LLM Validation Framework

End-to-End Blackbox Testing from the User Perspective

Contents

1 What We Measure and Why

We measure the system **as a whole** (blackbox testing) from the **point of view of the user**. This end-to-end approach validates the complete user experience, similar to how the audit team identified issues in VoiceBot by testing from the user's perspective.

1.1 Complementary to Development Team Testing

Our validation approach is complementary to what the development team measures:

Development Team **(Component/Dev Validation)**	**Our Team** **(E2E/Blackbox Testing)**
Unit tests for individual components	Full system integration testing
Model-specific benchmarks	User-facing quality assessment
Internal API validation	External behavior validation
Performance profiling	Response quality evaluation

Table 1: Comparison of validation approaches

We expect some overlap between component/dev validation and e2e/blackbox testing. This is **normal and predictable**, and is not a bad thing—it is natural that there will be overlaps. The overlap provides additional confidence when both approaches identify the same issues, and catches different types of problems when they don't.

2 Six Dimensions of Validation

2.1 Retrieval Validation

What It Is

Retrieval validation measures whether the system successfully retrieves all relevant pieces of knowledge needed to answer a query. Poor retrieval means the LLM cannot provide accurate answers, regardless of its generation capabilities.

Primary Metric: Retrieval Recall

Measures whether the system successfully retrieves all relevant pieces of knowledge needed to answer a query.

$$\text{Retrieval Recall} = \frac{|D_{\text{retrieved}} \cap D_{\text{relevant}}|}{|D_{\text{relevant}}|} \tag{1}$$

Implementation Notes: Requires a labeled test set mapping queries to relevant documents. Compare retrieved document IDs against ground truth using exact matching or semantic similarity thresholds.

Additional Metrics

- **Precision@k** — Proportion of top-k retrieved documents that are relevant
- **Mean Reciprocal Rank (MRR)** — Average of reciprocal ranks of the first relevant result

- **Hit@k** — Binary indicator for at least one relevant document in top-k

2.2 Q&A Validation

What It Is

Q&A validation measures whether the system provides correct, complete, and relevant answers to user questions. This evaluates the core functionality of the system—its ability to understand questions and generate appropriate responses.

Primary Metrics

Context Relevancy — Whether the retrieved knowledge matches the user's query intent.

Answer Relevancy — Whether the generated answer directly addresses the user's question.

$$\text{Answer Relevancy} = \frac{1}{n}\sum_{i=1}^{n}\cos(\mathbf{e}_q, \mathbf{e}_{q'_i}) \tag{2}$$

where $\mathbf{e}_q$ is the query embedding and $\mathbf{e}_{q'_i}$ are embeddings of questions generated from the answer.

Accuracy — Whether the final answer is factually correct and aligns with business approved response. **Threshold:** > 0.95

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\nVdash[y_i = \hat{y}_i] \tag{3}$$

Implementation Notes: Context relevancy uses an LLM-as-judge to evaluate each retrieved chunk against the query. Answer relevancy generates synthetic questions from the answer and measures embedding similarity to the original query. Accuracy uses an LLM-as-judge to evaluate semantic equivalence between generated and ground truth answers.

Additional Metrics

- **Completeness** — Proportion of required information present in response
- **Facts Covered** — Proportion of ground truth facts mentioned
- **ROUGE-L** — Longest common subsequence overlap with reference
- **F1 Score** — Harmonic mean of precision and recall at token level

2.3 Tone/Politeness Validation

What It Is

Tone validation ensures the system communicates in an appropriate manner consistent with brand guidelines and user expectations. This includes formality level, empathy, professionalism, and overall communication style.

Primary Metric: Tone Compliance

Measures whether the response aligns with the expected communication style of the bank (formal, respectful, concise, non-emotional).

$$\text{Tone Compliance} = P(t = t_{\text{target}} \mid r) \tag{4}$$

where $t_{\text{target}} \in \{\text{formal}, \text{respectful}, \text{concise}, \text{non-emotional}, \ldots\}$

Implementation Notes: Define target tone attributes from brand guidelines and use an LLM-as-judge or fine-tuned classifier to evaluate each attribute. Calculate compliance as the proportion of attributes satisfied.

Additional Metrics

- **Lexical Diversity** — Vocabulary richness (Type-Token Ratio)
- **Average Sentence Length** — Readability indicator
- **Word Count Comparison** — Response verbosity vs. expected length

2.4 Risk/Guardrails Validation

What It Is

Risk validation verifies that the system's safety mechanisms are functioning properly, filtering harmful, offensive, or inappropriate content. This protects users and the organization from reputational and legal risks.

Primary Metric: Toxicity/Safety Filter

Ensures the response avoids rude, harmful, biased, or offensive content. Checks that the model blocks unsafe behaviors by enforcing policies around PII, abuse, self-harm, compliance, and restricted topics. **Threshold: 0**

$$\text{Toxicity Score} = \max_{c \in \mathcal{C}} P(c \mid r) \tag{5}$$

where $\mathcal{C} = \{\text{PII}, \text{abuse}, \text{self-harm}, \text{compliance}, \text{restricted topics}, \ldots\}$

Implementation Notes: Run responses through safety classifiers for each risk category (PII, abuse, self-harm, restricted topics). Flag as unsafe if any category exceeds threshold. Combine regex patterns, NER, and LLM-as-judge for comprehensive detection.

Additional Metrics

- **Safety Rate** — Proportion of responses that pass safety filters

2.5 Hallucination Validation

What It Is

Hallucination validation detects when the system generates information that is not grounded in the provided context or factual knowledge. Hallucinations erode user trust and can lead to serious consequences in high-stakes domains.

Primary Metric: Faithfulness/Hallucination

Whether the response stays strictly grounded in the retrieved context without adding unsupported facts.

$$\text{Faithfulness} = \frac{|C_{\text{supported}}|}{|C_{\text{total}}|} \tag{6}$$

Alternatively expressed as **Hallucination Rate**:

$$\text{Hallucination Rate} = 1 - \text{Faithfulness} \tag{7}$$

Implementation Notes: Extract factual claims from the response and verify each against the provided context. Use an LLM-as-judge or NLI model to check if the context entails each claim. Faithfulness is the proportion of supported claims.

Additional Metrics

- **Has Contradictions** — Detection of internally contradictory statements
- **Facts Covered** — Verification that stated facts match ground truth

2.6 Bad-Actor Attack Validation

What It Is

Bad-actor validation tests the system's resilience against adversarial inputs designed to manipulate, bypass safety measures, or extract unintended behaviors. This includes prompt injection, jailbreaking attempts, and other attack vectors.

Primary Metric: Adversarial Robustness

Measures how well the system resists and safely handles adversarial inputs such as prompt injection attempts, manipulation, jailbreaks, or malicious user instructions.

$$\text{Adversarial Robustness} = \frac{1}{N}\sum_{i=1}^{N} \nVdash [f(x_i) = f(x_i + \delta_i)] \tag{8}$$

where δ_i represents adversarial perturbations (prompt injections, manipulation, jailbreaks, malicious instructions).

Implementation Notes: Curate adversarial test cases across attack categories (prompt injection, jailbreaks, data extraction, manipulation). Evaluate if the system maintains safe behavior using pattern matching and LLM-as-judge. Robustness is the proportion of attacks successfully defended.

Additional Metrics

- **Attack Success Rate (ASR)** — Proportion of attacks that change model behavior

3 Summary

Dimension	**Primary Metric**	**Focus Area**	**Threshold**
Retrieval	Retrieval Recall	Finding relevant information	—
Q&A	Context Relevancy, Answer Relevancy, Accuracy	Response correctness	Accuracy > 0.95
Tone/Politeness	Tone Compliance	Communication style	—
Risk/Guardrails	Toxicity/Safety Filter	Content safety	0
Hallucination	Faithfulness/Hallucination	Factual grounding	—
Bad-Actor Attack	Adversarial Robustness	Security resilience	—

Table 2: Summary of validation dimensions and primary metrics

A Overlap Metrics: Mathematical Definitions

A.1 ROUGE-1

Unigram overlap between generated and reference text.

$$\text{ROUGE-1 Recall} = \frac{|\text{unigrams}(R) \cap \text{unigrams}(C)|}{|\text{unigrams}(R)|} \tag{9}$$

$$\text{ROUGE-1 Precision} = \frac{|\text{unigrams}(R) \cap \text{unigrams}(C)|}{|\text{unigrams}(C)|} \tag{10}$$

$$\text{ROUGE-1 F1} = \frac{2 \cdot P \cdot R}{P + R} \tag{11}$$

A.2 ROUGE-2

Bigram (two consecutive words) overlap.

$$\text{ROUGE-2 Recall} = \frac{|\text{bigrams}(R) \cap \text{bigrams}(C)|}{|\text{bigrams}(R)|} \tag{12}$$

$$\text{ROUGE-2 Precision} = \frac{|\text{bigrams}(R) \cap \text{bigrams}(C)|}{|\text{bigrams}(C)|} \tag{13}$$

A.3 ROUGE-L

Longest Common Subsequence (LCS) based measure capturing sentence-level structure.

$$R_{\text{lcs}} = \frac{\text{LCS}(R, C)}{|R|}, \quad P_{\text{lcs}} = \frac{\text{LCS}(R, C)}{|C|} \tag{14}$$

$$\text{ROUGE-L} = F_{\text{lcs}} = \frac{(1 + \beta^2) \cdot R_{\text{lcs}} \cdot P_{\text{lcs}}}{R_{\text{lcs}} + \beta^2 \cdot P_{\text{lcs}}} \tag{15}$$

where β controls relative importance of recall vs. precision (typically $\beta = 1.2$).

A.4 BLEU

Precision-based n-gram overlap with brevity penalty.

Modified n-gram precision:

$$p_n = \frac{\sum_{g \in \text{n-grams}(C)} \min(\text{count}_C(g), \max_{R_j} \text{count}_{R_j}(g))}{\sum_{g \in \text{n-grams}(C)} \text{count}_C(g)} \tag{16}$$

Brevity Penalty:

$$\text{BP} = \begin{cases} 1 & \text{if } |C| > |R| \\ \exp\left(1 - \frac{|R|}{|C|}\right) & \text{if } |C| \leq |R| \end{cases} \tag{17}$$

BLEU Score:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{18}$$

A.5 F1 Score

Token-level harmonic mean of precision and recall.

$$\text{Precision} = \frac{|T_C \cap T_R|}{|T_C|}, \quad \text{Recall} = \frac{|T_C \cap T_R|}{|T_R|} \tag{19}$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{20}$$

B Quality Metrics: Mathematical Definitions

B.1 Context Relevancy

Measures whether the retrieved knowledge matches the user's query intent.

$$\text{Context Relevancy} = \frac{1}{n} \sum_{i=1}^{n} \nVdash [\text{relevant}(c_i, q)] \tag{21}$$

where c_i is a retrieved context chunk and q is the user query. Alternatively, using semantic similarity:

$$\text{Context Relevancy} = \frac{1}{n} \sum_{i=1}^{n} \cos(\mathbf{e}_q, \mathbf{e}_{c_i}) \tag{22}$$

B.2 Answer Relevancy

Measures whether the generated answer directly addresses the user's question.

$$\text{Answer Relevancy} = \frac{1}{n} \sum_{i=1}^{n} \cos(\mathbf{e}_q, \mathbf{e}_{q'_i}) \tag{23}$$

where $\mathbf{e}_q$ is the embedding of the original query and $\mathbf{e}_{q'_i}$ are embeddings of synthetic questions generated from the answer. High similarity indicates the answer addresses the original query intent.

B.3 Faithfulness

Proportion of claims grounded in context.

$$\text{Faithfulness} = \frac{|C_{\text{supported}}|}{|C_{\text{total}}|} \tag{24}$$

A claim c is supported if $P(\text{entailment} \mid \text{context}, c) > \tau$ for threshold τ.

B.4 Completeness

Proportion of required information present.

$$\text{Completeness} = \frac{|I_{\text{covered}}|}{|I_{\text{required}}|} \tag{25}$$

B.5 Facts Covered

Proportion of ground truth facts mentioned.

$$\text{Facts Covered} = \frac{|\mathcal{F}_{\text{response}} \cap \mathcal{F}_{\text{GT}}|}{|\mathcal{F}_{\text{GT}}|} \tag{26}$$

B.6 Accuracy

Correctness against ground truth.

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \nVdash[y_i = \hat{y}_i] \tag{27}$$

For semantic equivalence:

$$\text{Semantic Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \nVdash\left[\text{sim}(y_i, \hat{y}_i) \geq \tau\right] \tag{28}$$

B.7 Toxicity

Presence of harmful content.

$$\text{Toxicity Score} = \max_{c \in \mathcal{C}} P(c \mid r) \tag{29}$$

where $\mathcal{C} = \{\text{hate}, \text{harassment}, \text{violence}, \text{self-harm}, \ldots\}$

$$\text{Safety Rate} = \frac{1}{N} \sum_{i=1}^{N} \nVdash\left[\text{Toxicity Score}_i \leq \theta\right] \tag{30}$$

B.8 Tone Compliance

Adherence to specified style.

$$\text{Tone Compliance} = P(t = t_{\text{target}} \mid r) \tag{31}$$

Using embedding similarity:

$$\text{Tone Compliance} = \cos(\mathbf{t}_{\text{target}}, \mathbf{t}_{\text{response}}) \tag{32}$$

B.9 Adversarial Robustness

Stability under adversarial perturbations.

$$\text{Adversarial Robustness} = \frac{1}{N} \sum_{i=1}^{N} \nVdash [f(x_i) = f(x_i + \delta_i)] \tag{33}$$

Attack Success Rate:

$$\text{ASR} = \frac{1}{N} \sum_{i=1}^{N} \nVdash [f(x_i + \delta_i) \neq f(x_i)] \tag{34}$$

C Retrieval Metrics: Mathematical Definitions

C.1 Retrieval Recall

Measures whether the system successfully retrieves all relevant pieces of knowledge needed to answer a query.

$$\text{Retrieval Recall} = \frac{|D_{\text{retrieved}} \cap D_{\text{relevant}}|}{|D_{\text{relevant}}|} \tag{35}$$

where $D_{\text{retrieved}}$ is the set of documents returned by the retrieval system and D_{relevant} is the set of all documents relevant to the query.

C.2 Mean Reciprocal Rank (MRR)

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \tag{36}$$

C.3 Precision@k

$$\text{Precision@}k = \frac{|\{d \in D_k : d \text{ is relevant}\}|}{k} \tag{37}$$

C.4 Recall@k

$$\text{Recall@}k = \frac{|\{d \in D_k : d \text{ is relevant}\}|}{|D_{\text{relevant}}|} \tag{38}$$

C.5 Hit@k

$$\text{Hit@}k = \nVdash [\exists\, d \in D_k : d \text{ is relevant}] \tag{39}$$

D Text Metrics: Mathematical Definitions

D.1 Average Sentence Length

$$\text{Avg Sentence Length} = \frac{\sum_{i=1}^{n} |s_i|}{n} \tag{40}$$

D.2 Lexical Diversity (Type-Token Ratio)

$$\text{TTR} = \frac{|V|}{N} \tag{41}$$

Mean Segmental TTR:

$$\text{MSTTR} = \frac{1}{k}\sum_{j=1}^{k} \text{TTR}_j \tag{42}$$

D.3 Has Contradictions

$$\text{Has Contradictions} = \nVdash [\exists\, (c_i, c_j) : \text{entails}(c_i, \neg c_j)] \tag{43}$$

$$\text{Contradiction Score} = \max_{i \neq j} P(\text{contradiction} \mid c_i, c_j) \tag{44}$$

D.4 Word Count Diff (%)

$$\text{Word Count Diff (\%)} = \frac{|T_{\text{response}}| - |T_{\text{reference}}|}{|T_{\text{reference}}|} \times 100 \tag{45}$$