Elliot “Li” Bearden

li.bearden@proton.me · libearden.dev · linkedin.com/in/elliot-bearden · github.com/LiBearden · 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. Active capstone research on sycophancy detection and demographic variation in eval validity, targeting arXiv preprint and workshop submission (SoLaR, SafeAI@AAAI) by September 2026.

RESEARCH

LLM Sycophancy Detection & Evaluation Framework Nov 2025 – Present

MSCS capstone, Grand Canyon University · Targeting completion July 2026

- Designing a reproducible evaluation harness to measure sycophancy vs. accuracy across controlled prompt variants — domain and paraphrase variation, seed and temperature sweeps, and benchmark-style aggregation across mitigation strategies.
- Investigating demographic and contextual validity of alignment evaluations — whether eval signals that hold for the median user fail systematically for users outside the training distribution.
- Methodology: controlled prompt-variant design; sycophancy-accuracy tradeoff measurement; mitigation benchmarking. Stack: Python, HuggingFace, OpenAI Evals, W&B, vLLM, Ray.
- Target venues: arXiv preprint (Sept 2026), SafeAI@AAAI, SoLaR, NeurIPS workshop track. Project page: 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, a production data pipeline for ingestion, transcoding, and labeling of customer training audio. Slack and MLflow integration; LLM agent for serving labeling requests.
- Designed and shipped NAVI, a retrieval-grounded LLM agent for tier-1 customer support, indexing live Deepgram documentation and reference material. Testing UI, Slack integration, feedback mechanisms.
- 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 Summer 2020

Coding it Forward federal civic-tech fellowship · Pandemic-era public health data infrastructure

- Designed and implemented Common Data Elements (CDEs) for federal pandemic data infrastructure, contributing to demographic reporting standards across NIH systems — direct exposure to how institutional measurement systems get designed, contested, and operationalized inside federal organizations, which informs my interest in the institutional dynamics of AI safety review.

EDUCATION

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

Capstone: LLM evaluation methodology, sycophancy detection, eval validity · DS4A Data Engineering Graduate (Correlation One) · Endowed Scholar

B.S. Information Technology — Grand Canyon University 2021

SKILLS

Research methods: LLM evaluation design · controlled prompt-variant experiments · sycophancy and honesty evaluation · eval validity 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