder over corpus embeddings as an interpretable concept layer. Mined cross-document connections with PMI certificates, using the corpus citation graph as the null model and a pre-registered placebo control.
- LangGraph client (ask, trace, AMA triage) in which the graph topology, rather than the prompt, keeps the synthesizing model from seeing which passage was the certified surplus. Streamlit UI and a public index-only AMA page.

Attractor: Interactive Token-Level Decoding Instrument

2026

- Interactive decoding tool that exposes a model's full next-token distribution at every step and allows intervention at the logit level (suppress the top candidate by its own margin, boost a mid-rank band, select by rank). Every session is seeded and logged as a replayable JSONL score.
- Duet mode runs two models (e.g., base vs. LoRA fine-tune) in lockstep and annotates each candidate with the second model's rank and the Jensen–Shannon divergence between the two distributions, so training-induced differences can be inspected per token.
- FastAPI/WebSockets backend: models loaded once, per-model KV caching with lazy batched catch-up for the second model, streaming token events, single-seat concurrency control and session limits. Single-file HTML front end, deployed through a Cloudflare Tunnel.

Mindscape Explorer & AMA Checker | exploremindscape.preferencespace.com · amamindscape.preferencespace.com

2026

- Shipped a public semantic "has this been asked?" service over 2,581 AMA question–answer pairs (FastAPI, Qwen3 embeddings, ~150 ms/query, no LLM), served from a Mac through a Cloudflare tunnel under launchd. Explorer site on Cloudflare Pages.
- Collected 444 posts (7.4M words) from a WordPress API into a speaker-labeled, timestamped corpus (robots-compliant, cached, incremental re-runs). The parser covers five transcript layouts and records every exception in a manifest rather than dropping it.
- Make-style DAG runner (16 stages, fingerprinted on corpus, code, and parameters) rebuilds only what changed. The site is its final stage.
- NMF topic models at three resolutions (27/50/92) recover the host's hand-assigned tags at AP 0.65 vs. 0.08 chance and 0.55 for a TF-IDF baseline under nested CV. Serendipity recommender evaluated with beyond-accuracy metrics and bootstrap CIs.

TECHNICAL SKILLS

- **ML & AI:** PyTorch, MLX, Hugging Face Transformers, SFT, LoRA/QLoRA, GRPO/RLHF, distillation, quantization, speculative decoding, model surgery, sparse autoencoders, embeddings, sentence-transformers, RAG, vector search, reranking, RecSys, scikit-learn, NLP, RL
- **LLM Systems:** Claude API, tool use / function calling, MCP servers, LangGraph, LLM agents, prompt engineering and evaluation, controlled-experiment and statistical evaluation design
- **Languages:** Python (expert), SQL (expert), Spark (PySpark and Spark SQL), R, Bash, HTML/CSS
- **Platforms & Tools:** Databricks, AWS, Azure, DuckDB, PostgreSQL, Docker, FastAPI, Cloudflare Pages/Tunnel, Airflow/Prefect, DBT, Git, GitHub Actions, Terraform, CI/CD, Linux, Claude Code, Cursor, Codex

EDUCATION

University of Maryland, Baltimore County

December 2022

Master of Professional Studies in Data Science | GPA: 3.8/4.0

Washington University in St. Louis

May 2018

Bachelor of Arts in International Relations, Philosophy, and Chinese