AI Output Quality Evaluation Framework

Problem statement

The rapid adoption of advanced large language models (LLMs) and other generative AI systems means that many business and societal decisions now rely on machine-generated text. Unlike traditional software systems, these models produce *probabilistic* outputs and may vary across runs, so there is no single “correct” output to test against. Researchers and regulators have also shown that generative models sometimes generate hallucinations statements that are plausible but factually incorrect or fabricated, which undermines the reliability and trustworthiness of the system. Because of their stochastic nature, AI outputs can be correct but incomplete or internally inconsistent. Moreover, deploying inaccurate or unreliable AI systems increases negative AI risks and reduces trustworthiness. Hence, organizations need a structured evaluation framework that goes beyond single metrics and assesses the *quality* of AI‑generated outputs across several complementary dimensions.

The National Institute of Standards and Technology (NIST) emphasizes that developing trustworthy AI requires “reliable measurements and evaluations of underlying technologies.” The NIST AI Risk Management Framework warns that accuracy and robustness contribute to the validity of AI systems and should always be measured with clearly defined test sets and methodologies. International data‑quality standards such as ISO/IEC 25012 also recognize that data quality is multi‑dimensional and define separate characteristics for accuracy, completeness, and consistency. These observations motivate the development of a quality evaluation framework that systematically measures these dimensions for AI outputs. Such a framework helps organisations identify hallucinations, omissions and contradictions, understand trade‑offs among dimensions, and decide whether an AI system is suitable for a specific context.

Quality dimensions

Quality of AI‑generated text cannot be captured by a single measure. Drawing on data‑quality standards and AI safety literature, the following three dimensions form the foundation of the evaluation framework:

1. **Accuracy** – ***Are the statements correct?*** ISO/IEC 25012 defines accuracy as “**the degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event**”. The NIST AI RMF further notes that accuracy is the “closeness of results to the true values” and that measurements should be paired with realistic test sets and disaggregated across data segments. In the context of generative text, **accuracy** refers to **factual correctness** and **faithfulness** to source data or domain knowledge. It includes syntactic accuracy (correct format) and semantic accuracy (correct meaning).

2. **Completeness** – ***Are all required elements present****?* ISO/IEC 25012 defines completeness as the degree to which subject data associated with an entity has values for all expected attributes and related entity instances. In AI output evaluation, **completeness measures whether the response covers all aspects of the user’s prompt or task.** A model may be accurate but incomplete if it omits key constraints, parameters, or context. Measuring completeness requires a clear specification of required information for the task and may involve checklists or recall‑oriented metrics.

3. **Consistency** – ***Is the output logically coherent and stable?*** ISO/IEC 25012 defines consistency as the degree to which data are free from contradiction and coherent with other data in a specific context. For AI text, **consistency has two facets: intra‑response consistency, where the output should not contain self‑contradictory statements, and inter‑response consistency, where repeated runs with similar prompts should not produce conflicting answers unless the prompt or context changes.** Consistency also includes temporal stability and the absence of logical contradictions across related outputs.

These dimensions are complementary. A response can be accurate but incomplete, or complete but inconsistent. Evaluating all three dimensions helps organizations understand trade‑offs and make informed decisions about deploying AI systems. Other trustworthiness characteristics such as reliability and robustness are also important; NIST defines reliability as the ability of a system to perform as required without failure and notes that ongoing testing and monitoring are needed to ensure validity. However, for output‑quality evaluation, accuracy, completeness and consistency are the primary focus.

Evaluation scope

The scope of the evaluation framework encompasses textual outputs generated by AI systems, including:

- Responses from large language models (e.g., GPT‑4, Claude, Gemini) used for question answering, summarization, instruction following, classification, decision support and recommendations.

- Outputs from retrieval‑augmented generation (RAG) systems, where a model retrieves documents and generates answers based on them.

- Domain‑specific generative models used in medical, legal, financial, educational and scientific contexts.

The evaluation is model‑agnostic, meaning it does not depend on a specific AI architecture or provider. It applies to outputs in any language and across domains, but the context of use is critical: NIST notes that measurement and evaluation methods must consider the conditions of expected use and the severity of potential risks. Therefore, evaluations should be tailored to the task and domain; for example, a clinical decision support system requires stricter accuracy and completeness thresholds than a casual chatbot.

The framework evaluates individual outputs rather than entire systems, but it can be extended to assess system‑level performance by aggregating results across many outputs and tasks. It is not intended to replace broader AI risk management (which includes privacy, safety and bias), but rather to serve as a component of those efforts. Stakeholders such as developers, testers, auditors, domain experts and end‑users can use the framework to assess outputs during model development, before deployment and during ongoing monitoring.

Framework overview

The AI output quality evaluation framework follows a structured process built on the three dimensions:

1. **Task definition and expected content** – For each evaluation task, define the user prompt or query, the context, and the expected elements (e.g., facts, parameters, constraints) that the AI output should include. Clear definitions reduce subjectivity and ensure that completeness and accuracy can be measured. This aligns with NIST guidance that accuracy measurements must be paired with clearly defined test sets and methodologies.

2. **Output collection and normalization** – Generate the AI output(s) under evaluation. Where inter‑response consistency is relevant, produce multiple outputs using the same or slightly varied prompts. Normalize outputs by removing irrelevant formatting to focus on substantive content.

3. **Accuracy evaluation** – Compare the output's factual statements against trusted sources or ground‑truth data. Compute metrics such as precision, recall, and F1‑score, or use task‑specific measures like exact match for question answering. For RAG systems, evaluate the faithfulness of the answer to the retrieved documents and compute a hallucination rate (percentage of statements that are unsupported or false). NIST stresses that accuracy measurements should be accompanied by realistic test sets and may require disaggregation across data segments.

4. **Consistency evaluation** – Assess whether the output is internally coherent and whether repeated outputs contradict each other. Methods include logical contradiction detection, self‑consistency measures (agreement among multiple chains of reasoning), and semantic similarity metrics to quantify stability. For temporal stability, compare outputs over time as models are updated or as the prompt context changes.

5. **Completeness evaluation** – Determine whether the output covers all required elements defined in Step 1. Create a checklist or annotation schema to mark each required component. Calculate a coverage score (ratio of covered elements to required elements) or recall‑based metrics. Use human annotators or subject‑matter experts when domain knowledge is critical.

6. **Aggregation and reporting** – Summarise the metrics across outputs and present them in a report or dashboard. Identify trade‑offs (e.g., high accuracy but low completeness) and highlight any hallucinations or contradictions. Provide qualitative comments from human reviewers and note any systemic issues. The evaluation should inform decisions about model deployment, prompt design and risk mitigation. Continuous monitoring is advisable, as NIST recommends ongoing testing to maintain reliability.

By following this framework, organisations can systematically evaluate AI‑generated outputs across accuracy, completeness and consistency. The use of standards‑based definitions (ISO/IEC 25012 and NIST AI RMF) ensures that evaluation criteria are consistent with international best practices and facilitates communication among stakeholders. The framework provides a foundation for developing metric definition tables, scoring schemes and advanced evaluation tools in subsequent phases.