HILLARY NANA YAA OWUSU

hnyowusu@cs.umd.edu | linkedin.com/in/hillarynanayaaowusu | College Park, MD

EDUCATION

Ph.D. in Computer Science, University of Maryland, College Park 2023 - Present

B.S. in Telecommunications Engineering, University of Maryland, College Park 2017 - 2021

TECHNICAL SKILLS

Languages: Python, C, R, Bash, LaTeX

ML Frameworks: PyTorch, HuggingFace Transformers, JAX, DeepSpeed

Methods: Causal Intervention, Activation Steering, Mechanistic Interpretability, TransformerLens, NNSight

Tools: Linux, SLURM, Git, Weights and Biases, Docker

RESEARCH EXPERIENCE

Ph.D. Researcher | *CLIP Lab, University of Maryland* 2023 - Present

Advisor: Prof. Naomi Feldman - Quantitative analysis of LLM internals, causal ML methods, large-scale NLP

- Designed empirical studies on behavioral robustness in language models, applying statistical methods (OLS, LME, causal intervention) to characterize how internal representations predict model behavior across multiple model families and training paradigms.
- Applying activation patching and causal tracing to localize computational subgraphs within MLP and Attention layers responsible for specific behaviors; developing mitigation strategies via activation steering and targeted component modification.

PUBLICATIONS & RESEARCH

Under Review: *Anchoring Depends on Confidence and Post-Training in Language Models* - ACL 2026

- Evaluated anchoring susceptibility across 6 models (Llama and Qwen, 3 training paradigms, N = 600) using TVD, pooled OLS, and Linear Mixed-Effects models with HC3 robust standard errors.
- Internal certainty is a significant negative predictor of anchoring susceptibility (beta = -0.052, $p < .001$); factual accuracy is non-significant ($p = .48$), decoupling what a model knows from how easily it is influenced by numerical primes.
- Post-training amplifies the certainty-robustness coupling: instruction-tuned and distilled models show significantly steeper slopes (beta_int = -0.072, $p = .008$; beta_int = -0.054, $p = .021$) vs. base models ($p = .54$). TVD-anchoring correlation $r = 0.63$, 82.5% directional consistency.

In Preparation: *Same Bug, Different Wiring: How Training Paradigm Shapes Anchoring Bias in LLMs* - COLM 2026

- Using causal intervention (activation patching, causal tracing) to localize anchoring circuits within transformer layers, with comparative circuit-level analysis across base, instruct, and distilled models. Evaluating activation steering as a mitigation strategy.

Preprint: *Bias-Aware AI Chatbot for Engineering Advising at the University of Maryland A. James Clark School of Engineering* - arXiv:2510.09636 (2025)

- Supervised a team of 4 undergraduate researchers as corresponding author; guided project scoping, experimental design, and evaluation methodology from inception to publication.
- Project developed a RAG-based academic advising chatbot with integrated bias detection, achieving mean accuracy of 9.76/10, relevance 9.56/10, and 0% stereotypical bias rate across 75 evaluated prompts.

SELECTED PROJECTS

GSM++ - LLM Reasoning Robustness Benchmark Framework

- Built an evaluation framework extending the 8.5K-problem GSM8K benchmark with 5 semantically-equivalent variant test sets (paraphrased, redacted, thematic, multilingual) to stress-test reasoning robustness in LLaMA-2-7B and LLaMA-3.1-8B.
- LLaMA-3.1-8B accuracy dropped from 71% to 55.5% under Hindi translation with up to 26.5% of correct answers flipping incorrect; fewer than 12.3% of LLaMA-2-7B responses remained consistently correct across paraphrased variants despite BLEU-confirmed semantic preservation.

Membership Inference Attack Analysis - GPT-2 Shadow Model

- Fine-tuned GPT-2 (117M params) on 501,197 Common Crawl data points over 9 epochs as a GPT-3 shadow model; built a TF-IDF attacker (156,474 test samples) achieving 97% accuracy (F1: 0.97) at distinguishing training members from non-members.
- ~80% of 10,000 queried prompts classified as training members, quantifying memorization as the primary attack surface; proposed differential privacy and adversarial training as mitigations.

Gender Bias in English to Twi Neural Machine Translation

- Constructed 3,080 minimal pairs across 7 templates and 30 occupations to audit gender bias in English-Twi NMT; 7.3% of pairs (226/3,080) yielded structurally divergent translations, with feminine outputs systematically longer and reattributing physical descriptors vs. character traits.
- BLEU degradation confirmed quality loss in divergent pairs (0.0764 vs. 0.0810); unique token analysis revealed systematic gendered semantic drift across adjective and occupation categories.

ACADEMIC SERVICE & TEACHING

Reviewer: ACL 2026 | NeurIPS Reliable ML Workshop (2025)

Teaching Assistant: Algorithms (CMSC351), Data Science (CMSC320), C Programming (CMSC106)

AI Research Mentor: UMD Esteem-SerQUEST Summer Research Program - guided incoming engineering students through independent AI research projects.