Bethel Hall

AI Researcher | AI Verification, LLM Evaluation & Formal Methods

Hoboken, NJ | +1 908-880-4895 | bhall2@stevens.edu
bethelhall.tech | github.com/bethelhall | linkedin.com/in/bethelhall | Google Scholar

SUMMARY

Ph.D. researcher building trustworthy AI systems through LLM evaluation, formal verification, and neurosymbolic reasoning. Designs evaluation benchmarks and solver-in-the-loop LLM pipelines across finance, healthcare, and cloud security, with prior experience in safety-constrained model training.

EDUCATION

Ph.D. in Computer Science, GPA: 3.93/4.00 — Expected 2028
Stevens Institute of Technology; Advisor: Prof. William Eiers — *Hoboken, NJ*
B.S. in Software Engineering, Addis Ababa University — 2017–2022

RESEARCH EXPERIENCE

Graduate Research Assistant — Sep 2022–Present
Stevens Institute of Technology, Department of Computer Science — *Hoboken, NJ*

- **VeriFin – LLM evaluation and benchmark design:** Led the design and implementation of a solver-backed data and evaluation pipeline that grounds operands in SEC XBRL filings, derives reference calculations independently of the answer model, and verifies LLM-generated financial claims with Z3. Built a **667-question evaluation benchmark** and harness spanning XBRLFiling and FinanceBench; benchmarked **six LLMs** with **zero false accepts**, reaching **100% precision among accepted claims at 98.8% coverage** on a fixed Claude Haiku evaluation set.
- **VERIMED – AI reasoning and tool-augmented formalization:** Built a multi-stage LLM-to-SMT workflow for evaluating natural-language software requirements, cross-checking independent model formalizations with a solver and returning counterexamples to the model for repair on disagreement. Formalized 64/64 hemodialysis requirements, uncovered **12 latent ambiguities**, and improved verified QA accuracy from **55.4% to 98.5%**.
- **CloudFix – verified program repair with solver-in-the-loop LLMs:** Led development and evaluation of a reusable repair pipeline combining SMT fault localization, targeted LLM edits, and formal re-verification for AWS IAM policies. Open-sourced the implementation and released a **282-policy evaluation benchmark**; achieved **84.0% repair accuracy versus 48.2%** for an LLM-only baseline on 10-request tasks.
- **PolicySummarizer – semantic evaluation at scale:** Co-developed and open-sourced a semantic validation pipeline for LLM-generated access-control summaries using finite-state automata and model counting. Evaluated **746 cloud policies** from AWS, Azure, and GCP, achieving **0.93 mean semantic similarity** and improving difficult policy-review accuracy from **39% to 93%** in a user study.

SELECTED PUBLICATIONS

- **Hall, B.**, Ungaro, O., and Eiers, W. "CloudFix: Automated Policy Repair for Cloud Access Control Policies Using Large Language Models." *Research Track, SANER 2026.*
- Vatsa, A., **Hall, B.**, and Eiers, W. "Neurosymbolic Characterization for Reliable Access Control Policy Analysis." *Accepted at ISSRE 2026.*
- **Hall, B.**, Shome, S., and Eiers, W. "VeriFin: A Neurosymbolic Framework for Verifying LLM-Generated Financial Claims." Manuscript, 2026.
- **Hall, B.** and Eiers, W. "Neurosymbolic Auditing of Natural-Language Software Requirements." arXiv:2605.13817, 2026.

MODEL TRAINING, ADDITIONAL RESEARCH & TEACHING

- **University of Michigan, Undergraduate Research Assistant (2021) – safety-constrained model training:** Integrated formal safety constraints into the neural-network training loop in a simulated aviation environment, implementing a safety-augmented loss to enforce safety-compliant behavior.
- **Tennessee Tech University, Undergraduate Research Assistant (2021):** Analyzed bias in trauma-triage models and trained GANs for synthetic-data augmentation.
- **Teaching and research communication:** Graduate Teaching Assistant for Machine Learning and Java Programming, Stevens Institute of Technology (2025–Present); presented research at AIMACS/CAV 2026.

TECHNICAL SKILLS

AI and machine learning: LLM evaluation, model evaluation, evaluation benchmark design, agentic and tool-augmented LLM pipelines, model training, safety-constrained training, autoformalization, PyTorch, TensorFlow, scikit-learn, GANs, retrieval, zero/few-shot learning
Formal methods: Z3, SMT solving, formal verification, program analysis and repair, constraint solving, finite-state automata, model counting, UNSAT cores, fault localization
Research engineering and data: Python, Java, Git, Linux, data pipelines, XBRL, AWS IAM, Azure RBAC, GCP IAM, LaTeX
Honors: Global UGRAD Scholar, U.S. Department of State (2021) | Palantir Global Impact Scholar (2020)