From Metrics to Meaning: Estimating User Feedback Using LLM-Based Evaluation

Wanqun Zhao
Amazon
wanqun@amazon.com

Adam Patterson
Amazon
adapattk@amazon.com

Cibi Chakravarthy Senthilkumar
Amazon
sentocb@amazon.com

Yulong Wang
Amazon
yulongwa@amazon.com

Anuraag Gupta
Amazon
ganuraag@amazon.com

Shahriar Sadighi
Amazon
shahrias@amazon.com

Naumaan Nayyar
Amazon
nayyarnn@amazon.com

1 Introduction

Large language models (LLMs) are increasingly deployed in real-world applications such as chatbots, writing assistants, and text summarization tools. As these applications become more central to user-facing tasks, robust evaluation of their performance becomes critical, not only for ensuring quality but also for guiding continuous improvement. Traditional evaluation approaches rely on intrinsic metrics computed from model outputs, such as groundedness and relevance (Es et al. 2025, Ru et al. 2024). While these metrics offer scalability and generalizability, their relationship to extrinsic business outcomes, such as user satisfaction, remains unclear.

User in-product feedback, such as thumbs-up/thumbs-down ratings, provides direct insight into user experience. However, this data is often sparse and potentially biased. In our production chatbot, for instance, fewer than 5% of responses receive any user feedback. Moreover, self-selection bias may occur as users are more likely to provide feedback when they have strong reactions. This creates a fundamental challenge: while intrinsic metrics are abundant but difficult to interpret practically, user feedback is meaningful but scarce.

To address this challenge, we propose applying a surrogate index framework (Athey et al. 2019) that leverages rich intrinsic metrics as surrogate variables to estimate the effect of model changes on user feedback outcomes. This framework treats user feedback as a downstream outcome and employs LLM-based evaluation metrics as intermediate surrogates, enabling us to estimate the treatment effect of model updates on user satisfaction, even when direct feedback is limited.

We validate our approach using observational data from a production chatbot. Our analysis demonstrates that surrogate index estimates align directionally with direct estimates of user satisfaction while showing more conservative magnitudes, suggesting potential mitigation of self-selection bias in voluntary feedback. Through thematic analysis of user questions, we examine heterogeneous treatment effects across question categories, providing actionable insights for targeted improvements. Additionally, we demonstrate how this framework enables prospective offline evaluation of model updates by comparing new responses against existing ones using only LLM metrics and the trained surrogate model.

Our work contributes a practical methodology for bridging intrinsic model metrics with extrinsic business outcomes, offering a scalable solution to the evaluation challenge in production AI systems. This research also enhances the interpretability of LLM-based metrics, elevating them from internal diagnostics to impact estimation tools.

2 Metrics for GenAI Chatbot Evaluation

To evaluate the performance of the GenAI chatbot, we consider three categories of metrics: 1) Intrinsic metrics, which capture the internal quality of the model's output, such as the groundedness of the chatbot's responses; 2) Extrinsic metrics, which reflect business outcomes like user feedback ratings; and 3) User engagement metrics, which assess how users interact with the chatbot in practice. In this section, we provide details on each type of metrics, and motivate the need for the research presented in the following section by highlighting how each metric captures a different facet of the chatbot performance and how aligning them can enhance the interpretability and practical relevance of intrinsic evaluations.

2.1 Intrinsic Metrics - THELMA

THELMA (Task-based Holistic Evaluation of Large Language Model Applications) is a reference-free evaluation framework for RAG (Retrieval Augmented Generation) based question answering applications (Kapoor et al. 2024). The framework comprises six interdependent metrics: Groundedness measures the degree to which responses adhere to provided sources while avoiding speculation, fabrication, or reliance on intrinsic knowledge; Source Precision and Response Precision assess the relevance of the source and response to the query, respectively; Source Query Coverage and Response Query Coverage evaluate how completely the source and response address all query components; and Response Self-Distinctness captures the non-redundancy of information within a response.

Using the above metrics, practitioners can evaluate their RAG QA applications in a fine-grained manner without requiring labelled sources or reference responses. We validated the THELMA metrics through human evaluations. During the evaluation, we presented the subject matter experts (i.e., human raters) a set of chatbot responses, and asked them to rate the chatbot responses along the six dimensions defined by THELMA. The result demonstrated reasonable alignment between THELMA's automated assessments and human ratings. We apply this framework for the intrinsic evaluation of our chatbot. For each model-generated response to user questions, THELMA generates scores between zero and one (inclusive) across all six metrics. For example, a score of one in source query coverage indicates a source containing answers to all query components whereas zero specifies a source containing no answers to the query components. We track these metrics monthly to monitor performance trends and document month-over-month changes, as shown in Table 1. However, interpreting the practical significance of these decreases remains challenging. For example, a -0.23 change in source query coverage could indicate either a minor issue with source quality or a substantial problem such as missing sources. Moreover, it is difficult to determine whether differences in magnitude (e.g., a -0.23 vs. -0.28 decrease) carry meaningful distinctions in chatbot performance. This challenge in interpretation highlights a core limitation of this type of intrinsic metrics - their lack of clear interpretability in practical contexts.

Table 1: THELMA metric reporting example

Metric	Feb	Mar	MoM Diff
source query coverage	0.71	0.48	-0.23
response query coverage	0.79	0.56	-0.20
groundedness factscore	0.79	0.57	-0.20
response precision	0.35	0.27	-0.08
source precision	0.74	0.56	-0.16
distinctness claim	0.37	0.44	0.06

2.2 Extrinsic Metrics - User Feedback

To capture the extrinsic measures of chatbot performance through business outcome, we use direct user feedback collected in-product. After each question and response, users are given an option to rate the response with a thumbs-up or thumbs-down button. This feedback mechanism is completely voluntary, so users may choose not to provide any ratings. The thumbs-up is interpreted as a positive signal of chatbot response satisfaction, and a thumbs-down as a negative one. We monitor the trends in thumbs-up and thumbs-down ratings month-over-month. Typically, only a small fraction of questions received explicit feedback - roughly 3% in both months. This highlights the scarcity of direct user feedback. In addition, interpreting user satisfaction from this feedback data requires careful consideration of its limitations. Users who choose to provide feedback may represent only a particular subset of the user population - those with particular positive or negative experiences. This introduces a potential self-selection bias and make it difficult to interpret the positive or negative feedback rate as a representative measure of overall user satisfaction or dissatisfaction.

Table 2 summarizes the volume of chatbot questions and corresponding feedback rates for February and March. For each month, we report the total number of user questions, the percentage that received feedback, and the breakdown of positive or negative feedback. As shown, only a small fraction of questions received explicit feedback - roughly 3% in both months. This highlights the scarcity of direct user feedback. In addition, interpreting user satisfaction from this feedback data requires careful consideration of its limitations. Users who choose to provide feedback may represent only a particular subset of the user population, those with particular positive or negative experiences. This introduces a potential self-selection bias and make it difficult to interpret the positive or negative feedback rate as a representative measure of overall user satisfaction or dissatisfaction.

Table 2: User feedback report example

	Feb	Mar
Question Count	1413	6616
Feedback Count	58	212
Positive Feedback Count (%)	10 (0.7%)	31 (0.5%)
Negative Feedback Count (%)	48 (3.4%)	181 (2.7%)

2.3 User Engagement Metrics - Thematic Analysis

To monitor user engagement, we analyze the types of questions users ask the chatbot. We leverage LLMs to extract thematic patterns from these questions using QualIT, a tool we developed that combines LLMs with clustering techniques. QualIT systematically addresses the limitations of traditional topic modeling methods and generates more interpretable topic representations from free-text data like the chatbot questions (Kapoor et al. 2024). The tool has so far helped us by identifying key pain points that our customers face and determining primary topics our customers discuss, enabling us to improve source coverage and general model performance.

Since January 2025, we have used QualIT to conduct monthly thematic analyses of chatbot questions and user feedback. Table **??** presents the question themes identified for March 2025. The tool has so far helped us by identifying key pain points that our customers face and determining primary topics our customers discuss, enabling us to improve source coverage and general model performance.

3 Methodology

To investigate how intrinsic metrics relate to downstream user feedback, we adopt a surrogate index modeling approach. In this framework, we treat LLM-based THELMA metrics as intermediate variables (surrogates) that mediate the effect of model improvements on user ratings. Although the model updates do not follow randomized experimentation setup, we rely on three key assumptions of surrogate index model to support causal inference. First, the surrogacy assumption posits that the treatment (i.e., model update or month indicator in our context) affects user feedback only through its impact on the surrogates. This is satisfied in our setting as user ratings are not directly influenced by the version of the model itself, but rather by how that version changes observable qualities of the chatbot responses such as response groundedness or source query coverage. Second, we assume unconfoundedness. In our setting, users are not aware of the chatbot version they're interacting with, and they don't choose into one version or another. This means the version assignment is as good as random from the user's perspective. Third, we rely on the comparability assumption, which states that the relationship between THELMA metrics and user feedback is stable across different datasets - specifically, between data where user feedback is observed (our observational dataset) and where it is missing (analogous to the experimental dataset in standard surrogate index literature). Figure 1 illustrates the framework.

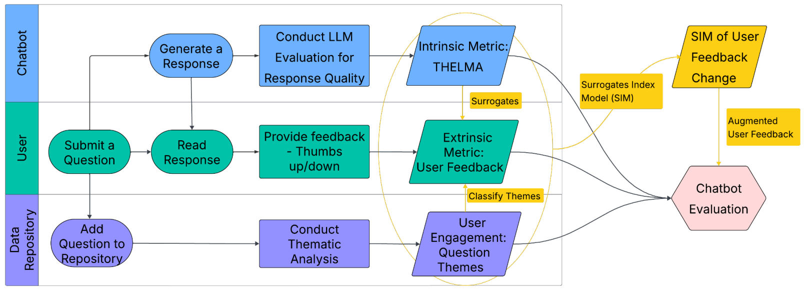

Figure 1: GenAI chatbot evaluation framework

4 Results: Impact Evaluation of Model Updates

We apply the surrogate index framework to validate the impact of a model update we recently made to the chatbot, using data from February to March. The model update includes two components: a general update to improve overall model performance, and a targetted update on Theme 1 related questions by incorporating more source articles to expand the knowledge base in handling Theme 1 related requests. Table 3 presents both the direct estimates[1] and the surrogate index estimates.

Table 3: Comparison of model update effects using different estimators

	Estimator Results					
Theme	Estimate Type	Estimate	SE	p-val	95% CI	N
Theme 1	Direct	-0.362***	0.053	0.000	[-0.466, -0.259]	200
	Surrogate Index	-0.029**	0.014	0.033	[-0.056, -0.002]	1264
All questions	Direct	0.164***	0.062	0.009	[0.041, 0.286]	134
	Surrogate Index	-0.059***	0.016	0.000	[-0.090, -0.0269]	3302

Notes: * p<0.1, ** p<0.05, *** p<0.01. Robust standard errors. Surrogate Estimators and confidence interval bootstrapped with 500 iterations.

For Theme 1 questions, both the direct estimates and the surrogate index estimates indicate a statistically significant decline in user satisfaction. The effects are negative and consistent in direction. However, an important note is that, as expected, the surrogate index estimates are much smaller in magnitude (-0.029 vs. -0.362), reflecting adjustment for the self-selection bias inherent in the direct feedback data (and hence the direct estimate). These findings suggest that the targeted improvements made to Theme 1 questions were not successful in improving user's satisfaction of the chatbot responses.

We also analyze overall model performance outcome without the thematic analysis filtering. For this analysis, due to the large volume of questions, we take a random stratified sub-sample of questions based on question themes. The result indicates that the two estimation methods diverge. The direct estimate shows a statistically significant positive change in feedback, whereas the surrogate index estimate remains negative and significant. This discrepancy again highlights the potential bias in voluntary feedback data. Because only a small fraction (less than 5%) of users left feedback, and those who did may disproportionately reflect strong or polarized experiences, direct estimate can misrepresent overall trends. The surrogate index approach, by integrating LLM-based evaluation metrics across the full population of chatbot questions, offers a more comprehensive and less biased view of the model update impact on user satisfaction rate.

To better understand the underlying factors contributing to observed decrease in user satisfaction, we also examine the direct estimates of the model update on each of the six THELMA metrics. Table 4 presents the average treatment effect for each metric. Among those, we observe statistically

[1] Regression setup for the direct estimate: $y_i = \beta T_i + X_i + \varepsilon_i$, where y_i is the user feedback, T_i indicates model update status, being 1 if the question was responded by the updated model. X_i are the covariates like user persona and company tenure. β is the direct estimate.

Table 4: Impact of model updates on THELMA metrics (surrogates)

Outcome	OLS Treatment Estimates					
	Estimate	SE	t-stat	p-value	95% CI	R^2
source query coverage	0.047	0.195	0.242	0.809	[-0.335, 0.429]	0.000
response query coverage	-0.514***	0.193	-2.669	0.008	[-0.892, -0.137]	0.018
groundedness factscore	-0.481**	0.212	-2.269	0.023	[-0.897, -0.066]	0.016
response precision	-0.506**	0.245	-2.060	0.039	[-0.987, -0.024]	0.018
source precision	0.283*	0.157	1.802	0.072	[-0.025, 0.590]	0.006
distinctness claim	0.182	0.178	1.022	0.307	[-0.167, 0.531]	0.002

Notes: * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Robust standard errors.

Table 5: Estimated changes in user feedback ratings from model V1.0 to V1.5

Theme	Estimator Results					
	Estimate Type	Estimate	SE	p-val	95% CI	N
Theme 1	Surrogate Index	0.016	0.029	0.593	[-0.041, 0.072]	236
All	Surrogate Index	0.029**	0.013	0.027	[0.003, 0.054]	1412

Notes: * $p<0.1$, ** $p<0.05$, *** $p<0.01$. Robust standard errors. Surrogate Estimators bootstrapped with 500 iterations.

significant decrease in response query coverage, groundedness factscore and response precision, and statistically significant improvement in source precision. These findings are largely consistent with the general results of decreasing user satisfaction. In addition, they suggest the directions for future model improvements.

Overall, the results presented in this section serve as a validation of our approach. Despite sparse and potentially biased voluntary user feedback, the surrogate index framework provides consistent and easy-to-interpret estimates that align with observed outcomes and intuitive expectations based on the THELMA intrinsic metrics. The directional match between the surrogate estimate and the direct estimates within the targeted theme demonstrates that the LLM-based evaluation metrics can serve as effective surrogates for user in-product feedback. Additionally, the divergence in results based on samples highlights the importance of thematic analysis, as we expect to observe heterogeneity in the treatment effect based on model updates and user interaction patterns with the chatbot.

5 Extensions: Prospective Evaluation of Model Improvement

Beyond retrospective analysis, the surrogate index framework can also be used for prospective evaluation of model improvements before deployment. Informed by the negative results from Section 4, we implemented a targeted refinement to the model's groundedness factor and improve the source retrieval process. We call this version V1.5. To assess the potential impact of this quick fix, we conduct a counterfactual analysis. Using the exact same set of questions received post model update (response generated by model V1.0), we generate new responses using the V1.5 model and compute the associated THELMA scores. We treat the V1 responses as the control group, and the V1.5 responses as the treatment group. By applying the surrogate index model that trained with historical feedback, we estimate the expected change in user satisfaction for the V1.5 improvements.

Table 5 presents the estimates from our analysis. We consider such counterfactual analysis a logical extension of this evaluation framework, as it enables us to conduct evaluation of potential model updates using only offline data. We can effectively test model improvement without the need of user exposure. It facilitates rapid iteration in model developments, and enables more informed decision-making prior to deployment. Furthermore, this framework shows potential for optimizing model development based on business outcomes.

References

Athey, Susan et al. (2019). *The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely*. Tech. rep. National Bureau of Economic Research.

Es, Shahul et al. (2025). *Ragas: Automated Evaluation of Retrieval Augmented Generation*. arXiv: 2309.15217 [cs.CL]. URL: https://arxiv.org/abs/2309.15217.

Kapoor, Satya et al. (2024). *Qualitative Insights Tool (QualIT): LLM Enhanced Topic Modeling*. arXiv: 2409.15626 [cs.IR]. URL: https://arxiv.org/abs/2409.15626.

Ru, Dongyu et al. (2024). *RAGChecker: A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation*. arXiv: 2408.08067 [cs.CL]. URL: https://arxiv.org/abs/2408.08067.