

ALVIN ZHANG

Nine East 33rd Street, Baltimore, Maryland, 21218
(443) 722 7886, bzhang90@jhu.edu

Education

Johns Hopkins University, Baltimore, MD

August 2023 –Present

Junior Major: Applied Mathematics and Statistics

Minor: Computer Science

- **GPA:** 4.0/4.0

- **Selected Awards and Honors:** Dean's List: Every Semester Until Now; CRA Outstanding Undergraduate Research Awards Honorable Mention ([CRA Outstanding Undergraduate Researchers - CRA Education](#))

Course Works: *Calculus III; Linear Algebra; Differential Equations and Applications; Probability; Honor Mathematical Statistics; Bayesian Statistics; Introduction to Optimization; Time Series Analysis; Gateway Computing (Java); Data Structures; Introduction to Data Science; Machine Learning: Deep Learning; Natural Language Processing; Natural Language Processing: Self-Supervised Models; AI Agents; Honor Discrete Mathematics; Research and Design in Applied Mathematics; Reading the Mind: Computational Cognitive Neuroscience of Vision; AI and Democracy*

Publications

Dongwei Jiang*, **Alvin Zhang***, · · · & Daniel Khashabi, *Feedback Friction: LLMs Struggle to Fully Incorporate External Feedback*, Advances in Neural Information Processing Systems (NeurIPS), 2025. (<https://arxiv.org/abs/2506.11930>)

Zhaoyi Joey Hou, **Alvin Zhang**, · · · & Daniel Khashabi, Xiang Lorraine Li, *CreativityPrism: A Holistic Benchmark for Machine Creativity*, Submitting for TMLR (<https://arxiv.org/pdf/2510.20091>)

Anushri Suresh*, **Alvin Zhang***, · · · & Daniel Khashabi, *Compute When Worth It: Risk Control for Reasoning on a Compute Budget*, Advances in International Conference on Machine Learning (ICML), 2026; NeurIPS 2025 Workshop on Efficient Reasoning ([\[2602.03814\] Conformal Thinking: Risk Control for Reasoning on a Compute Budget](#))

Alvin Zhang*, Arda Uzunoglu*, Daniel Khashabi, *Trust Functions: Near Lossless Weak-to-Strong Generalization by Learning to Trust the Weak Teacher*, Advances in International Conference on Machine Learning (ICML), 2026; arxiv coming soon.

Zheyuan Zhang, Zehao Wen, **Alvin Zhang**, Andrew Wang, Jianwen Xie, Daniel Khashabi, Tianmin Shu, *AgentOdyssey: Open-Ended Text Game Generation for Test-Time Continual Learning Agents*, Submitting to NeurIPS2026 (<https://agentodyssey.github.io/>)

* indicates equal contribution

Research Interests

Natural Language Processing; Machine Learning; Deep Learning; Improving LLM reasoning abilities, particularly in the domains of self-correction; LLM confidence calibration; LLMs as judges; LLMs creativity and systematic ways of evaluating LLM creativity; Improving LLM's mathematical reasoning ability; Inference

ALVIN ZHANG

Nine East 33rd Street, Baltimore, Maryland, 21218
(443) 722 7886, bzhang90@jhu.edu

time scaling; Efficient Reasoning; Agents.

Research Experience

Center for Language and Speech Processing, Intelligence Amplification Lab, Johns Hopkins University, Baltimore, MD

Undergraduate Research, Supervisor: Professor Daniel Khashabi

August 2024 – Present

Project 1: Investigating LLM’s Self-Correction Ability, Trying to Improve with Inference-Time Scaling Methods and RL Training

Investigated LLMs’ limitations in incorporating near-perfect feedback and their tendency toward stubbornness in multi-turn settings.

Developed inference-time scaling methods to improve feedback incorporation and self-correction behavior across dialogue turns.

Co-First-authored paper accepted to **NeurIPS 2025**, currently extending the work with **multi-turn reinforcement learning (RL)** to enhance small models’ ability to learn from external feedback.

Project 2: Risk-Controlled Early Stopping for Efficient Reasoning Models

Designed a **risk-control pipeline** that leverages confidence, entropy, and hidden-state probes to trigger early stopping during inference.

Achieved significant token savings for large reasoning models while maintaining solution quality.

Co-First-Authored Work **accepted at NeurIPS 2025 Workshop on Efficient Reasoning**; currently refining the pipeline for fine-grained signal extraction and preparing an arXiv submission. In submitting to ICML 2026.

Project 3: Evaluating LLM Creativity

Benchmarked large models’ creativity and accuracy trade-offs in structured reasoning domains.

Designed and evaluated inference-time scaling methods to boost creative reasoning while maintaining correctness.

Paper **under submission to TMLR** focusing on quantifying creativity in diverse tasks.

Project 4: Confidence Probing and Hidden-State Steering in LLMs

Developing lightweight **probes and polytopes** to extract per-token correctness signals from hidden states.

Applying **hidden-state steering** when the probe detects low-confidence reasoning trajectories.

Work in progress toward building general purpose steering vectors to steer a base model behaving like a reasoning model, testing if steering vectors are transferable across model families, currently in progress.

Project 5: LLMs Weakness Under Question Perturbation

Designing automatic frameworks in perturbing reasoning questions with verifiable answers.

Evaluating weakness of LLMs under perturbed questions.

Trying Perturber-Solver RL training to improve LLMs robustness, currently in progress.

Project 6: Weak-to-Strong Generalization in LLM Training

Studying how weak teachers and weakly supervised data can bootstrap strong student models across different

ALVIN ZHANG

Nine East 33rd Street, Baltimore, Maryland, 21218
(443) 722 7886, bzhang90@jhu.edu

domains, from MCQ to Mathematical Reasoning.

Exploring efficient transfer mechanisms and generalization bounds for weak-to-strong adaptation, currently in progress. In submitting to ICML 2026.

Project 7: Self-Correction Abilities of Small-Scale Vision-Language Models (VLMs)

Explored the self-correction capacity of VLMs when given feedback on visual reasoning errors.

Implemented and fine-tuned VLMs using targeted feedback to assess adaptability.

Presented results as part of a **class research project** highlighting finetuning strategies for feedback incorporation.

Project 8: Evaluating LLM Agents in Open Text-Based World

Examine how Agents perform in open-ended text-based world on memorization and reasoning.

Implemented model training and helping building the evaluation framework.

In submitting to ICML 2026. Will be arxiv soon

Internship Experience

Software Develop Engineer Intern

June 2024 – July 2024

Caterpillar.inc, Simulation Intern, WuXi, JiangSu Province, China

- Developed Recurrent Neural Networks in simulating the current and voltage in electrical motors.
- Explored possible ways to replace traditional simulation methods with neural networks.

Teaching Experience

Undergraduate Teaching Assistant:

Lecturer: Donniell Fishkind **Class:** EN553.361 Introduction to Optimization (August 2024 – December 2024)

- Teaching TA sessions for class recap and solve practice problems with students; Grading homework and exams.

Lecturer: Jason Esiner **Class:** EN 601.465 Natural Language Processing August 2025 – December 2025

- Course Assistant for homework grading and office hours.

Skills

- Language: Chinese, English.
- Programming: Python, Java, Matlab, R, Wolfram Mathematica.
- ML/NLP Frameworks: PyTorch, Transformers, VeRL, vLLM, SGLang.
- Systems: Windows, Linux, Experience with hypercomputing clusters remotely.
- Machine Learning, Deep Learning, Reinforcement Learning.