Activation-Level Preference Unlearning for Robust LLM Recommenders

Final Project Proposal

ECE5554 Trustworthy Machine Learning | Fall 2025

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

*Equal Contributors

Motivation:

Large Language Models (LLMs) are increasingly used to power personalized recommender systems, offering natural interaction and transparent reasoning. However, recent studies show that the "preferences" expressed by these models are unstable and can be easily influenced by user phrasing or context. For instance, a slight change from *"Which movie should I watch?"* to *"Do you have any movie recommendations?"* can produce entirely different suggestions. These shifts reveal that LLMs do not merely reflect user's intent; they can also **reshape it**.

This phenomenon, known as **preference manipulation**, raises fundamental questions about how LLMs internalize, represent, and express user preferences. Understanding the mechanisms behind these shifts is essential before we can design reliable and trustworthy recommendation systems. Do LLMs amplify subtle linguistic cues? Are specific activations more sensitive to context framing? And how do conversational histories or reasoning chains contribute to preference drift?

Our work aims to **analyze and characterize** these dynamics by observing how model activations and outputs vary under semantically equivalent but linguistically distinct prompts. By combining prompt perturbation experiments with activation-level inspection, we aim to identify where and how manipulative behaviors emerge within the model. Through this exploration, we strive to provide insights into the boundaries between personalization, alignment, and manipulation, laying out the groundwork for future approaches to make LLM-powered recommenders both effective and trustworthy.

Related Work:

Recent studies (Wu, Zheng, & Qiu, 2024) demonstrate a shift from using LLMs as discriminative extractors (DLLM4Rec) to generative recommenders (GLLM4Rec), which directly produce personalized outputs. This evolution improves personalization but also increases sensitivity to prompt phrasing and interaction patterns, making manipulation easier. Frameworks like A-LLMRec (Kim et al., 2024) combine collaborative filtering with LLM world knowledge to handle different scenarios efficiently. However, such a tight coupling means that even slight changes to prompts or contexts can shift preferences.

To address these issues, recent works have begun exploring methods to enhance prompt robustness and mitigate manipulation risks. For instance, PromptRobust (Zhu, Wang, Zhou, & Xie, 2024) introduces adversarial prompt benchmarks to assess how minor textual perturbations impact model behavior. GANPrompt (Li, Zhao, Zhao, Wu, & He, 2024) utilizes GAN-generated diverse prompts to enhance the stability of LLM-based recommenders against changes in phrasing.

In parallel, the broader machine-unlearning literature provides valuable context for understanding how models internalize and potentially alter user preferences. Techniques such as Selective Synaptic Dampening and Saliency-Guided Unlearning [Foster et al., 2024; Fan et al., 2024] reveal that knowledge can be localized to specific neuron groups or activation pathways. Similarly, gradient projection and random labeling loss methods demonstrate how targeted updates can distort or erase learned associations without full retraining. These findings suggest that preference manipulation in LLM recommenders may emerge from comparable internal mechanisms - where subtle prompt cues activate overlapping neural representations linked to prior user contexts or biases.

Recent works on LLM unlearning, including Representation-based Machine Unlearning (RMU) and Negative Preference Optimization (NPO) [Ko et al., 2025; Jang et al., 2025], further highlight that modifying activation patterns through gradient ascent or regularization can change model behavior, sometimes unintentionally. Studying these internal dynamics may therefore shed light on how preference drift occurs - offering a foundation for later efforts to identify, interpret, and eventually mitigate manipulation within generative recommender systems.

Early-Stage Ideas:

CoT-Rec (Liu et al., 2025) consists of two stages: (1) *personalized information extraction*, where user preferences and item perceptions are derived, and (2) *personalized information utilization*, where these features are incorporated into the LLM-powered recommendation process. While this approach strengthens personalization, it also raises new questions: How stable is the extracted information across prompts? Does the model overfit linguistic framing rather than true intent?

To investigate this, we propose treating preference manipulation as a targeted unlearning problem - not to erase data, but to understand *which internal activations* and *linguistic features* cause unstable behavior. Drawing from recent research on selective forgetting (Kurmanji et al., 2023; Foster et al., 2024; Ko et al., 2025), our idea is to:

1. **Characterize instability**: Create prompt perturbation experiments where semantically equivalent queries produce divergent recommendations. Quantify "preference drift" using ranking correlation (e.g., Kendall's τ) and embedding similarity.
2. **Localize manipulation**: Use *saliency maps* or *Fisher Information Matrix (FIM)* analysis to identify which neurons or attention heads contribute most to phrasing sensitivity - like *synaptic dampening* in unlearning literature.
3. **Analyze activation overlap**: Examine how shared neuron clusters encode multiple "classes" of preferences. Overlapping activations may explain why minor prompt changes alter recommendations.
4. **Test gradient effects**: Experiment with *gradient ascent/descent contrasts* (as in RGD and SalUn) to probe whether specific gradient directions correspond to manipulated versus stable preferences.
5. **Relate to alignment**: Connect findings to alignment tuning - if manipulative cues activate distinct pathways, we can eventually design retention mechanisms to reinforce user intent rather than erasing instability.

By combining these methods, we aim to uncover *how* LLM recommenders internalize and distort preference signals, laying out the foundation for trustworthy personalization.

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