MAKING LLM-BASED RECOMMENDER SYSTEMS RELIABLE: TESTING, UNDERSTANDING, AND IMPROVING CONSISTENCY

Same user. Same request. Different answer?

Can we trust AI recommenders every time? We test them with Metamorphic Testing.

Introduction

• LLM recommendations can change when the same preferences are expressed differently.

• Accuracy alone does not establish consistent or appropriate behaviour.

• Metamorphic testing examines expected relationships under controlled input changes.

Unchanged prompts: high agreement

98.14%

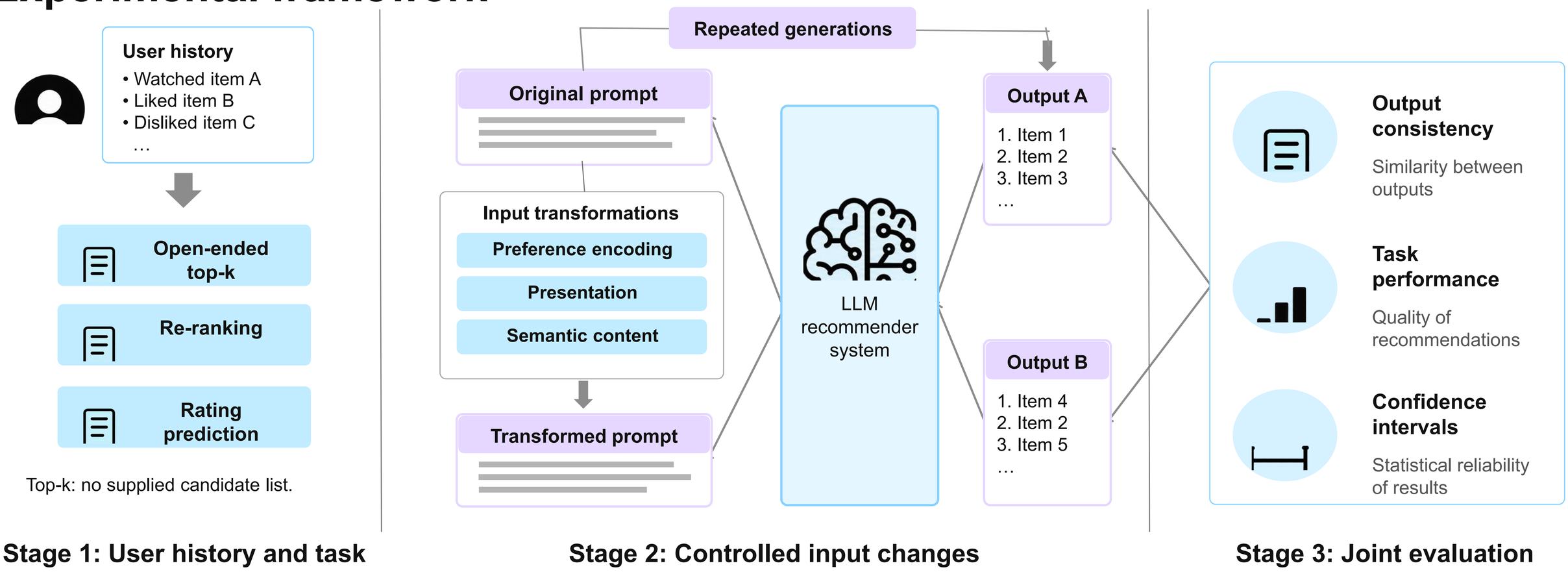
Llama · exact-set agreement
95% CI: 97.71–98.55%
n = 2,026

99.72%

Qwen · exact-set agreement
95% CI: 99.62–99.81%
n = 7,296

Amazon Arts, top-5. Different user cohorts.
User confidence intervals condition on recorded generations.

Experimental framework



Preliminary results: repeatable does not mean robust

Presentation changes can disrupt list agreement

	Unchanged	Typos
Llama	0.9896	0.0009
Qwen	0.9994	0.0001

RBO similarity, Amazon Arts top-k. 0 = dissimilar; 1 = identical.

High repeatability does not guarantee prompt robustness.

In the reported presentation test,
small edits reduce RBO almost to zero.

Repeatability and perturbation sensitivity
must be evaluated separately.

What the findings mean

Changed output needs interpretation

The results show insconsistency in LLM based RS which needs to be further investigated.

Next, test which item the user prefers across small and large rating gaps in prediction and re-ranking.

Next steps and health relevance

Future directions

Targeted fine-tuning and activation patching
to investigate and address observed failures.

Test transfer to health

The proposed testing method can be applied across disciplines, including healthcare, where LLMs are used to assess consistency.

Madhurima Khirbat 1, 2, Yongli Ren 1, Estrid He 1, Mark Sanderson 1, 2

[1]RMIT University,
[2]ARC Centre of Excellence for Automated Decision Making + Society

This research was conducted by the ARC Centre of Excellence for Automated Decision-Making and Society (CE200100005) and funded fully by the Australian Government through the Australian Research Council.