The Blessing of Reasoning: LLM-Based Contrastive Explanations in Black-Box Recommender Systems

Yuyan Wang[1], Pan Li[2], Minmin Chen[3]
[1]Stanford University, [2]Georgia Institue of Technology, [3]Google DeepMind

Abstract

Modern recommender systems use ML models to predict consumer preferences based on consumption history. Although these ``black-box'' models achieve impressive predictive performance, they often suffer from a lack of transparency and explainability. While explainable AI research suggests a tradeoff between the two, we demonstrate that combining large language models (LLMs) with deep neural networks (DNNs) can improve both. We propose **LR-Recsys**, which augments state-of-the-art DNN-based recommender systems with LLMs' *reasoning* capabilities. LR-Recsys introduces a contrastive-explanation generator that leverages LLMs to produce human-readable positive explanations (why a consumer might like a product) and negative explanations (why they might not). These explanations are embedded via a fine-tuned AutoEncoder and combined with consumer and product features as inputs to the DNN to produce the final predictions. Beyond offering explainability, LR-Recsys also improves learning efficiency and predictive accuracy. To understand why, we provide insights using high-dimensional multi-environment learning theory. Statistically, we show that LLMs are equipped with better knowledge of the important variables driving consumer decision-making, and that incorporating such knowledge can improve the learning efficiency of ML models.

Extensive experiments on three real-world recommendation datasets demonstrate that the proposed LR-Recsys framework consistently outperforms state-of-the-art black-box and explainable recommender systems, achieving a 3–14% improvement in predictive performance. This performance gain could translate into millions of dollars in annual revenue if deployed on leading content recommendation platforms today. Our additional analysis confirms that these gains mainly come from LLMs' strong reasoning capabilities, rather than their external domain knowledge or summarization skills.

Research Question

Existing literature in Explainable AI shows that there is a trade-off between explainability and accuracy => ***No*** evidence that explanations help model performance.

Research Question:

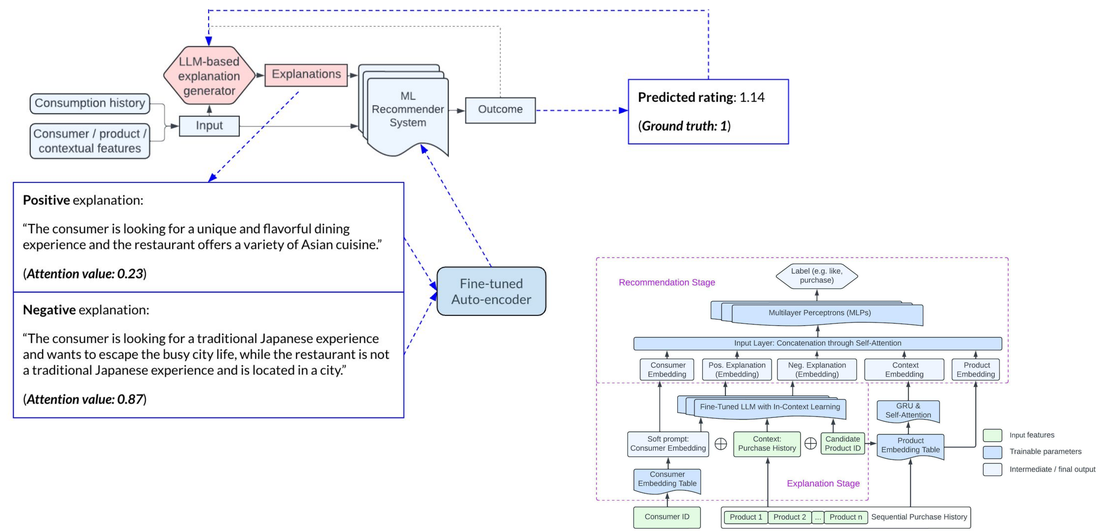
Can (LLM-generated) **explanations improve** the performance of black-box recommender systems?

– *Likely, by improving learning efficiency*

Challenges:

1. LLMs themselves are **NOT** good recsys
   - Generative tasks (LLMs) vs. discriminative tasks (Recsys)
   - Underperforms DNNs when used directly for prediction [Tsai et al. 2024, Ye et al. 2025]
2. The value of **unstructured natural language explanations** to DNN-based Recsys is unclear
   - Existing solutions convert reasoning into structured formats (e.g., graphs) [Wang et al. 2024]

Proposed Framework: LLM-Reasoning-Powered Recommender System (LR-Recsys)



Motivation: Leveraging LLMs for Recsys Explanation

Toy example: LLMs can explain consumer choices based on consumption history:

Positive explanation

Purchase history: apple, orange, watermelon, …
New purchase: orange juice in a paper box with a sun as the logo
Reason for purchase?

The consumer purchased this product because the consumer regularly buys fruits including oranges, and the product is an orange juice which aligns with her existing preferences.

↓

Identifies associations between purchased items (e.g., oranges → orange juice)

Negative explanation

Purchase history: apple, orange, watermelon, …
Did NOT purchase: orange juice in a paper box with a sun as the logo
Reason for **no** purchase?

The consumer did not purchase this product because the consumer may prefer whole fruits over processed juices, and the product is an orange juice, which may not align with her preference for whole, natural fruits.

↓

Highlights nuanced differences (whole vs. processed)

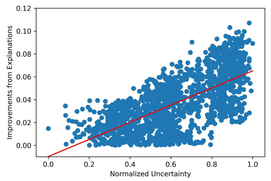
⇒ LLMs can identify associations in a ***zero-shot*** fashion (i.e. without training data), unlike traditional recsys that require thousands of examples.

Results

	TripAdvisor			Yelp			Amazon Movie		
	RMSE ↓	MAE ↓	AUC ↑	RMSE	MAE	AUC ↑	RMSE ↓	MAE ↓	AUC ↑
LR-Recsys (Ours)	**0.1889** (0.0010)	**0.1444** (0.0008)	**0.7289** (0.0018)	**0.2149** (0.0010)	**0.1685** (0.0009)	**0.7229** (0.0017)	**0.1673** (0.0010)	**0.1180** (0.0009)	**0.7500** (0.0018)
% Improved	+5.36%***	+15.11%***	+2.88%***	+11.31%***	+18.64%***	+3.01%***	+20.30%***	+33.33%***	+3.65%***

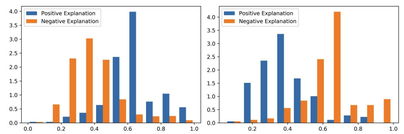
"Harder" examples benefit more:

Hypothesis: Reasoning provides greater value on ***harder*** examples (measured by uncertainty) where consumer decisions are less obvious.



Value of pos vs. neg explanations:

Our model learns to focus on the right type of explanation depending on the outcome…



(a) Positive examples (Yelp Dataset). (b) Negative examples (Yelp Dataset).

Left: Positive examples → higher attention on positive explanations
Right: Negative examples → higher attention on negative explanations

… Mirroring how humans justify decisions.

Theoretical Insights

S* - the subset of important variables that predicts the outcome (as in high-dimensional statistical learning)
*(e.g. for orange juice: "**ingredient**", "packaging", "logo", …)*

Lemma 1: LLMs have better knowledge of S* than the training data itself.

$$\mathbf{P}(supp(\hat{\beta}_L) = S^*) \to 1$$

- LLMs have "seen" similar decision-making in **multiple** environments (orange juice purchase in supermarkets / online) ⇒ more likely to identify relevant variables **across** contexts (*multi-env learning*)

Lemma 2: Better knowledge of S* leads to better model performance.

- Without the knowledge of S*: $o_P\left(\sqrt{\frac{s \log p}{n}}\right)$
- With the knowledge of S*: $o_P\left(\sqrt{\frac{s}{n}}\right)$

Faster convergence, lower generalization error

Contact

Yuyan Wang: yuyanw@stanford.edu
Assistant Professor of Marketing, Kevin J. O'Donohue Family Faculty Scholar for 2024-2025

Stanford Graduate School of Business
655 Knight Way
Stanford, CA 94305
United States

References

1. Tsai, Alicia Y., et al. "Leveraging LLM Reasoning Enhances Personalized Recommender Systems." arXiv preprint arXiv:2408.00802 (2024).
2. Ye, Zikun, Hema Yoganarasimhan, and Yufeng Zheng. "Lola: Llm-assisted online learning algorithm for content experiments." Marketing Science (2025).
3. Wang, Xiang, et al. "Explainable reasoning over knowledge graphs for recommendation." Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.