Bowen Wei | Resume

Fairfax, VA · +1 (434) 254-9053 · bwei2@gmu.edu

Website: https://weibowen555.github.io/
LinkedIn: https://www.linkedin.com/in/bowen-wei-9485a1192/

Research Interest

My research spans **trustworthy and interpretable AI** for large language models. I develop prototype-based, symbolic, and explanation-driven methods to make model behavior transparent, fair, and robust, enabling users to understand and trust AI decisions in high-stakes settings. In parallel, I study **RL** and **post-training** techniques that distill multi-agent reasoning into single, verifiable agents—improving reasoning quality, evidence attribution, and causal grounding. Together, these directions aim to advance AI systems that are both **interpretable in their inner logic** and **competent in autonomous, evidence-based reasoning**.

Education

- Ph.D. in Computer Science, George Mason University, Fairfax, VA (Expected 2028)
- M.S. in Computer Science, University of Virginia (2021–2023)
- B.S. in Computer Science, Xidian University (2016–2021)

Publications

- **AAAI 2026** Oral **Bowen Wei**, Ziwei Zhu.
Making Sense of LLM Decisions: A Prototype-based Framework for Explainable Classification. Acceptance: 4,176 / 23,680 ≈ 17.6%.

- **ACL 2025** Main **Bowen Wei**, Ziwei Zhu.
ProtoLens: Advancing Prototype Learning for Fine-Grained Interpretability in Text Classification. Acceptance: 1,699 / 8,360 ≈ 20.3%.

- **NeurIPS 2025 LAW Workshop** **Bowen Wei**, Yuan Shen Tay, Howard Liu, Jinhao Pan, Kun Luo, Ziwei Zhu, Chris Jordan.
CORTEX: Collaborative LLM Agents for High-Stakes Alert Triage.

- **WACV 2026** Mehrdad Fazli, **Bowen Wei**, Ziwei Zhu.
CAAC: Confidence-Aware Attention Calibration to Reduce Hallucinations in Large Vision-Language Models.

- **ICLR 2026 (under review)** **Bowen Wei**, Ziwei Zhu.
Neural Symbolic Logical Rule Learner for Interpretable Learning.

- **In submission** Chahat Raj, **Bowen Wei**, Ziwei Zhu.
VIGNETTE: Socially Grounded Bias Evaluation for Vision-Language Models.

- **M.Sc. Thesis** **Bowen Wei**, Yiling Jia, Hongning Wang.
An Empirical Study of Neural Contextual Bandit Algorithms.

Current Projects

Evidence-Attribution Reinforcement Learning (EA-RL) - Target: ICML 2026 Oct 2025 – Present

- Leading a project to design multi-agent LLM distillation with explicit **evidence attribution and reliance**
- Proposing a new paradigm where models are rewarded not just for being correct, but for **using and depending on the right evidence**
- Building on the CoA framework by introducing **evidence-aware post-training**
- Aiming to develop **faithful, verifiable single-agent reasoning systems** that can explain both *what* and *why*
- Designed to bridge outcome accuracy and causal faithfulness, establishing a new standard for trustworthy reasoning in LLMs

NeuroSymbolic Autoencoder for Interpretable Recommendation - Target: SIGIR 2026 Oct 2025 – Present

- Developing a **NeuroSymbolic Autoencoder** that integrates neural representation learning with symbolic reasoning for transparent RecSys
- Employing the **Rule Network** as both encoder and decoder to learn **logical rule-based latent spaces**
- Aiming to create **interpretable, rule-grounded recommendation systems** that unify neuro-symbolic learning and explainable personalization

Internship

AI Agents Developer - Fluency Security Jun 2025 – Aug 2025

- Proposed and built **CORTEX**, a role-specialized multi-agent LLM architecture for SOC alert triage
- Work accepted to the NeurIPS 2025 LAW Workshop as ***CORTEX: Collaborative LLM Agents for High-Stakes Alert Triage***

GenAI Engineer - GoEngage Jun 2025 – Aug 2025

- Implemented a semantic search engine that improved retrieval accuracy over keyword matching
- Developed an agentic chatbot that autonomously queries backend APIs and generates analytical reports for non-technical users

Professional Service

- Reviewer / Sub-reviewer: ARR (Dec 2024; Feb 2025—ACL; May 2025—EMNLP), KDD 2024, ACML 2024–2025, SSCI 2025, CAIS.