Evaluating Prompting Strategies for Reliable LLM-Based User Simulation in Information Retrieval

EEMCS, Delft University of Technology, The Netherlands

Ziliang Zhang
z.zhang-78@student.tudelft.nl
Supervisor: Avishek Anand
Delft University of Technology

Background Information

Current Limitations:

- **Offline Evaluation (e.g., nDCG, MAP):** Computationally efficient and reproducible, but relies on sparse human judgments that abstract away real user behavior and satisfaction.
- **Online Evaluation (e.g., A/B testing):** Captures genuine user interactions (clicks, dwell time), but is expensive, slow, and difficult to reproduce at scale.

The Gap: This creates a persistent mismatch between what we can easily measure offline and what users actually experience online.

LLM-Based Simulation as a Bridge:

- **Promise:** Large language models can generate synthetic user interactions at scale, avoiding the cost and privacy concerns of human studies.
- **Critical Challenge:** The behavior of these simulators is not inherent; it is fundamentally shaped by their prompt instructions. Minor variations yield vastly different outcomes.

Research Question

Which prompting strategy produces the most reliable LLM-based user simulations for online IR evaluation?

Methodology

Research Goal: Systematically evaluate how prompting affects simulation reliability, providing actionable guidelines for IR practitioners. **Evaluation Criteria:** We define reliability as consistency across runs, sensitivity to system quality differences, and realism compared to human patterns. **Three Prompting Strategies Compared:**

- **One-Shot (OS):** Minimal guidance. Provides only a core instruction and a single, generic example. The LLM must infer the complete interaction pattern, leading to reliance on its internal priors.
- **Multi-Shot (MS):** Example-rich guidance. Provides multiple distinct behavioral examples, illustrating varied responses to documents of differing quality. Offers abstract principles without over-constraining.
- **Offline-Evaluation-informed (OE):** Data-driven guidance. Explicitly injects relevance signals (e.g., graded judgments, nDCG scores) into the prompt, instructing the LLM to use them as the primary basis for generating clicks and dwell times.

Reliability Assessment Framework: To move beyond single-metric comparisons, we evaluate each strategy across three core dimensions:

- **Stability (Consistency):** Measures how consistently the simulator behaves across different IR systems under the same prompting strategy (low variance is desired).
- **Discrimination (Sensitivity):** Assesses whether the simulator can reliably distinguish between retrieval systems of different quality (high sensitivity is desired).
- **Plausibility (Realism):** Evaluates alignment with established patterns from real user interaction studies (e.g., typical CTR ranges, position bias).

A composite reliability score weights Stability and Discrimination most heavily (0.4 each), as they are critical for comparative evaluation.

Sensitivity Analysis: We tested weight variations (0.3-0.5 for each dimension) and found MS consistently outperformed OS and OE across all reasonable weightings, confirming the robustness of our conclusion.

Experimental Setup & Online Evaluation Results

Experimental Setup: Using TREC DL19/DL20 with 52 diverse retrieval systems, we simulate ∼300k user interactions via **Mistral-7B**. Five key online metrics—**Click-Through Rate (CTR), Average Dwell Time (ADT), Success Satisfaction Rate (SSR), Session Abandonment Rate (SAR), Zero Result Rate (ZRR)**—are computed from the simulated behavior, with **prompting strategy** as the sole controlled variable.

Metric Patterns by Strategy:

	OS	MS	OE
SSR (%)	62–69	80–88	88–96
CTR	0.175–0.185	0.180–0.190	0.190–0.200
ADT (s)	28–52	30–58	22–40
Variance	**High**	**Moderate**	**Low**

Interpretation:

- **Guidance ↑ Engagement ↑:** SSR rises steadily OS→MS→OE
- **Trade-off:** OE boosts engagement but reduces behavioral diversity
- **Optimal Balance:** MS offers variability to differentiate systems + reasonable consistency
- **Consistency Check:** SAR and ZRR remain closely aligned (difference < 2%)

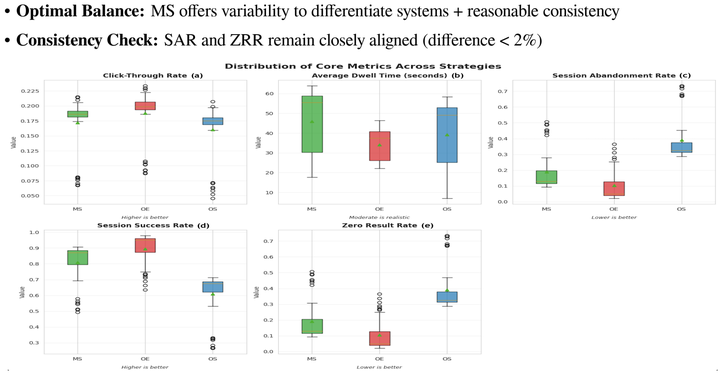

Fig 1: Distribution of five metrics across strategies.

Behavioral Coherence Analysis:

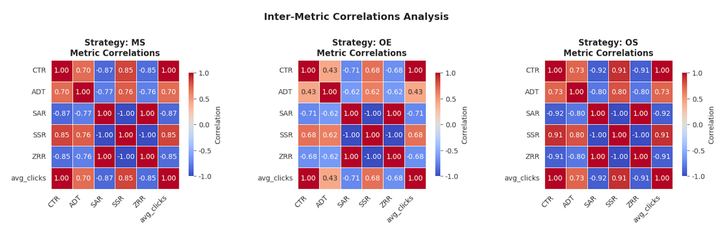

Fig 2: Correlation matrices showing internal metric relationships.

Key Correlation Insights:

- **MS:** Coherent logic (CTR-SSR: 0.85) → integrated, user-like decisions
- **OE:** Weaker correlations (0.62) → compartmentalized metric generation
- **OS:** Extreme correlations (-0.92) → rigid, binary rules

Connection: These internal logics directly explain reliability outcomes.

Reliability Result & Conclusions

Reliability Scores:

	OS	MS	OE
Stability	0.665	**0.751**	0.706
Discrimination	0.485	**0.566**	0.514
Plausibility	0.899	0.912	**0.942**
Composite	0.615	**0.709**	0.655

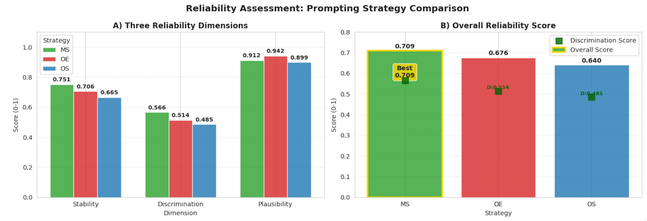

Fig 3: Stability, discrimination, plausibility, and composite scores.

Conclusions:

- **Multi-Shot prompting achieves highest overall reliability** (composite score 0.709), outperforming Offline-Evaluation-informed by 8.3% and One-Shot by 15.3%.
- **Optimal balance across dimensions:** Excels in stability (0.751) and discrimination (0.566)—the most critical aspects for comparative evaluation—while maintaining strong plausibility (0.912).
- **Prompting fundamentally shapes simulation behavior:** Our 3D assessment reveals distinct behavioral patterns across strategies (MS: coherent logic, OE: compartmentalized, OS: rigid rules).
- **Practical implication:** For scalable, trustworthy LLM-based user simulation in IR, adopt Multi-Shot as the default prompting strategy.

Limitations & Future Work

Limitations:

- **Scope constraints:** Evaluation limited to Mistral-7B on TREC datasets; generalization to larger models and diverse domains needed.
- **Simulation realism:** While plausible, simulated behaviors may not fully capture real user complexity and noise patterns.
- **Prompting strategies:** Tested three foundational approaches; advanced techniques (chain-of-thought, self-refinement) remain unexplored.

Future Directions:

- **Generalization studies:** Extend evaluation to larger LLMs (GPT-4, Claude) and application domains (e-commerce, healthcare, legal search).
- **Technical integration:** Embed reliability assessment directly into simulation pipelines for real-time quality monitoring.
- **Ethical auditing:** Systematically evaluate simulation outputs for bias, fairness, and representativeness across user groups.

Final Recommendation: Adopt **Multi-Shot prompting** as the default strategy for LLM-based user simulation in IR—it offers the optimal balance of consistency, sensitivity, and realism for reliable comparative evaluation.

References:

H. Zamani, J. R. Trippas, J. S. Culpepper, F. Radlinski. "Simulating User Interactions for Evaluation of Conversational Information Retrieval." *SIGIR* 2020.

J. Wei, X. Wang, D. Schuurmans, et al. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." *NeurIPS* 2022.

P. Thomas, A. S. Cofield, M. Czerwinski, et al. "Methods for Evaluating Offline Evaluation Metrics." *WSDM* 2022.