Whom to Trust? Analyzing the Divergence Between User Satisfaction and LLM-as-a-Judge in E-Commerce RAG Systems

Arif Türkmen
Trendyol
arif.turkmen@trendyol.com

Kaan Efe Keleş
Trendyol
efe.keles@trendyol.com

Abstract

We study retrieval-augmented generation (RAG) evaluation in the Trendyol QA Assistant using 150k real e-commerce interactions. Our framework combines user satisfaction labels, *LLM-as-a-judge* scoring, and factor-based diagnostics to separate retrieval from generation errors. We find that judge models broadly reflect user satisfaction trends, though important nuances of dissatisfaction are often missed. Factor-level analysis highlights systematic error patterns across query types and context quality, demonstrating that hybrid evaluation, combining multiple LLM judges with direct user feedback offers the most reliable assessment strategy for production RAG systems.

1 Introduction

Retrieval-augmented generation (RAG) has risen as an effective paradigm for grounding large language models (LLMs) in customer-facing applications such as help desks, product Q&A, and enterprise copilots, by conditioning generation on retrieved domain knowledge—for greater factuality and relevance [Lewis et al., 2021, Yu et al., 2025]. Yet, evaluating RAG systems remains challenging because overall quality depends jointly on retrieval quality, context use, and generation [Lewis et al., 2021, Yu et al., 2025]. Recent benchmarks like *RAGBench* introduce explainable and modular metrics to assess retrieval and generation components [Friel et al., 2025], while retrieval-centric methods (e.g., *eRAG*) quantify document contributions to final answers [Yu et al., 2025]. Complementing these, the emerging "LLM-as-a-judge" paradigm allows scalable, automatic evaluation by having stronger models score responses—but systematic biases and blind spots remain, as highlighted in recent survey and bias studies [Gu et al., 2025, Ye et al., 2024].

We examine these evaluation modalities in a large-scale, real-world deployment: the *Trendyol QA Assistant*. Trendyol—Türkiye's largest e-commerce platform—handles hundreds of thousands of product questions daily, many of them similar to earlier question-answer pairs. The assistant addresses this by retrieving semantically similar historical Q/A pairs from past interactions with domain specific embedding model[1], and conditioning a domain-specific LLM on this retrieved evidence[2] to generate a concise, grounded answer. In effect, the system automates seller responses by reusing and synthesizing prior knowledge, reducing latency and seller workload while preserving answer quality.

Our work makes three primary contributions. First, we present what is, to our knowledge, the first large-scale analysis of user satisfaction in a deployed e-commerce RAG system, using 150k real interactions from the Trendyol QA Assistant. Second, we provide the first empirical study demonstrating the divergence between LLM judge predictions and human preferences in Turkish-language generative outputs. Third, we document systematic evaluation failures of popular judge models in the e-commerce domain, identifying context-dependent blind spots that limit their reliability as standalone judges.

2 Evaluation Framework for User Satisfaction

A central challenge in evaluating the Trendyol QA Assistant lies in capturing end-user satisfaction at scale. Online A/B tests provide high-quality ground truth but are expensive and slow to iterate, motivating scalable alternatives. We therefore adopt a three-pronged evaluation methodology, combining direct human feedback with structured LLM-based approaches.

[1] TY-ecomm-embed-multilingual-base-v1.2.0
[2] Trendyol-LLM-8B-T1

Proceedings of the First Workshop on Multilingual Multicultural Evaluation, pages 189–195

2.1 Direct User Feedback

Ground-truth performance is obtained from user-reported satisfaction collected during live usage. After each interaction, users are asked in a thumbs-up or thumbs-down fashion whether they were satisfied with the assistant's answer. If dissatisfied, they may optionally select one of four categorical reasons: *irrelevant*, *insufficient/incomplete*, *unclear*, or *misleading/incorrect*. This feedback provides the most accurate measure of user experience, though it is costly to scale and limited in experimental coverage.

2.2 LLM-as-a-Judge Simulation

To complement human feedback, we employ *LLM-as-a-judge* techniques, in which stronger models are prompted to evaluate QA interactions automatically using a fixed, structured judging prompt (Appendix A). For each interaction, the judge receives four inputs: (i) the user query, (ii) retrieved similar Q/A pairs, (iii) the assistant's final response, and (iv) the assistant's internal prompts used for generation (system prompt and base user prompt with examples). We use a few-shot version of this prompt and experimented with multiple candidate few-shot sets; a configuration with 10 diverse examples (spanning query types and common failure modes) yielded the best agreement on a held-out validation subset. The judge returns (a) a binary satisfaction decision and (b) if dissatisfied, one of the same four standardized reasons provided to users.

To assess alignment with users, we report two complementary measures. (i) *Satisfaction Agreement*: the exact match rate between the predictions of the judge and the satisfaction of the user in all interactions, contextualized with the expected chance agreement and Cohen's κ [Cohen, 1960]. (ii) *Dissatisfaction Breakdown*: the categorical distribution of dissatisfaction reasons, enabling direct comparison between user-reported and judge-assigned error types.

These measures jointly assess sentiment alignment and dissatisfaction modeling, identifying where judge models fail to match real user feedback.

2.3 Factor-Based LLM Analysis

Finally, we use a granular, multi-prompt technique where the LLM isolates and evaluates specific factors. Prompts target different aspects to categorize question-answer pairs:

- Query Classification: It determines the topic of the question. Is the user asking about the product or the seller?
- Intent Analysis: It identifies the user's goal. Is the user asking a genuine question or making a demand?
- Contextual Relevance: Did the information retrieved actually contain the necessary details to address the user's query?
- Persona Consistency: Does the answer maintain the assistant's intended style (formal tone, third-person narration) throughout?

Alongside these binary LLM outputs, we use string length as a verbosity measurement. This provides a consistent, model-agnostic metric, unlike token counts, which vary between different tokenizers. This factor-level decomposition helps identify whether user unsatisfaction arises from retrieval, grounding, or generation, providing actionable insights.

Our analysis is grounded in approximately 150k QA interactions from Trendyol's latest production environment to ensure stability. Each instance contains the user query, retrieved Q/A pairs, the assistant's response, and binary user satisfaction labels; dissatisfied users additionally provided categorical reasons. It is important to note that feedback collection relies on user initiative; while this means coverage is not universal, it provides a realistic representation of production dynamics where explicit feedback is sparse. This supervision enables systematic comparison between human feedback, LLM judges, and factor-based diagnostics at scale.

3 Results

Table 1 summarizes alignment between LLM-as-a-judge predictions and human feedback. The human-reported satisfaction prevalence is 77.2%. The two judges differ in calibration to this baseline (65.0% vs. 76.3%), with *o4-mini* closer to the target. However, agreement with user labels is only modest: exact-match rates are 64.5% (GPT-4o) and 72.5% (o4-mini), which translate to Cohen's κ of 0.15 and 0.23, respectively. Given the class imbalance, these κ values indicate only slight-to-fair agreement beyond chance. We therefore treat LLM-as-a-judge as a useful heuristic for binary satisfaction but not a calibrated substitute for direct user feedback.

Table 1: Calibration and agreement of LLM-as-a-judge vs. human labels on 150k QA interactions. Δ is model prevalence minus human prevalence. Agreement is exact match rate. Cohen's κ adjusts for chance agreement using observed marginals.

Model	Satisfaction (%)	Δ vs Human (pp)	Agreement (%, 95% CI)	Cohen's κ
GPT-4o	65.00	−12.2	64.48 [64.24, 64.72]	0.15
o4-mini	76.34	−0.83	72.53 [72.30, 72.76]	0.23
Gemini 2.5 Pro	80.60	+3.40	73.47 [72.60, 74.33]	0.20
Claude 4 Sonnet	78.72	+1.52	74.97 [74.12, 75.82]	0.27

Human satisfaction prevalence: 77.2%. Expected agreement by chance from marginals: 58.1% (GPT-4o), 64.3% (o4-mini), 66.6% (Gemini 2.5 Pro), 65.6% (Claude 4 Sonnet).

Table 2 compares the categorical breakdown of dissatisfaction reasons between users and LLM judges, each measured over their own dissatisfied subsets.

User feedback is dominated by insufficient/incomplete (62.3%), with meaningful fractions of unclear (15.7%) and misleading/incorrect (14.2%). However, the LLMs do not approximate this distribution. Since they can access retrieved contexts unlike users, they identify misleading/incorrect cases more aggressively.

These discrepancies reveal that while LLM judges may track overall satisfaction rates, they struggle to accurately model user behavior and may misrepresent user attitudes. By over-assigning "incorrectness" and overlooking issues like clarity or relevance, they provide a distorted diagnostic view. As a result, relying solely on judge distributions risks missing the user-centered issues that matter most for customer impact.

Table 2: Distribution of dissatisfaction reasons (%) across user feedback and LLM judges. Each row shows the categorical prevalence among all dissatisfied cases for that source.

	Irrelevant	Insufficient/ Incomplete	Unclear	Misleading
Users	7.8	62.3	15.7	14.2
GPT–4o	0.4	37.4	0.2	62.0
o4–mini	1.8	45.1	0.3	52.9
Gemini 2.5 Pro	0.7	45.7	0.8	52.8
Claude 4 Sonnet	0.9	59.1	0.8	39.2

As shown in Table 3, all models except GPT-4o achieve higher satisfaction with short responses than with long responses. Notably, Gemini 2.5 Pro attains the highest satisfaction under the long-response stratum (75.3%), which likely contributes to its strong overall satisfaction (80.6%) in this evaluation.

We zoom in on the two slices with the lowest human satisfaction: low-relevance contexts and seller-related queries. As shown in Table 4, when context is weak, agreement significantly declines. Since users do not see the ground truth or relevant contexts on the UI; their judgments on these types of questions are not reliable. LLM judge systems have access to retrieved contexts and they can identify the misinformation and give worse scores than users, as shown in Table 3.

Table 3: Satisfaction rates (%) across different interaction strata. The categories analyze performance based on query type, context quality, and response length.

Category	Human	GPT-4o	o4-mini	Gemini	Claude
Query Type					
Product-Related	77.9	65.2	76.5	80.9	79.3
Seller-Related	64.9	62.9	70.4	73.7	67.5
Factual Q&A	77.5	65.3	76.3	79.4	77.1
User Commands	73.7	63.6	74.9	87.6	87.9
Context Quality					
High-Relevance	79.6	69.7	80.8	84.0	83.0
Low-Relevance	61.1	35.4	46.3	53.4	44.0
Response Length					
Short Response	80.3	64.9	78.5	81.9	80.7
Long Response	66.6	65.9	68.1	75.3	70.8

Table 4: Agreement of LLM-as-a-judge vs. human labels on questions that came with weakly related context.

Model	Agreement (%, 95% CI)	Cohen's κ
GPT-4o	49.89[49.16, 50.62]	0.06
o4-mini	57.47[56.75, 58.19]	0.16
Gemini 2.5 Pro	57.92[55.03, 60.82]	0.15
Claude 4 Sonnet	56.88[53.95, 59.81]	0.15

Human satisfaction prevalence: 61.1%. Expected agreement by chance from marginals: 46.5% (GPT-4o), 49.1% (o4-mini), 52.1% (Gemini), 49.2% (Claude).

Another important slice is **seller-related queries** (Table 5). In a marketplace like Trendyol, many products have multiple sellers; user questions about shipping or packaging are therefore seller-specific, and the "correct" answer varies by seller. In this slice, o4-mini's κ score sees a modest deterioration

from 0.23 to 0.21. However, Claude 4 Sonnet impacted heavily by seller-related questions, a major κ score decline 0.27 to 0.15. By contrast, GPT-4o shows a modest improvement, with Cohen's κ rising from 0.15 to 0.19. These results underscore that LLM-judge performance is dependent on various factors and must be interpreted in context.

Table 5: Agreement of LLM-as-a-judge vs. human labels on seller-related questions.

Model	Agreement (%, 95% CI)	Cohen's κ
GPT-4o	62.48[61.35, 63.60]	0.19
o4-mini	65.71[64.61, 66.81]	0.21
Gemini 2.5 Pro	62.93[58.66, 67.20]	0.15
Claude 4 Sonnet	62.70[58.45, 66.96]	0.17

Human satisfaction prevalence: 64.9%. Expected agreement by chance from marginals: 53.5% (GPT-4o), 56.1% (o4-mini), 57.0% (Gemini), 54.7% (Claude).

4 Discussion

Our findings highlight both the promise and pitfalls of LLM-as-a-judge for evaluating production RAG systems. Among the four judges evaluated, Claude 4 Sonnet achieves the highest agreement with users with 74.97% agreement and $\kappa = 0.27$, suggesting stronger alignment with human judgment patterns. A notable finding is the performance gap between GPT-4o and o4-mini, both from the same model family but differing in their reasoning capabilities. While GPT-4o achieves only 64.48% agreement ($\kappa = 0.15$), o4-mini—an inference-time reasoning model—improves to 72.53% agreement ($\kappa = 0.23$). This suggests that using reasoning models in evaluation tasks may help to simulate user judgment patterns, particularly when assessing complex QA interactions.

Our stratified analysis reveals that no single judge model performs uniformly well across all interaction types. In seller-related queries, Claude 4 Sonnet experiences the largest relative decline. These findings caution against deploying a single judge model for all query types and advocate for segment-aware evaluation strategies.

Beyond binary agreement, our analysis of dissatisfaction categories (Table 2) reveals a critical blind spot: LLM judges systematically over-attribute failures to "Misleading/Incorrect" (39–62%) while under-reporting "Insufficient/Incomplete" (37–59% vs. user-reported 62.3%). This asymmetry likely stems from judges' access to retrieved contexts, enabling them to detect factual inconsistencies invisible to users. Teams relying solely on LLM-judge metrics risk optimizing for factual accuracy while neglecting completeness, a mismatch with actual user pain points.

The findings above collectively argue against relying on any single LLM judge for a real system evaluation. Each model exhibits blind spots and all judges systematically misattribute user dissatisfaction categories. We therefore advocate for a *model ensemble* approach, where diverse judges vote to mitigate individual biases. More fundamentally, our results suggest that the most reliable evaluation strategy is a *hybrid* system combining LLM-judge with direct human feedback. LLM judges offer scalability and consistency for detecting factual errors and retrieval failures, while human ratings capture the subjective dimensions of satisfaction that models systematically miss.

5 Limitations

Our study is limited to a single production system within Trendyol, reflecting the characteristics of Turkish e-commerce queries. User feedback is binary with a restricted set of dissatisfaction categories, which constrains expressiveness. LLM-as-a-judge results are sensitive to prompt design and model choice, and we evaluate only four specific models.

References

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2025. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *Preprint*, arXiv:2407.11005.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and

Xiangliang Zhang. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *Preprint*, arXiv:2410.02736.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2025. *Evaluation of Retrieval-Augmented Generation: A Survey*, page 102–120. Springer Nature Singapore.

A Evaluation Prompt for LLM-as-a-Judge

LLM-as-a-Judge Prompt (System + User)

System prompt

```
You are an impartial expert evaluator for an e-commerce Q&A assistant.
You will receive the assistant's internal prompts, the user's question,
the retrieved similar question/answer pairs, and the assistant's final answer.
Your job is to critically assess the retrieval quality and the final answer,
and then produce a concise, strictly structured JSON evaluation.

Be precise, objective, and consistent. Do not invent facts beyond the provided content.
```

User message

```
Evaluation context (for transparency, not to be judged for style):

Assistant system prompt: [SYSTEM_PROMPT]
Assistant base user prompt (including few-shot examples): [USER_PROMPT]

Inputs to evaluate:
1) User question: [QUESTION_TEXT]
2) Retrieved similar question/answer pairs: [RETRIEVED_QAS]
3) Assistant final answer to the user: [LLM_ANSWER]

Evaluation requirements:
- Judge whether retrieved QA pairs are semantically relevant to the user's question.
- Judge whether the assistant's final answer directly addresses the user's question using only
↪  retrieved information.
- Penalize hallucinations or unsupported additions.
- If retrieval is irrelevant or insufficient, the correct assistant behavior is:
  "Üzgünüm bilmiyorum" (I am sorry, I do not know). Penalize deviations.

Output format (strict JSON object):
- question_analysis: string
- retrieved_answers_analysis: string
- llm_answer_analysis: string
- satisfaction_feedback_analysis: string
- satisfaction_feedback_boolean: boolean
- satisfaction_feedback_negative_reason: string
  (if dissatisfaction: one of "Irrelevant", "Insufficient/Incomplete",
  "Unclear", "Misleading/Incorrect"; else "None").

Guidance:
- Be concise but specific; reference concrete retrieved evidence.
- Treat greetings or seller politeness in retrieved QA pairs as noise.
- Reflect violations (fabrication, irrelevance, improper style) in analysis.
- For seller requests (e.g., "send red color"), the correct behavior is to decline with
  "I am sorry, I don't know.". Reward this.

Few-shot examples (redacted):
The judge prompt includes 10 such examples in practice.
To keep the appendix short, we do not show all examples here.
```

Redacted few-shot example (one shown)

One Redacted Example (Format Illustration)

```
Product context (English, brand-redacted):
"21V Brushless Impact Rotary Hammer Drill & Grinder & Drill & Nut Tightening/Loosening 4-piece
↪  Set"

Inputs:
1) User question: [EXAMPLE_QUESTION_TEXT]
2) Retrieved similar question/answer pairs: [EXAMPLE_RETRIEVED_QAS]
3) Assistant final answer to the user: [EXAMPLE_LLM_ANSWER]

Expected evaluation output (strict JSON object):
{
```

```
  "question_analysis": "[...]",
  "retrieved_answers_analysis": "[...]",
  "llm_answer_analysis": "[...]",
  "satisfaction_feedback_analysis": "[...]",
  "satisfaction_feedback_boolean": [true/false],
  "satisfaction_feedback_negative_reason": "[None | Irrelevant | Insufficient/Incomplete |
  ↪  Unclear | Misleading/Incorrect]"
}
```