Jehyeok Yeon

(619) - 513 - 1181 · jehyeok2@illinois.edu · https://www.linkedin.com/in/jehyeoky

Education

University of Illinois Urbana-Champaign August 2022 – December 2025

Bachelor of Sciences in Computer Science + Linguistics. GPA 3.92.

- President, Codable Student Organization: Led a team of over 40 students to organize coding workshops and events along with running semester-long projects, fostering student engagement across different skill levels and backgrounds.
- Relevant Courses: Machine Learning, Data Mining, Database Systems, Robot Manipulation, Corpus Linguistics, Computational Linguistics, Algorithms, Data Structures, Compilers, Discrete Structures

Research Interests

- Trustworthy AI, Explainable AI, Formal Methods, Generative AI, Multimodal Models, Agentic AI

Publications

- Hangoo Kang*, **Jehyeok Yeon***, Gagandeep Singh. *TRAP: Targeted Redirecting of Agentic Preferences.* NeurIPS 2025. (*indicates equal contribution)
- **Jehyeok Yeon**, Isha Chaudhary, Gagandeep Singh. *Certifying Robustness of Agent Tool-Selection Under Adversarial Attacks.* (Under Review)
- **Jehyeok Yeon**, Yifan Wu, Federico Cinus, Luca Luceri. *GSAE: Graph-Regularized Sparse Autoencoders for Robust LLM Safety Steering.* (Under Review)
- **Jehyeok Yeon**, Lawrence Angrave. *The Power of Friendship: Analyzing Leadership and Adversarial Attacks in Multi-Agent Collaboration.* ACM Collective Intelligence 2025 Poster Acceptance.

Ongoing Projects

- **Jehyeok Yeon**, Isha Chaudhary, Gagandeep Singh. *When Context Breaks Representation: Understanding Layer–Input Failures in LLMs (Working Title).* To be submitted for publication in 2026.
- **Jehyeok Yeon**, Federico Cinus, Luca Luceri. *Temporal Tomography of Conceptual Learning: An SAE-based Analysis of LLM Pre-training (Working Title).* To be submitted for publication in 2026.
- **Jehyeok Yeon**, Hyeonjeong Ha. *All Roads Flow to Rome: Securing Multimodal Models Against Cross-Modality Attacks (Working Title).* To be submitted for publication in 2026.

Research Experience

Max Planck Institute for Intelligent Systems January 2026 – August 2026

Visiting Researcher

- Will be conducting research at the AI Safety and Alignment group under Maksym Andriushchenko about understanding and improving safety of computer use AI agents.

FOCAL Lab@UIUC November 2024 – Current

Research Assistant

- Achieved 100% ASR on SoTA vision-language models via novel embedding-level semantic injection and diffusion decoding (first-author, advised by Prof. Gagandeep Singh).
- Developed the first statistical certification framework for agentic AI tool selection under adversarial scenarios via LLM-based adversarial distributions (first-author, advised by Prof. Gagandeep Singh).

ISI@USC February 2025 – Current

Research Assistant

- Designed a graph-based analysis method to uncover features related to LLM refusal behavior using graph Laplacian regularization on sparse autoencoders (first-author, advised by Prof. Luca Luceri)
- Built a steering mechanism that uses a dual gating system with hysteresis to enable control over safety behaviors while maintaining strong utility, performing 40% better than previous safety steering methods.

Industry Experience

Intradiem May 2025 – August 2025

Machine Learning Intern

- Trained and tuned a feedforward neural network to predict agent burnout and attrition; optimized for imbalanced classes, achieving 17% F1 lift and robust generalization to unseen shifts.
- Deployed an agent burnout and attrition prediction model as Spring Boot API; served 12k QPS at <180 ms latency, automating retraining and drift detection using Temporal and Mlflow for 99.95% uptime.

Hanwha Life Insurance June 2024 – August 2024

AI Data Scientist

- Developed a hybrid semantic-lexical retrieval architecture for complex tabular data using ElasticSearch and ChromaDB, improving recall on legal document QA datasets by 427%.
- Built a retrieval augmented generation chatbot by integrating a cross-encoder reranking, multi-step query decomposition, and Hypothetical Document Embeddings, raising MRR by 38%.

ATLAS May 2023 – August 2024

Machine Learning Intern

- Developed a high-fidelity forecasting system using real-time Chicago city sensor data, fusing multi-modal IoT streams and high-res imagery via LSTM-TCN hybrids in PyTorch; achieved a 27% reduction in RMSE and 0.91 F1 on severe event prediction.
- Engineered cross-modal attention for feature alignment and validated robustness with adversarial stress tests, maintaining 93% accuracy under noisy sensor dropout and outperforming baseline uni-modal models by 18%.

Honors and Awards

- Get Experience Scholarship (2023)