Structured & Tiered Agentic Response Evaluation (STARe): LLM-Judge Framework for Conversational Shopping Agents

Akash Kumar Singh[1], Digvijay Singh Bhadouria[1], Flint Luu[2], Pierre Yves Vandenbussche[3], Pinchu Sabu[1], Sanket Thapliyal[1], Vikram Vishal Singh[1], Vishal Shivhare[1] and Wenjia Xie[3]

[1]eBay Inc, India
[2]eBay Inc, USA
[3]eBay Inc, Netherlands

Abstract

Conversational shopping agents built on large language models increasingly support users across the full e-commerce journey, from discovery and comparison to transaction support and post-purchase resolution. Evaluating such systems remains difficult because response quality in multi-turn interactions depends not only on factual correctness, but also on context retention, intent fulfillment, policy compliance, and communication quality. We propose STARe (Structured & Tiered Agentic Response Evaluation), a modular LLM-judge framework designed to evaluate response quality based on user input and context. Rather than collapsing quality into a single score, STARe organizes evaluation into structured decision stages that separate objective failures from intent-dependent judgments and communication weaknesses. This improves diagnostic interpretability and better reflects user experience.

Keywords

LLM-as-a-judge, conversational shopping, e-commerce, conversational agents, evaluation

1. Introduction

The long-standing dominance of keyword-based search is currently being challenged by the integration of large language models (LLMs) and the subsequent rise of natural language processing in daily interactions[1]. In e-commerce, conversational interfaces are beginning to complement or replace traditional keyword-based search by enabling users to express richer, more contextual needs in natural language.

Rather than issuing short keyword queries, users can now ask for support in forms such as *"help me find a spring wedding outfit"* or *"Hi, wass up ?"* or *"is this charger compatible with my laptop?"* This shift has created a new class of conversational shopping agents that support users across discovery, evaluation, purchase, and post-purchase journeys. By interpreting nuanced preferences and situational context, these AI agents act as digital concierges that guide the entire customer journey[2].

This shift, amplified by lack of diverse cold start datasets creates a fundamental evaluation challenge. In conversational shopping, response quality is not determined solely by local relevance or fluency. Following are the challenges which make conversational dialogue evaluation vastly complex:

1. **Response Quality Is Multidimensional, Intent-Dependent, and Shaped by Proactive AI Behavior:** A strong conversational response must adapt to user intent, context, and changing needs. For example, a user saying, "I am sad because my dog died and I don't know how to cancel my pet food order," needs both empathy and efficient task help at once. Balancing these qualities, while knowing when to proactively guide the interaction, makes evaluation difficult[**?**].
2. **Traditional Dialogue Metrics Are Inadequate:** Overlap-based metrics like BLEU and ROUGE are poorly suited to conversation, where many valid responses may look different. Likewise, task-oriented metrics (TOD) such as Inform and Success rates focus on surface outputs or final outcomes, missing hallucinations, partial failures, and mid-dialogue breakdowns[3, 4].

ECOM'26: SIGIR Workshop on eCommerce, Jul 24, 2026, Melbourne, Australia

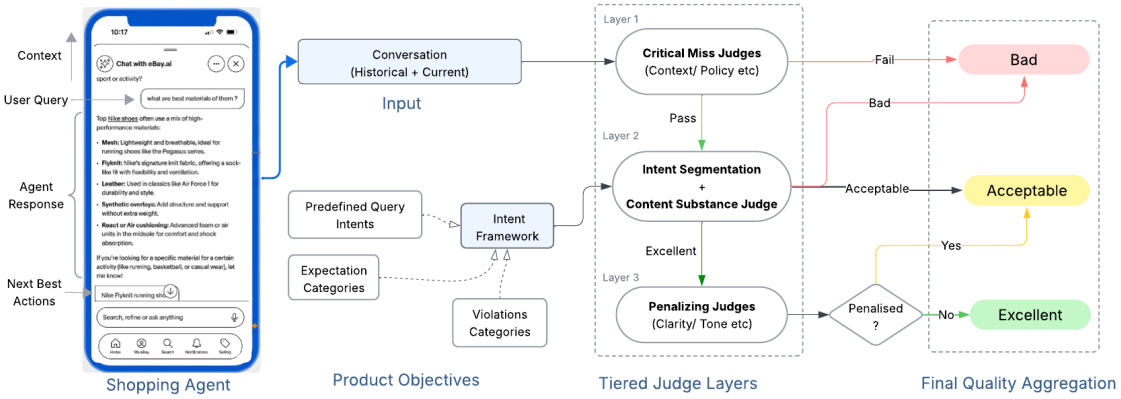

Figure 1: Components of shopping assistant along with STARe LLM Judge Framework

3. **Compounding Errors and Fragmented Evaluation:** Small errors can snowball across turns, while evaluating system components in isolation misses their combined impact on the overall user experience[**?**].
4. **Human Evaluation Has Important Limitations:** Although often treated as a gold standard, human evaluation is weakened by annotator disagreement, fatigue, bias, and difficulty maintaining consistency across long and complex dialogues. Human judges persistently disagree on subjective qualities like coherence, empathy, or helpfulness, creating noisy and inconsistent ground-truth data[**?**].
5. **Bias and Interfactor Coupling Undermine Reliability:** In monolithic LLM-as-a-judge setups, multiple evaluation dimensions and user needs are often assessed together, causing interfactor coupling where distinct subjective dimensions are inadvertently conflated. This makes scores harder to interpret and less actionable, while also increasing susceptibility to biases such as rewarding verbosity or socially appealing responses over well-grounded ones[**? ?**].

For example, eBay Conversational Search had "Resolution" as the monolithic judge of overall quality [scale 1-5] and was defined as the extent to which the response directly addresses and satisfactorily fulfills the user's query and underlying intent. However, like many monolithic judges, Resolution was not a diagnostically actionable north star: it collapsed heterogeneous intents into a single notion of success, made failure modes hard to localize, and was susceptible to verbosity bias, sometimes rewarding longer or more socially polished answers over grounded, intent-appropriate ones. For example, a small-talk exchange such as **"hey there!"**, a friendly reply that gently redirects the user toward shopping may score poorly on Resolution just because it does not explicitly complete a task.

To address some of these challenges, we propose STARe (Structured & Tiered Agentic Response Evaluation), an LLM-judge framework designed for multi-turn conversational shopping agents. The core idea is to delve into interpretable decision layers which improve objectivity, diagnostic interpretability and holistic view of overall conversation quality. STARe decomposes response quality into **three tiered layers** (Fig-1):

- **Layer 1 - Objective Critical Miss :** It first identifies critical failures through binary checks, such as context loss, intent misunderstanding, policy violations, and hallucinations.
- **Layer 2- Intent-Conditioned Quality:** It then assesses substantive response quality using an intent-conditioned judge, which evaluates the response against query-specific expectations and violations on a scale of Bad, Acceptable, and Excellent.
- **Layer 3 - Penalizing Metrics:** Third, it applies penalizing metrics to capture secondary metrics like communication-level weaknesses, weak next-step guidance, or partial intent coverage.

- **Final Aggregation:** These signals are then combined through a hierarchical aggregation rule to produce an interpretable final quality label.

We present STARe as a case-study framework for structured evaluation of conversational shopping agents, and show that its decomposed design yields useful failure localization and competitive human alignment on several rubric dimensions.

2. Related Work

2.1. Multi-Turn Dialogue Evaluation

Recent benchmarks and surveys show that evaluating multi-turn systems requires more than single-turn response scoring. Beyond surface quality, evaluation must account for task completion, memory and context retention, planning, tool use, and dialogue-level coherence[5]. TD-EVAL is especially relevant because it argues that dialogue-level success alone can mask intermediate failures and proposes turn-level analysis combined with dialogue-level comparison[4].Like TD-EVAL, STARe builds on this principle, but differs in how it operationalizes it through further decomposition into structured and tiered layers.

2.2. Reliability and Scoring Mechanics of LLM Judges

LLM judges are attractive because they scale, but recent work highlights substantial reliability issues. CheckEval shows that decomposing broad criteria into Boolean checklist items can improve robustness relative to unconstrained Likert-style evaluation. More recent work also shows that evaluator consistency depends on scoring granularity and protocol choice[6, 7, 8].

Existing work provides a strong foundation, but three gaps remain: limited separation of objective failures from intent-dependent judgments, reliance on a single holistic score that weakens diagnosis, and insufficient support for conversational shopping, where quality depends on intent, policy, safety, and journey stage. Our framework addresses these through intent-aware rubrics, modular LLM judging, and hierarchical aggregation.

3. THE FRAMEWORK - STARe

3.1. Framework Design Principles

Taking some inspiration from Edward Tufte who once said, “Overload, clutter, and confusion are not attributes of information, they are failures of design.”, following are the design principles of evaluation framework :

- **Principle 1- Clarity Through Intentional Design:** Evaluation framework should achieve clarity through deliberate architectural design, rather than by oversimplifying complex quality judgments.
- **Principle 2- Separation of Distinct Evaluation Signals:** Objective constraint violations and multidimensional quality assessments should be treated as distinct signals to improve interpretability and reliability.
- **Principle 3- Contextual and Intent-Aware Evaluation:** To reduce subjectivity, response quality should be assessed relative to user intent and expected behavioral categories, rather than through a single generalized judgment.
- **Principle 4- Diagnostic Interpretability:** Evaluation outputs should provide granular, qualitative reasoning for failure modes instead of relying only on opaque scalar scores.
- **Principle 5- Conservative Quality Aggregation:** Aggregation should ensure that superficial strengths cannot offset fundamental functional failures.

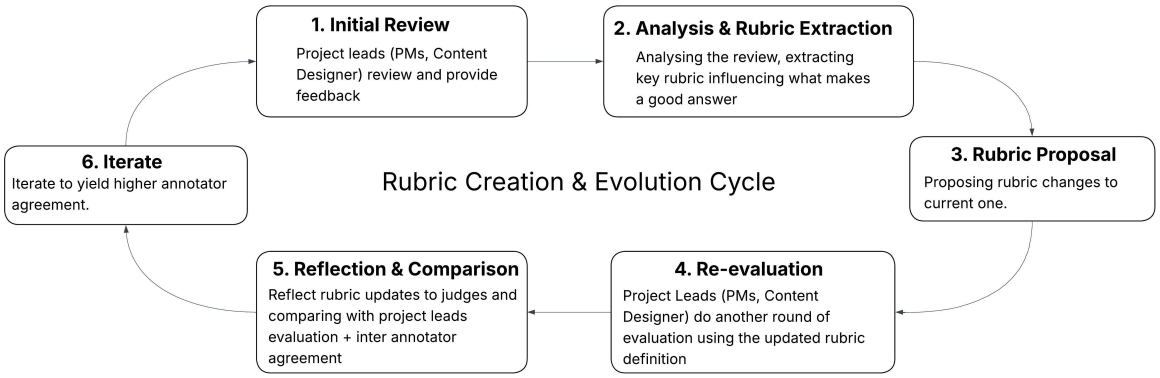

Figure 2: Iterative Rubric Optimization and Evaluation Framework

3.2. Rubric Development

The rubric was developed through an iterative process of drafting metrics based on common failures and user goals. These criteria were tested against diverse examples and refined through expert feedback to resolve ambiguities. This cycle continued until the definitions were stable enough to ensure consistent manual and automated scoring (Fig-2).

3.3. Metric Decomposition and LLM Judge Architecture

3.3.1. Layer 1- Objective Critical Miss Judge

For eBay Conversational Search use case, the pipeline begins with **Layer 1- the critical-miss judge,** which checks the current response against prior dialogue history. It detects hard failures such as context loss, intent misunderstanding, hallucinations, unsupported claims, inappropriate refusal, or policy violations. Because these failures directly affect trust, correctness, and safety, any triggered condition immediately assigns a final label of Bad. This segregation of objective critical metrics into binary labels has advantages over likert scale[**?**].

3.3.2. Layer 2 - Intent-Conditioned Quality

If the response passes the critical-miss layer, it moves to the **Layer 2 which starts with intent segmentation judge,** which predicts the user's intent from the query and dialogue context using a finite set of intent types, such as discovery, comparison, compatibility, post-purchase support, or ambiguous intent. Based on product and business objectives, the predefined intent framework consists of two global category spaces: an **expectation space** containing what a good response may need to include, and a **violation space** containing what a response should avoid. For each intent, it specifies which expectation categories are required and which violation categories are disallowed.

Table 1 provides a compact summary of the intent framework with examples. The segmentation judge acts as a routing mechanism: it maps the current query to the most appropriate theme–intent category and ensures that the next stage evaluates the response using category-specific expectations. In this way, STARe distinguishes, for example, discovery-oriented requests that require curation and reasoning from small-talk or ambiguous queries that require brief, socially appropriate, or clarifying responses. This routing step ensures that each candidate response is judged against the right notion of what "good" looks like for that interaction type.

Next component of Layer 2, the **content-substance judge** evaluates the response against the expectations and violations defined for that intent. It determines whether the response actually fulfilled the user's need. Responses can be rated Excellent, Acceptable, or Bad depending on how well they satisfy intent-specific requirements. This is the core human-centric stage, since quality is judged relative to intent rather than generic helpfulness.

Table 1: Representative examples illustrating how the intent framework conditions evaluation.

Example query	Predicted intent	Response should include	Common failure signals
"help me find a spring wedding outfit"	Use-Case / Occasion / Gift	Occasion-aware curation, style reasoning, narrowing questions or filters.	Generic list only, ignores occasion/context, weak next-step guidance.
"does this game work on PS5?"	Fit / Compatibility / Sizing	Compatibility rules, platform/version check, clarification if needed.	Compatibility guessing, omitted constraints, overconfident claim.
"where is my order?"	Orders / Returns / Disputes	Empathy, status explanation, next support step.	User blame, fabricated internal details, vague advice.
"Hi, wass up?"	Small Talk	Brief natural response, optional soft redirect to shopping.	Overlong response, forced task completion, awkward tone.
"Apple"	Ambiguous	Explicit clarification of candidate meanings.	Premature interpretation.

3.3.3. Layer 3 - Penalizing Metrics

Only responses provisionally rated Excellent move to the Layer 3- **penalizing judge**. This final stage does not reassess correctness or intent fulfillment, but examines communication quality and other secondary metrics, such as structure, tone, next best action relevance etc. These are treated as downgrade factors, so a substantively strong response may be reduced from Excellent to Acceptable, but not treated as a full failure.

The final quality score is produced through **hierarchical aggregation** across these stages. Because each stage conditions the next, the framework is not just a set of separate judges but a connected decision system. This makes evaluation more interpretable, efficient, and diagnostic, revealing whether poor performance came from contextual failure, wrong intent handling, weak substance, or poor communication.

Importantly, we intentionally constrain outputs to either binary labels or coarse ordinal categories rather than fine-grained numerical scales. This design choice reduces ambiguity and avoids artificial precision, which is known to introduce instability in human annotation settings[9]. This approach has been shown to significantly improve rating reliability, reduce variance, and enhance correlation with human judgments compared to pointwise Likert-based methods like G-Eval[**?**].

STARe defines a general layered judging pattern; the eBay deployment instantiates this pattern with domain-specific intent taxonomies and brand constraints. Because the architecture is modular, teams can refine one component without redesigning the whole system.

4. Application of Framework for eBay Conversational Search

4.1. Cold Start Data Curation

We construct a synthetic evaluation dataset designed to reflect the diversity of conversational shopping interactions encountered in practice and taking into account the expected user behavior in the next six months to one year. To generate this, we use the Multi-Dimensional Sampling methodology. The proposed methodology for synthetic dataset generation follows a structured, four-stage pipeline designed to maximize categorical diversity while maintaining high linguistic fidelity. The process begins with the definition of key dataset dimensions, establishing the primary axes of diversity required to ensure the corpus spans a representative range of semantic behaviors. dataset is stratified along 5 axes: eBay business vertical, query intent (shopping, customer support, non-shopping general knowledge, multi intent journey etc.), query formulation style (keyword-based, natural-language, hybrid queries

etc.), interaction depth (single-turn and multi-turn instances), and contextual dependency (independent queries and context-dependent follow-ups).

Subsequently, dimension-specific datasets are generated through a "seed-and-expand" strategy, where real-world queries sourced from eBay are processed via Large Language Models (LLMs) using targeted prompts to produce nuanced synthetic variants. These individual subsets are then integrated into a unified dataset, which is subjected to a sampling protocol to align the final collection with a pre-defined query distribution. Finally, the framework incorporates a hygiene phase characterized by human-in-the-loop (HITL) validation to verify data integrity and the enrichment of the corpus with granular metadata to facilitate supervised learning and benchmarking.

4.2. Gold Dataset and Human Annotation

We build a gold-standard evaluation set of about 1,000 responses sampled across the full query intent space. Using the predefined rubric, multiple annotators including SMEs, PMs, a Content Designer, and Data Scientists—labeled the dataset, and disagreements were resolved through majority voting. A subset was annotated independently by 4 annotators to measure agreement, refine the rubric, and estimate a practical upper bound for automated judge performance.

4.3. Rubric and Intent Framework

Through iterative rubric development process the following metrics were arrived at,

1. **Critical Miss Metrics**
    a. Context Understanding: Checks Repetition error, memory loss, coreference error, and context pollution.
    b. User Message Understanding: Intent misunderstanding, missed clarification for ambiguous queries, and inappropriate refusal.
    c. Policy and Safety Compliance: Checks alignment with eBay policy and safety requirements.
    d. Hallucination: Detects unsupported or fabricated claims presented as facts.
2. **Content Substance:** Evaluates whether the response includes the required components for the predicted intent.
3. **Penalizing Metrics**
    a. Intent Partially Missed: Covers only part of the user request.
    b. Tone and Voice of eBay: Assesses alignment with eBay brand tone.
    c. Structure and Clarity: Evaluates organization and comprehensibility.
    d. Next Best Actions Relevance : Checks whether the response gives clear, context-appropriate next steps.

    The intent framework was predefined and built by the product team using learnings from eBay's Jobs to Be Done, an internal rubric development process, real user query data, and competitor research and analysis. It organizes conversational shopping behavior into roughly 20 intent types, each linked to clear expectations and failure patterns, so evaluation can assess whether a response is appropriate for that specific user need rather than judging quality in a generic way.

4.4. Prompt Design and Judge Implementation

We tested multiple prompting strategies, including zero-shot, few-shot, and chain-of-thought, and selected them based on agreement with human labels. In practice, no single prompting style worked best across all metric types. Chain-of-thought was most effective for content-substance evaluation, while binary checks performed better with more constrained prompts. All judge outputs were generated in a structured schema for reliable parsing, large-scale evaluation, auditability, and downstream aggregation. Model selection was treated as a calibration problem, comparing candidate judges on human agreement, output stability, schema adherence, and cost.

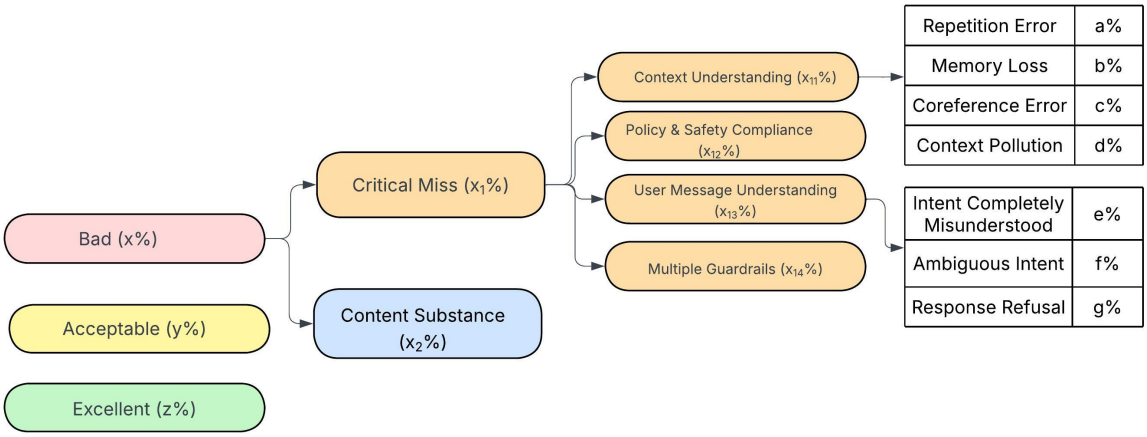

Figure 3: Example of Hierarchical Error Taxonomy and Diagnostic Breakdown

5. Results

We evaluate the effectiveness of the proposed system with its ability to provide granular diagnostic traces, by examining its alignment with human expert judgment, and its overall reliability across diverse shopping intents.

5.1. Diagnostic Interpretability

The framework supports both instance-level explainability and aggregate-level error analysis. At the single-conversation level, the judge output can identify the exact failure chain for a response; at the system level, the percentages attached to each branch allow researchers to quantify which error classes are most prevalent and therefore where model improvements should be prioritized (Fig-3).

A key advantage of this structured approach is its ability to contextualize evaluation relative to user intent, thereby avoiding misclassification of responses that are appropriate given the interaction type. For example, consider a simple conversational exchange: **"hey there!"** "Hello! I'm here to help you find great items on eBay. Let me know what kind of Jordans or other products you're interested in!" In prior monolithic evaluation setups, such responses were often penalized with low resolution scores, as they lack explicit task completion. However, under STARe, the intent segmentation layer correctly classifies this interaction as small talk. Given this intent, the response satisfies expectations of being polite, context-aware, and gently guiding the user toward shopping, and is therefore rated Excellent rather than penalized. This illustrates how intent-aware evaluation prevents false negatives arising from applying task-oriented criteria to non-task interactions.

Similarly, a broad discovery query example such as **"What are the best sounding wired earbuds?"**, a response that gives general recommendations without deeply narrowing user preferences may receive a low score under the old Resolution-based framework because it can appear vague or insufficiently specific. Under STARe, however, this is recognized as a broad discovery intent, so the response is judged **Acceptable**: it meets the baseline informational need, even if deeper narrowing or comparative reasoning would be needed for an Excellent rating.

The framework thus helps identify why a response did not reach an Excellent rating. It distinguishes whether the issue comes from content substance (i.e., the response did not fully meet intent expectations) or from penalizing metrics such as tone, structure, or weak next best action relevance. When the issue is related to content substance, we can further analyze it at the intent level to see where the gaps are. Because of the structured design of the framework, the LLM judge also explicitly lists the missing expected elements and violations that led to the final rating, making it easier to debug and improve the system.

	Human Annotators Agreement		Human + LLM Judge Agreement	
Judge	HJ Fleiss Kappa (κ)	95% CI band	HJ+LLM Fleiss Kappa	95% CI band
Context Understanding	0.93	[0.84, 0.98]	0.90	[0.82, 0.96]
Policy and Safety Compliance	0.96	[0.89, 1.00]	0.86	[0.80, 0.91]
Content substance	0.61	[0.54, 0.68]	0.61	[0.55, 0.68]
User Message Understanding	0.83	[0.77, 0.88]	0.86	[0.81, 0.92]
Segmentation	0.93	[0.87, 0.98]	0.85	[0.80, 0.91]
Tone and voice of eBay	1.00	[0.94, 1.00]	0.90	[0.84, 0.95]
Structure and Clarity	0.96	[0.92, 1.00]	0.80	[0.76, 0.84]
Intent Partially Missed	0.85	[0.78, 0.90]	0.79	[0.70, 0.86]
Next Best Action Relevance	0.68	[0.62, 0.72]	0.64	[0.59, 0.68]
Hallucination	0.97	[0.90, 1.00]	0.92	[0.84, 1.00]

Table 2
Inter-Annotator Agreement and Human–LLM Judge Alignment Across Evaluation Rubric

5.2. Judges Quality Achieved

We evaluate inter-rater reliability using **Fleiss' Kappa (κ),** measuring agreement among four human judges (SMEs) and between human judges and the LLM judge. Human–human agreement serves as a baseline against which Human–LLM agreement is evaluated.

Experimental Setup. We use a stratified sample of 500 instances, with oversampling applied to ensure representation of low-prevalence error categories (e.g., hallucination, policy violations), mitigating prevalence bias. The dataset is further segmented by intent complexity (lookup, comparison, exploratory) to assess robustness across diverse shopping scenarios. To estimate uncertainty, we compute 95% confidence intervals using bootstrap sampling.

The results show strong alignment between the LLM judge and human annotators on objective metrics such as Context Understanding ($\kappa = 0.93$), User Understanding ($\kappa = 0.83$), and Segmentation ($\kappa = 0.93$), validating the benefit of decomposing evaluation into structured components. For more subjective dimensions like Content Substance, agreement is lower ($\kappa = 0.61$) for both human–human and human–LLM comparisons, indicating that the judge operates close to the same ambiguity ceiling as human experts. This suggests that remaining variance is driven more by task subjectivity than by model limitations, and can be further improved by refining high-ambiguity intent categories.

6. Impact

The framework's main contribution is diagnostic interpretability, which is critical for AI products where non-deterministic behavior makes failures hard to trace and fix. Instead of collapsing quality into one opaque score, it identifies the specific failure path for each response and the error distribution across the system. This helps product teams pinpoint the cause of poor user experience enabling targeted improvement rather than blind tuning.

The framework offers three advantages over traditional evaluation. First, unlike reference-based metrics, it suits open-ended, multi-turn interaction because it does not assume one gold response or a single notion of correctness. Second, it improves on end-only evaluation by detecting harmful intermediate failures, such as missed constraints or misleading compatibility claims. Third, it improves over monolithic LLM-as-a-judge by reducing judge overload and separating distinct decisions, making evaluation more interpretable and more useful for error analysis while reducing signal dispersion in rubric generation[**?**].

Consistent with this, in A/B test analysis using random sample of user feedback, conversations with negative feedback showed a 1.63× higher prevalence of model-detected errors than conversations with positive feedback, suggesting that the framework captures signals meaningfully associated with real user experience. This signal should be interpreted cautiously, as the SEEK-based analysis reflects only a

limited feedback subset rather than the full traffic distribution.

As an instantiation on actual evaluation in conversational search, approximately 15% of the bad responses and 50% of the policy violations were attributed to routing failures. Furthermore, the new framework improves inter-rater agreement by more than 30% on similar instances as compared with previous monolithic judge framework.

7. Conclusion

We introduced STARe, a structured framework for evaluating conversational shopping agents by separating critical failures, intent-conditioned substance, and communication penalties instead of collapsing quality into one score. This makes evaluation more interpretable, more diagnostic, and more useful for product improvement.

In the eBay setting, STARe showed strong alignment with human judgment on objective dimensions, while keeping subjective content evaluation structured and bounded. This is especially valuable for conversational AI, where non-deterministic behavior makes user issues hard to trace and fix.

A key limitation is that STARe depends on a well-defined query intent framework, which requires significant upfront work from product and design teams. Future work includes incorporating broader search results page context into evaluation and scaling the framework for real-time, production-grade monitoring.

Acknowledgments

Core AI Team: Lamia Kasmi, Jyotsna Thapliyal, Leonid Ekimov
Design: Barbara Kofler
Product: Daniel Kaufman, Dhruv Varma
Leadership: Blair Ethington, Nitzan Mekel, Amit Sharma, Vipul Bahety

Declaration on Generative AI

During the preparation of this manuscript, the authors used ChatGPT for grammar and spelling checks and for sentence-level paraphrasing and rewording. All AI-assisted edits were reviewed and revised by the authors as needed, and the authors take full responsibility for the final content of the publication.

References

[1] Bloomreach, More than 60% of consumers have used conversational ai for shopping, 2025.
[2] J. Zeng, et al., Cite before you speak: Enhancing context-response grounding in e-commerce conversational llm-agents, arXiv preprint arXiv:2503.04830 (2025).
[3] F. Mo, C. Meng, M. Aliannejadi, J.-Y. Nie, Conversational search: From fundamentals to frontiers in the llm era, arXiv preprint arXiv:2506.10635 (2025).
[4] E. C. Acikgoz, et al., Td-eval: Revisiting task-oriented dialogue evaluation by combining turn-level precision with dialogue-level comparisons, arXiv preprint arXiv:2504.19982 (2025).
[5] S. Guan, et al., Evaluating llm-based agents for multi-turn conversations: A survey, arXiv preprint arXiv:2503.22458 (2025).
[6] Y. Lee, et al., Checkeval: A reliable llm-as-a-judge framework for evaluating text generation using checklists, arXiv preprint arXiv:2403.18771 (2025).
[7] N. Lee, J. Hong, J. Thorne, Evaluating the consistency of llm evaluators, in: Proceedings of the 31st International Conference on Computational Linguistics (COLING), 2025.
[8] N. Chen, et al., Evaluating conversational recommender systems via large language models: A user-centric framework, arXiv preprint arXiv:2501.09493 (2025).

[9] S. Mohammad, S. Kiritchenko, Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation, in: Proceedings of the Association for Computational Linguistics (ACL), 2017.
[10] A. Szymanski, et al., Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks, arXiv preprint arXiv:2410.20266 (2024).
[11] J. Wang, et al., Shoppingbench: A real-world intent-grounded shopping benchmark for llm-based agents, arXiv preprint arXiv:2508.04266 (2025).
[12] Google Blog, Ai mode in search update, 2025.
[13] A. Braggaar, et al., Evaluating task-oriented dialogue systems: A systematic review of measures, constructs and their operationalisations, arXiv preprint arXiv:2312.13871 (2024).
[14] S. Kiritchenko, S. Mohammad, Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation, in: Proceedings of the ACL, 2017.
[15] T. Nekvinda, O. Dušek, Shades of bleu, flavours of success: The case of multiwoz, in: Proceedings of the GEM Workshop, 2021.
[16] A. Szymanski, et al., Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks, in: Proceedings of the ACM Conference on Intelligent User Interfaces (IUI), 2025.

A. Appendix

A.1. Mathematical Decomposition of Judge Architecture

Let an evaluation instance at dialogue turn t be defined as:

$$x_t = (q_t, c_{<t}, r_t)$$

q_t : current user query , $c_{<t}$: conversation history till t-1 , r_t: current response

We define a structured evaluation function (judge):

$$J(x_t) \to \{M_c(x_t), M_s(x_t), M_p(x_t), y(x_t)\}$$

M_c = set of critical miss metric, M_s = content substance metric, M_p = set of penalizing metric, $y(x_t)$ = composite response quality score

A.1.1. Critical-miss metrics

$$M_c(x_t) = \{m_1^c(x_t), m_2^c(x_t), \ldots, m_K^c(x_t)\}$$

where each: $m_i^c(x_t) \in \{0, 1\}$ indicates (1) or absence (0) of a critical failure.

$$g_c(x_t) = \prod_{i=1}^{K} (1 - m_i^c(x_t))$$

A.1.2. Intent framework creation

The intent framework is defined prior to building the content judges. Let

$$I = \{i_1, i_2, \ldots, i_N\}.$$

For each intent $i \in I$, define

$$E(i) = \{e_1^i, e_2^i, \ldots, e_{n_i}^i\}, \qquad V(i) = \{v_1^i, v_2^i, \ldots, v_{m_i}^i\},$$

where $E(i)$ is the set of expected response characteristics for intent i, and $V(i)$ is the set of intent-specific substantive violations for intent i.

The full intent framework is the mapping

$$F : i \mapsto (E(i), V(i)).$$

A.1.3. Two-part content substance judge

A segmentation model assigns an intent using the current user query and prior dialogue context:

$$i_t = S(q_t, c_{<t}).$$

The content-substance judge then evaluates the current response against the expectation and violation categories associated with the predicted intent. Define the structured intermediate judge output as

$$\phi_t = G(q_t, c_{<t}, r_t; F(i_t)) = (u_t, w_t),$$

where

$$u_t : E(i_t) \to \{0,1\}, \qquad u_t(e) = 1 \iff \text{expectation } e \text{ is satisfied in } r_t,$$

and

$$w_t : V(i_t) \to \{0,1\}, \qquad w_t(v) = 1 \iff \text{violation } v \text{ is present in } r_t.$$

The content-substance score is then

$$M_s(x_t) = s(\phi_t \mid i_t) = s(u_t, w_t \mid i_t), \qquad M_s(x_t) \in \{\textit{Excellent}, \textit{Acceptable}, \textit{Bad}\}.$$

A.1.4. Penalizing metrics

Let the penalizing family contain L binary metrics:

$$M_p(x_t) = \{m_1^p(x_t), m_2^p(x_t), \ldots, m_L^p(x_t)\},$$

where each $m_i^p(x_t) \in \{0,1\}$ indicates the presence (1) or absence (0) of a non-critical deficiency.

These metrics are evaluated only if $M_s(x_t) = \textit{Excellent}$. Their aggregate pass indicator is

$$g_p(x_t) = \prod_{i=1}^{L} \left(1 - m_i^p(x_t)\right).$$

A.1.5. Composite Response Quality

Define

$$g_c(x_t) = \prod_{i=1}^{K} \left(1 - m_i^c(x_t)\right).$$

The final response quality label is

$$y(x_t) = A(g_c(x_t), M_s(x_t), g_p(x_t)),$$

with

$$\begin{aligned}
&y(x_t) = \text{Bad}, && \text{if } g_c(x_t) = 0,\\
&y(x_t) = \text{Bad}, && \text{if } g_c(x_t) = 1 \text{ and } M_s(x_t) = \text{Bad},\\
&y(x_t) = \text{Acceptable}, && \text{if } g_c(x_t) = 1 \text{ and } M_s(x_t) = \text{Acceptable},\\
&y(x_t) = \text{Acceptable}, && \text{if } g_c(x_t) = 1,\ M_s(x_t) = \text{Excellent},\ g_p(x_t) = 0,\\
&y(x_t) = \text{Excellent}, && \text{if } g_c(x_t) = 1,\ M_s(x_t) = \text{Excellent},\ g_p(x_t) = 1.
\end{aligned}$$

This decomposition yields a more interpretable and operational evaluation framework. As a result, the framework supports both more reliable LLM-based judging and more actionable diagnosis of agent failures.

A.2. Intent Framework Summary

Table 3 provides a compact summary of the intent framework used by STARe.

Table 3: Intent framework summary used by STARe.

Intent Name	Intent Definition (Routing Boundary)	Response Must Include	Must NOT Do
Discovery & Navigation			
Thematic and Broad Discovery (Inspire & Explore)	User wants ideas/trends/aesthetic/themes with no specific item OR only a vibe ("what's hot", "core", "style").	theme-decomposition; Reasoning for the recommendations, curation, Search Queries	Repetition, overwhelm, generic-recommendations, ignore-constraints
Use-Case / Occasion / Gift (Context-Driven Curation)	User goal is a project/hobby/occasion (eg. street photography, starter kit, retirement gift) where constraints are contextual.	theme-decomposition; Reasoning for the recommendations; Curation; Search Queries	Repetition, overwhelm, generic-recommendations, ignore-constraints
Specific Item Lookup (Known Target)	User names a specific model/item (often discontinued/collectible) and wants to find/verify it. Also includes filters like price, sort etc.	Search query recommendations, Reasoning for the recommendations	Repetition, overwhelm, generic-recommendations, ignore-constraint, irrelevant-substitutions, unverified-availability, fabricated-prices
Brand /Series/Collection/ Lineup Navigation	User wants to browse within a brand/collection and choose among lines/eras (not a single SKU).	Taxonomy-map, Search query recommendations, curation	Repetition, overwhelm, generic-recommendations, ignore-constraint, new-only-bias
Pre-Purchase Decision Support			
Attribute / Condition / Grading / Authenticity/ Usage Focus (Decision Support)	Query is primarily about quality state or specifications or collectible attributes (grading, "mint", first edition, sealed, authenticity).	Search query recommendations, Reasoning for the recommendations	Repetition, overwhelm, generic-recommendations, ignore-constraint, Authenticity-guarantees, counterfeit-enablement, ignore-constraints
Fit / Compatibility / Sizing (Mismatch Prevention)	User asks "will this fit/work with...?" including parts, measurements, platform compatibility, regional lock, vintage sizing.	compatibility-rules, specs-to-verify, minimal-clarifiers for mismatch prevention	compatibility-guessing, vintage-sizing-assumption
Compare / Recommend / Shortlist (Decision Support)	User is choosing between known options or a small set of listings; wants "best", "which one", "rank these".	Search query recommendations, Reasoning for the recommendations, comparison, rationale	overwhelm, ignore-constraints
Price / Value / Market Intel (Research Mode)	User asks about fair price, trends, rarity, timing, or investment-like questions.	market-factors, market-verification	fabricated-prices, financial-promises, fake-urgency
Pricing & Transactional			
Deals / Negotiation / Auction Strategy	User explicitly wants discounts, coupons, bundle deals, offer strategy, bidding tactics.	Reasoning	Repetition, overwhelm

Continued on next page

Table 3 (continued)

Intent Name	Intent Definition (Routing Boundary)	Response Must Include	Must NOT Do
Transaction-Ready (Checkout & Logistics Choices)	User is ready to purchase and asks about how to complete (shipping speed, address changes, payment steps, taxes/duties conceptually).	Reasoning	false-action-claims, fabricated-policy, overtechnical
Post-Purchase & Ownership			
Orders / Returns / Disputes (Post-Purchase Support)	Anything about an existing order: tracking, authentication delays, returns, item not as described, delivered-not-received, seller not shipped.	empathy, postpurchase-support	user-blame, fabricated-internal-details, guaranteed-outcomes, forced-selling
Ownership / Care / Repair / Parts (After You Have It)	User wants help maintaining/using an owned item (cleaning, storage, replacement parts).	care-guidance	unsafe-care-advice, unrelated-upsell
Seller & Platform Support			
Sell / List / Improve My Listing (Seller Journey)	User asks how to sell on eBay: pricing, photos, title/keywords, shipping, credibility, returns strategy.	Reasoning	overpromise-sale, deceptive-practices
Platform How-To / Account / App Troubleshooting	User needs help using features or resolving account/app issues (bidding limits, payment failure, notifications).	App-troubleshooting; Links to articles the answer is based on; Platform awareness (Is this for native or for web); Be transparent about limitations	overconfident-claims, fabricated-policy, Recommendations that don't match the platform
Policy / Restricted Items / Safety & Compliance	Questions about prohibited/restricted goods, replicas/counterfeits, weapons, endangered products, regional restrictions.	policy-guidance	policy-workarounds, illegal-or-harmful-enablement
Non-Shopping			
Non-Shopping General Knowledge	Not commerce-related but adjacent to hobbies	Commerce-bridge when shopping related, collector-context	forced-shopping-pivot
Edge Cases			
Small Talk	Casual info irrelevant to eBay	Brief-answer - Does not need to be actionable or specifically contextual.	overlong
Ambiguous	Intent unclear (Polysemy)	disambiguation	premature-interpretation
Non Sensical or unsupported language	Ill-formed query or unsupported language where intent cannot be reliably interpreted.	language-recovery, policy-guidance	premature-interpretation, illegal-or-harmful-enablement
Advanced			
Agentic Actions & Shopping Automation (Do/Manage for Me)	User asks the assistant to take or manage an in-product action (or simulate one): create alerts, manage saved searches, add to cart/watchlist, apply filters automatically, "track price," "notify me," "show my saved...". This is not advice—it's an execution/automation request.	automation-confirmation, automation-steps	false-action-claims, consentless-account-changes, spammy-alerts, sensitive-data-exposure