Elliot “Li” Bearden

li@libearden.dev · Portfolio · LinkedIn · Github · Chiang Mai, Thailand (US citizen)

SUMMARY

AI safety research engineer focused on LLM evaluation methodology: specifically how current eval infrastructure misses the failure modes that matter in production deployment. Five years of applied ML at Deepgram, with deep experience in production model evaluation, custom training pipelines, and the gap between benchmark performance and deployment behavior. MSCS capstone on run-level reproducibility of LLM sycophancy benchmarks, defended July 2026; arXiv preprint targeting August 2026.

RESEARCH

Run-Level Reproducibility of LLM Sycophancy Benchmarks **Nov 2025 – Present**

MSCS capstone, Grand Canyon University · Defended July 8, 2026

Public repo (v0.1.0): github.com/LiBearden/sycophancy-benchmark-reliability

- Built a reproducible evaluation harness testing whether single-run confidence intervals understate true measurement uncertainty in LLM sycophancy benchmarks — controlled prompt variants, seed/temperature sweeps, multi-model panel (Claude Sonnet-4.6, GPT-4.1, GPT-4o, Qwen2.5-72B, GPT-3.5-turbo).
- PARROT positive control replicated cleanly at K=5: rank stability holds across runs (Spearman 0.95–0.99), confirming within-run CIs are conservative for this metric class. SycEval reproduction is the pre-registered, still-open falsification test.
- Co-developed a denominator-taxonomy methodology for follow-rate measurement with capstone advisor Dr. Abdulaziz Alharbi (GCU), matching prior published methodology exactly and explicitly labeling divergent conventions rather than conflating them.
- External research funding secured on this work: BlueDot Impact Rapid Grant ($350, compute); Manifund grant recommended, pending final confirmation (project page)
- Target venues: arXiv preprint (Aug 2026), SafeAI@AAAI / SoLaR workshop track (Oct 2026). Project page: https://libearden.dev

EXPERIENCE

Applied Engineer, AI — Deepgram **Jul 2021 – Apr 2026**

Remote · Production speech and language model deployment, evaluation infrastructure, custom training pipelines

- Built and owned the production evaluation infrastructure for custom-trained speech models, including the automated accuracy and engine-load assessment pipeline that cut custom model delivery from 14 days to 1. Supported 150+ on-premise and managed-cloud deployments for enterprise customers with security and compliance requirements (SOC 2, HIPAA, FedRAMP-adjacent).
- Designed and shipped Midas (training-data ingestion/labeling pipeline) and NAVI (retrieval-grounded LLM support agent), both in production use across the customer base.
- Worked with 300+ enterprise customers across $100M+ ARR — direct exposure to the gap between model evaluation metrics and deployment-context performance, which now anchors my research.

Civic Digital Fellow — NIH National Library of Medicine **Jul 2020 – Feb 2021**

Coding it Forward federal civic-tech fellowship

- Designed and implemented Common Data Elements (CDEs) for federal pandemic data infrastructure — direct exposure to how institutional measurement systems get designed, contested, and operationalized inside federal organizations.

EDUCATION

M.S. Computer Science — Grand Canyon University **July 2026**

Capstone: LLM evaluation methodology, sycophancy detection, run-level reproducibility · DS4A Data Engineering Graduate (Correlation One) · Endowed Scholar

B.S. Information Technology — Grand Canyon University **2021**

AI SAFETY FELLOWSHIPS & PROGRAMS

BASE (Black in AI Safety and Ethics) ARENA cohort — attending · ML4Good Technical bootcamp, Hong Kong (Oct 2026) — applied · BlueDot Impact AGI Strategy course — completed · EAGxSingapore 2026 (Aug 29–30) — attending

SKILLS

Research methods: LLM evaluation design · controlled prompt-variant experiments · sycophancy and honesty evaluation · eval validity/reproducibility analysis · measurement infrastructure for behaviors that resist easy measurement (sandbagging, capability concealment, deployment-context drift)

Tools: Python, Rust, SQL · PyTorch, HuggingFace, OpenAI Evals, W&B, vLLM, Ray · Docker, Kubernetes, AWS