Activation-Level Preference Unlearning for Robust LLM Recommenders

Final Project Presentation

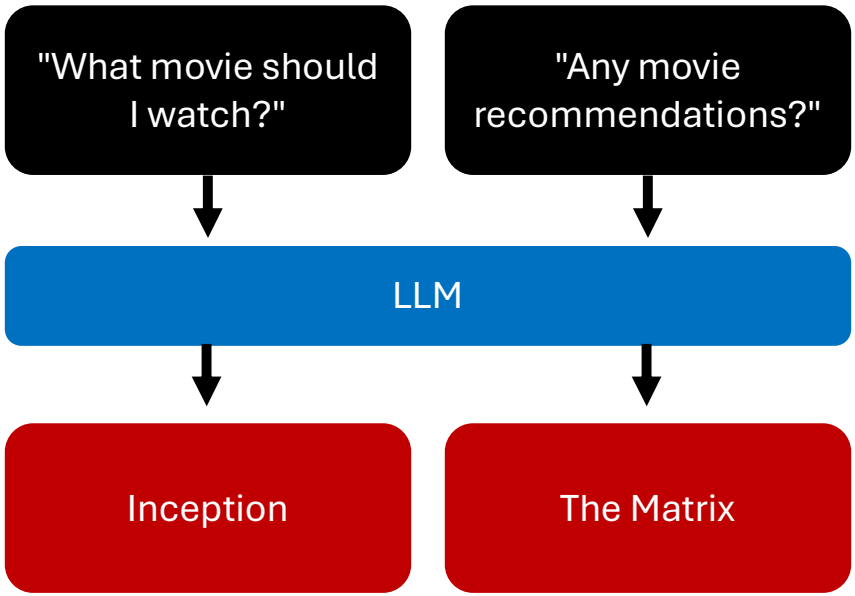
ECE6514 Trustworthy Machine Learning | Fall 2025

Team Members: Mack Werner, Gary Ruppert, Chuan Liu, Jacob Ramey*

*Equal Contributors

Motivation

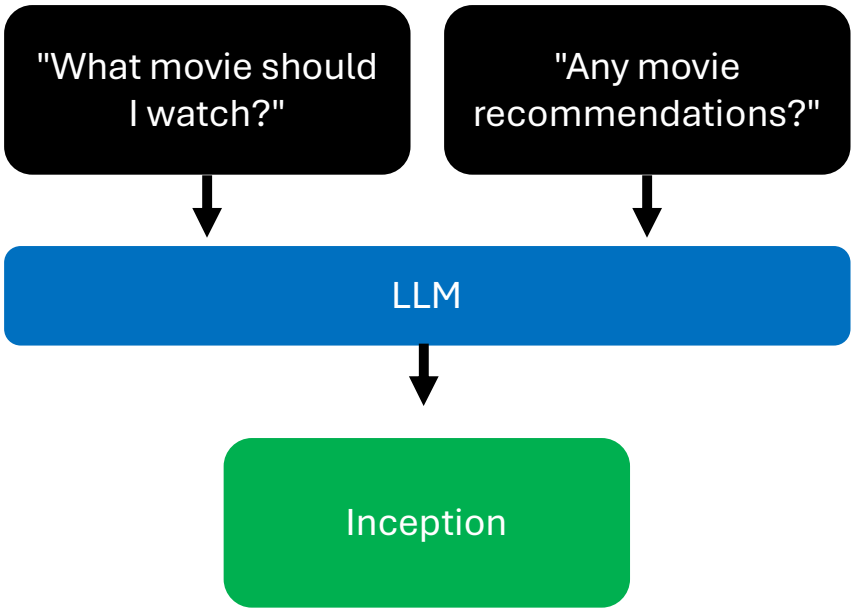
- LLM recommenders often give different answers for prompts that mean the same thing
- Small prompt changes → large output shifts



Similar prompt, different results

Motivation *Cont.*

Goal: Make recommendations *trustworthy* and *consistent*.

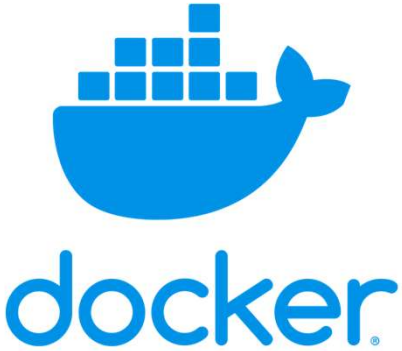


Method Overview

1. Prompt-perturbation experiments
2. Activation extraction & saliency maps
3. Identify sensitive neurons/heads
4. Apply selective dampening or gradient projection
5. Re-evaluate drift reduction

ARC Environment Setup / Github

- Tinkercliffs : A100_normal_q partition
- 8 CPUs / 1 GPU / 64GB Memory
- Docker image running via apptainer on ARC rameyjm7/llm-preference-unlearning | Docker Hub
- Github Repo: rameyjm7/llm-preference-unlearning
- Github Repo 2: https://github.com/Thekicker35/Trustworthy-Machine-Learning-Final-Project/tree/main



Prompt-perturbation experiments

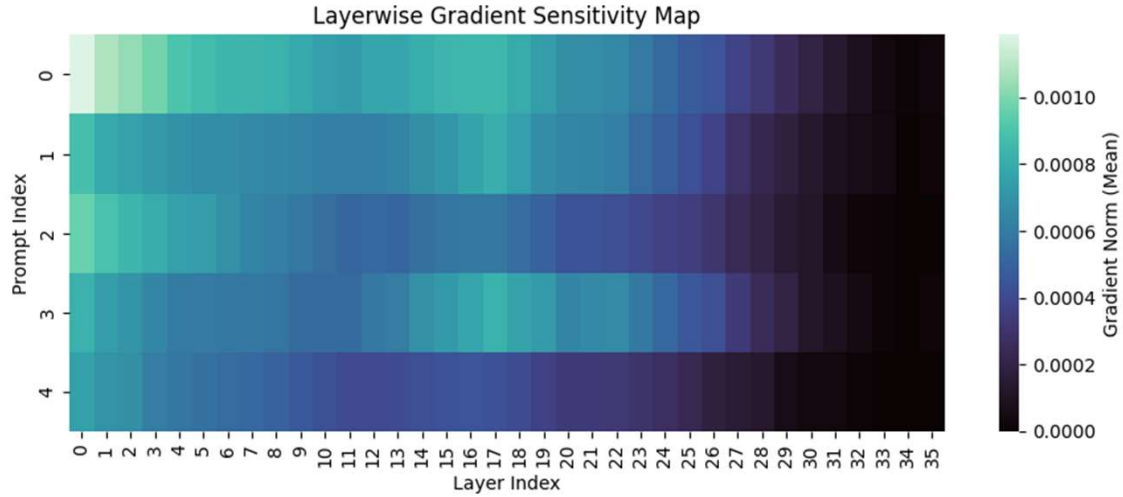
Lets ask the same question in different ways

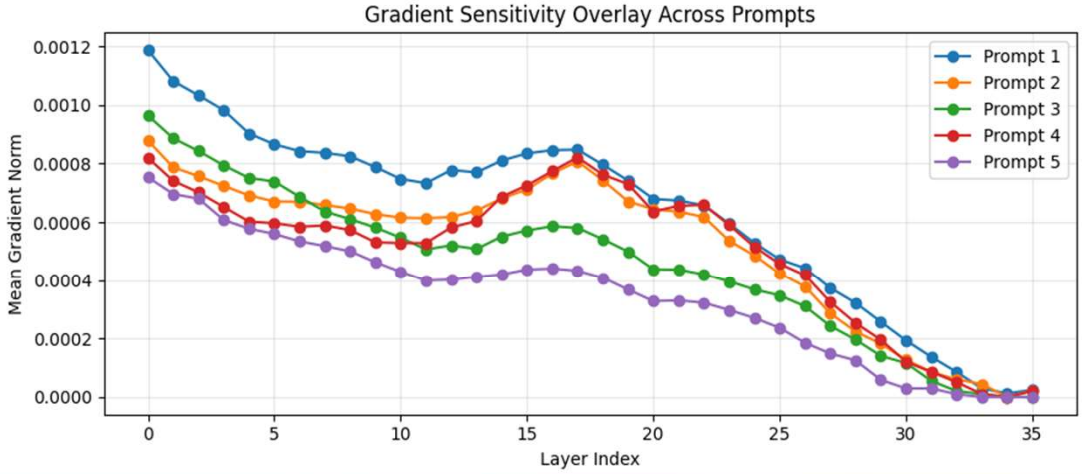
id	prompt
1	Tell me the most informative movie in the 2020–2025 range.
2	Which movie between 2020 and 2025 gives the most valuable real-world insights?
3	From 2020–2025, which film is the most educational or thought-provoking?
4	List a movie released between 2020 and 2025 that best informs audiences about real issues.
5	What's the most knowledge-rich or awareness-raising movie made from 2020 to 2025?

Activation extraction & saliency maps

Where is the model most sensitive to our inputs?

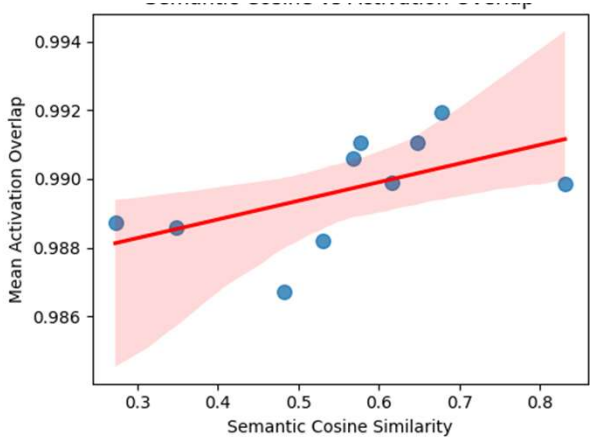
Identify sensitive neurons

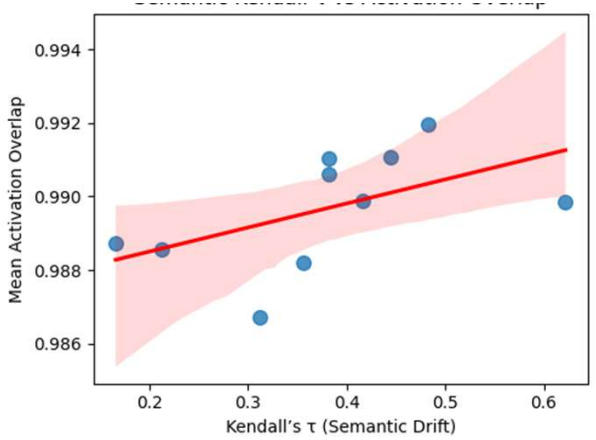


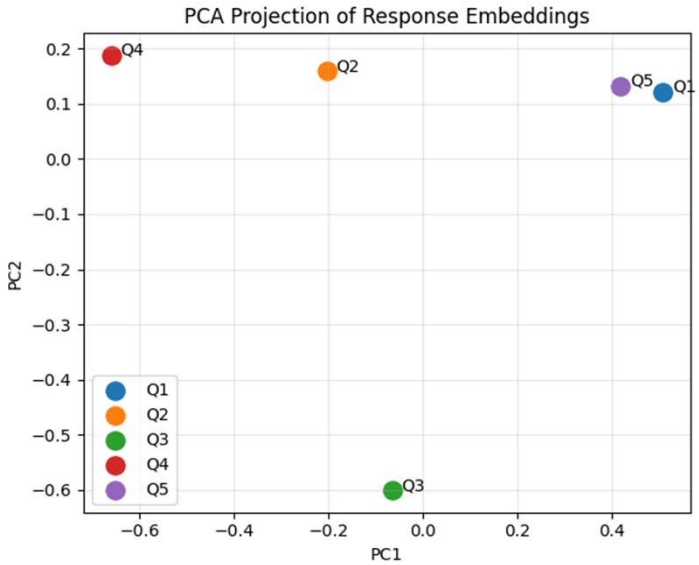


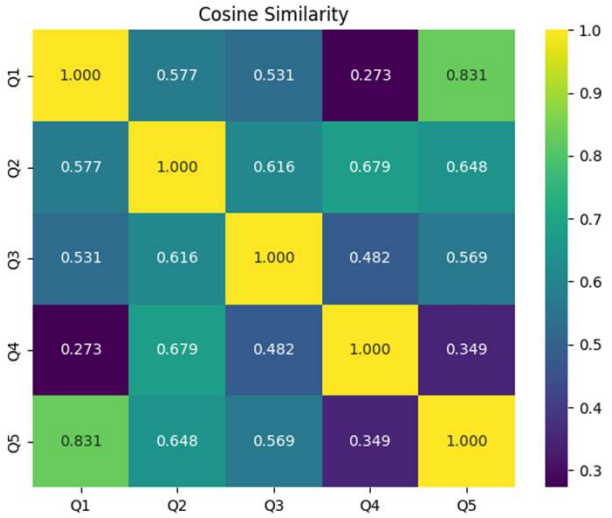
Other methods for similarity…

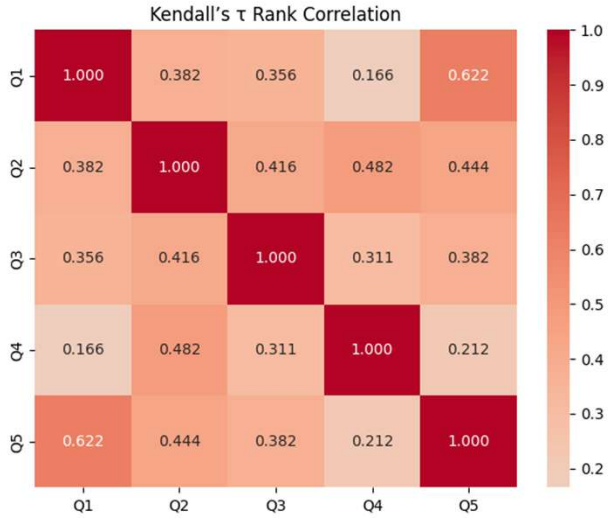
- Cosine Similarity
- Activation Overlap





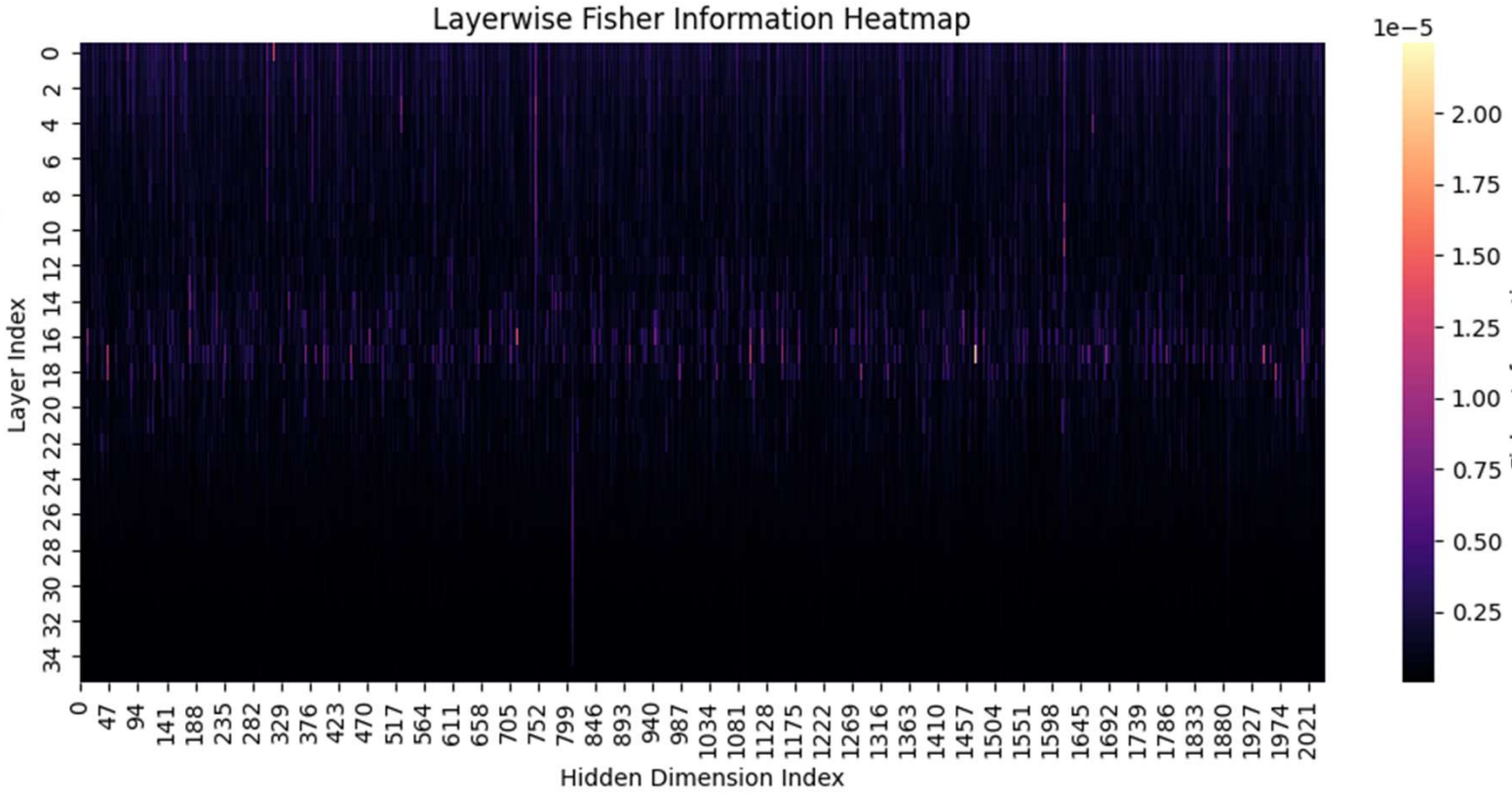






Other methods for similarity...

Fisher Information Heatmaps



Preliminary Results

"avg_embedding_similarity"	0.860328062375386
"avg_ranking_correlation"	-0.197889182058047

"Recommend 5 movies similar to The Matrix"
Inception (2010)
Blade Runner 2049 (2017)
Ex Machina (2014)
Snowpiercer (2013)
The Truman Show (1998)

"What are some movies resembling the Matrix? Top 5."
Inception (2010)
Blade Runner 2049 (2017)
Terminator Genisys (2015)
Ex Machina (2014)
Ghost in the Shell (2017)

"Can you recommend five films akin to The Matrix?"
Inception (2010)
Blade Runner 2049 (2017)
Ex Machina (2014)
Snowpiercer (2013)
Neon Genesis Evangelion (anime series)

"Suggest five films like The Matrix"
Inception (2010)
District 9 (2009)
Blade Runner 2049 (2017)
Ghost in the Shell (2017)
Ex Machina (2014)

"List 5 movies that are similar to The Matrix."
Inception (2010)
Blade Runner 2049 (2017)
Cloud Atlas (2012)
Ghost in the Shell (2017)
Atomic Blonde (2017)

"Top 5 movie suggestions similar to The Matrix"
Inception (2010)
Ghost in the Shell** (2017)
Blade Runner 2049 (2017)
Ex Machina** (2014)
Ready Player One** (2018)

Key Insights

- Fisher Information Matrices and Location of sensitive nuerons can provide valuable insights into how a model works
- LLM responses can vary greatly with slight variations in how prompts are written

Next Steps

- Re-evaluate and compare FIM from baseline to post-training

Questions?

References

- Fan, C., Zhao, Y., & Liu, S. (2024). SalUn: Gradient-Based Saliency Unlearning for Deep Neural Networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Retrieved from https://arxiv.org/abs/2403.04783
- Foster, J., Sattigeri, P., & Zhou, S. (2024). Selective Synaptic Dampening: Towards Efficient and Targeted Machine Unlearning. *AAAI Conference on Artificial Intelligence*. Retrieved from https://arxiv.org/abs/2402.08728
- Gandikota, R. (2023). Erasing Concepts from Diffusion Models. *IEEE/CVF International Conference on Computer Vision*. Retrieved from https://arxiv.org/abs/2303.07345
- Golatkar, A., Achille, A., & Soatto, S. (2020). Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Retrieved from https://arxiv.org/abs/1911.04933
- Hamad, M. (2025). *Decoupling Popularity Bias and User Fairness in LLM-Based Recommendation Systems*. Aalto University. Retrieved from https://aaltodoc.aalto.fi/items/68d19d5f-3e81-463a-93d8-9bdb52cfcf98
- Hu, J., Zhang, D., & Li, X. (2025). Relearning Attacks: Recovering Forgotten Knowledge in Machine Unlearning Models. *International Conference on Learning Representations*. Retrieved from https://arxiv.org/abs/2503.05821
- Jang, M., Kim, S., & Choi, E. (2025). *Negative Preference Optimization for LLM Unlearning*. arXiv. Retrieved from https://arxiv.org/abs/2502.09317
- Kim, S., Kang, H., Choi, S., Kim, D., Yang, M., & Park, C. (2024). Large Language Models Meet Collaborative Filtering: An Efficient All-Round LLM-Based Recommender System. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, (pp. 1395–1406). Retrieved from https://doi.org/10.1145/3637528.3671931
- Ko, H., Park, J., & Lee, S. (2025). Rethinking Gradient Directions for Machine Unlearning. *Neural Information Processing Systems*. Retrieved from https://arxiv.org/abs/2501.01234
- Kurmanji, M., Bernstein, J., & Ghalebikesabi, S. (2024). *Machine Unlearning: From Data Deletion to Concept Erasure*. arXiv. Retrieved from https://arxiv.org/abs/2408.08949
- Kurmanji, M., Ghalebikesabi, S., & Bernstein, J. (2023). Selective Disobedience in Neural Networks: Targeted Forgetting via Gradient Reversal. *Advances in Neural Information Processing Systems*. Retrieved from https://arxiv.org/abs/2311.17082
- Li, X., Zhao, C., Zhao, H., Wu, L., & He, M. (2024). *GanPrompt: Enhancing Robustness in LLM-Based Recommendations with GAN-Enhanced Diversity Prompts*. arXiv. Retrieved from https://arxiv.org/abs/2408.09671
- Liu, J., Yan, X., Li, D., Zhang, G., Gu, H., Zhang, P., . . . Gu, N. (2025). Improving LLM-powered Recommendations with Personalized Information. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 2560–2565). Retrieved from https://doi.org/10.1145/3726302.3730211
- Peng, Q., Liu, H., Huang, H., Yang, Q., & Shao, M. (2025). *A Survey on LLM-Powered Agents for Recommender Systems*. arXiv. Retrieved from https://arxiv.org/abs/2502.10050
- Sun, C., Liang, Y., Yang, Y., Xu, S., Yang, T., & Tong, Y. (2024). *Direct Preference Optimization for LLM-Enhanced Recommendation Systems*. arXiv. Retrieved from https://arxiv.org/abs/2410.05939
- Wang, Z., Gao, M., Yu, J., Gao, X., Nguyen, Q., Sadiq, S., & Yin, H. (2025). ID-Free Not Risk-Free: LLM-Powered Agents Unveil Risks in ID-Free Recommender Systems. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 1902–1911). Retrieved from https://doi.org/10.1145/3726302.3730003
- Wu, L., Zheng, Z., & Qiu, Z. (2024). A Survey on Large Language Models for Recommendation. *World Wide Web*. Retrieved from https://doi.org/10.1007/s11280-024-01291-2
- Zhu, K., Wang, J., Zhou, J., & Xie, X. (2024). PromptRobust: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts. *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, (pp. 57–68). Retrieved from https://doi.org/10.1145/3689217.3690621
- Zhu, X., & Jang, H. (2025). *Representation-Based Machine Unlearning for Large Language Models*. arXiv. Retrieved from https://arxiv.org/abs/2501.05542