Philadelphia, PA
cek99@drexel.edu
cekayan.com
in *cekayan*

Ceyhun Efe Kayan

Summary

PhD researcher in mechanistic interpretability of large language models, investigating how LLMs internally represent and execute reasoning, planning, and formalization. Employs activation steering, representation geometry, and circuit-level analysis to identify the mechanisms underlying these behaviors and to intervene on them. Industry research experience at IBM on applications of LLM hidden states. Builds interpretability pipelines on open-weight models using PyTorch, Hugging Face Transformers, and SAE Lens.

Education

2025 - 2028 **Ph.D. in Computer Science**
Drexel University, Philadelphia, Pennsylvania
Focus: Mechanistic Interpretability of LLM planner/formalizers.

2022 - 2025 **Masters in Machine Learning Engineering**
Drexel University, Philadelphia, Pennsylvania
Relevant Coursework: Statistics and Probability, Applied NLP, Reinforcement Learning, Convex Optimization

2018 - 2022 **Bachelors in Electrical and Electronics Engineering**
Izmir Institute of Technology, Izmir, Turkey
GPA: 3.76 | **Department Valedictorian**

Experience

Jun 2026 - Sep 2026 **Research Scientist Intern**, IBM, Silicon Valley Lab, San Jose, CA
- Built a router for a chat-completion endpoint driven by LLM activations that outperformed different large general-purpose models in accuracy (4.9-7.1%) across a broad multi-benchmark suite while requiring fewer FLOPs. **[Paper submitted]**

Sep 2025 - Present **Graduate Researcher**, HAZ Lab, Drexel University
- Applying and constructing activation steering techniques to probe and improve how LLMs reason, plan, and formalize in constraint-satisfaction and MDP-style tasks; developed a prototype-based dynamic steering method for reasoning [1]. **[EMNLP 2026, Main Conference]**
- Co-first author of a systematic evaluation of LLMs as formalizers versus end-to-end solvers on real-life constraint satisfaction problems, spanning 4 domains, 6 LLMs, and 2 types of formal languages [2]. **[COLM 2026 Efficient Reasoning Workshop, Spotlight]**

Jun 2021 - Jul 2022 **Undergraduate Researcher**, MIRALAB, Izmir Institute of Technology
- Integrated pre-trained CNNs with LSTMs for event classification in Φ-OTDR fiber-optic sensing systems.[3]
- Developed an adaptive resizer-based transfer learning framework to improve breast cancer detection from histopathology images.
- Combined CNNs and autoencoders for diagnosis of COVID-19 and viral pneumonia from chest X-rays.

Selected Publications

[1] Ceyhun Efe Kayan and Li Zhang. "Prototype-Based Dynamic Steering for Large Language Models". *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. To appear. 2026. URL: https://arxiv.org/abs/2510.05498.

[2] Rikhil Amonkar*, Ceyhun Efe Kayan*, Qimei Lai, Ronan Le Bras, and Li Zhang. "A Reality Check of Language Models as Formalizers on Constraint Satisfaction Problems". *COLM 2026 Workshop on Efficient Reasoning*. Spotlight; non-archival. *Equal contribution. 2026. URL: https://arxiv.org/abs/2505.13252.

[3] **Ceyhun Efe Kayan**, Kivilcim Yuksel Aldogan, and Abdurrahman Gumus. "Intensity and phase stacked analysis of a Φ-OTDR system using deep transfer learning and recurrent neural networks". *Appl. Opt.* 62.7 (Mar. 2023), pp. 1753–1764. DOI: 10.1364/AO.481757. URL: https://opg.optica.org/ao/abstract.cfm?URI=ao-62-7-1753.

Skills

Interpretability: Steering, Linear Probing, Representation Geometry, SAEs, Transformer Circuits
ML Tools: PyTorch, Hugging Face Transformers, TransformerLens, SAE Lens, vLLM, nnsight