Part 2 - Metric Definition Table

Introduction

The quality-evaluation framework established in **Part 1** identified accuracy, completeness and consistency as the foundational dimensions for assessing AI-generated text. **Part 2** operationalises these dimensions by defining concrete metrics that can be computed on individual outputs or aggregated across many tasks. The table below summarises each metric using short phrases—longer explanations and sources follow the table. Where possible, metrics draw from established standards (e.g. ISO/IEC 25012 for data quality) and well-known evaluation schemes (e.g. precision, recall and F1) to ensure conceptual soundness.

Metric definition table

Metric	Quality dimension	Definition (short)	How to measure (short)	Output
Factual correctness rate	Accuracy	Share of factual statements that are correct	Count correct vs incorrect statements	Percentage
Precision	Accuracy	Ratio of correct positive statements	TP / (TP + FP)	Value in [0, 1]
Recall	Accuracy & Completeness	Ratio of correct statements vs all relevant facts	TP / (TP + FN)	Value in [0, 1]
F1-score	Accuracy	Harmonic mean of precision & recall	2 × (precision × recall) / (precision + recall)	Value in [0, 1]
Hallucination rate	Accuracy	Share of statements unsupported by sources	Count hallucinated statements / total statements	Percentage
Coverage score	Completeness	Proportion of required elements present	# covered elements / # expected elements	Value in [0, 1]
RAG completeness	Completeness	How fully a response reflects relevant context	Compare answer against retrieved context	Score or percentage

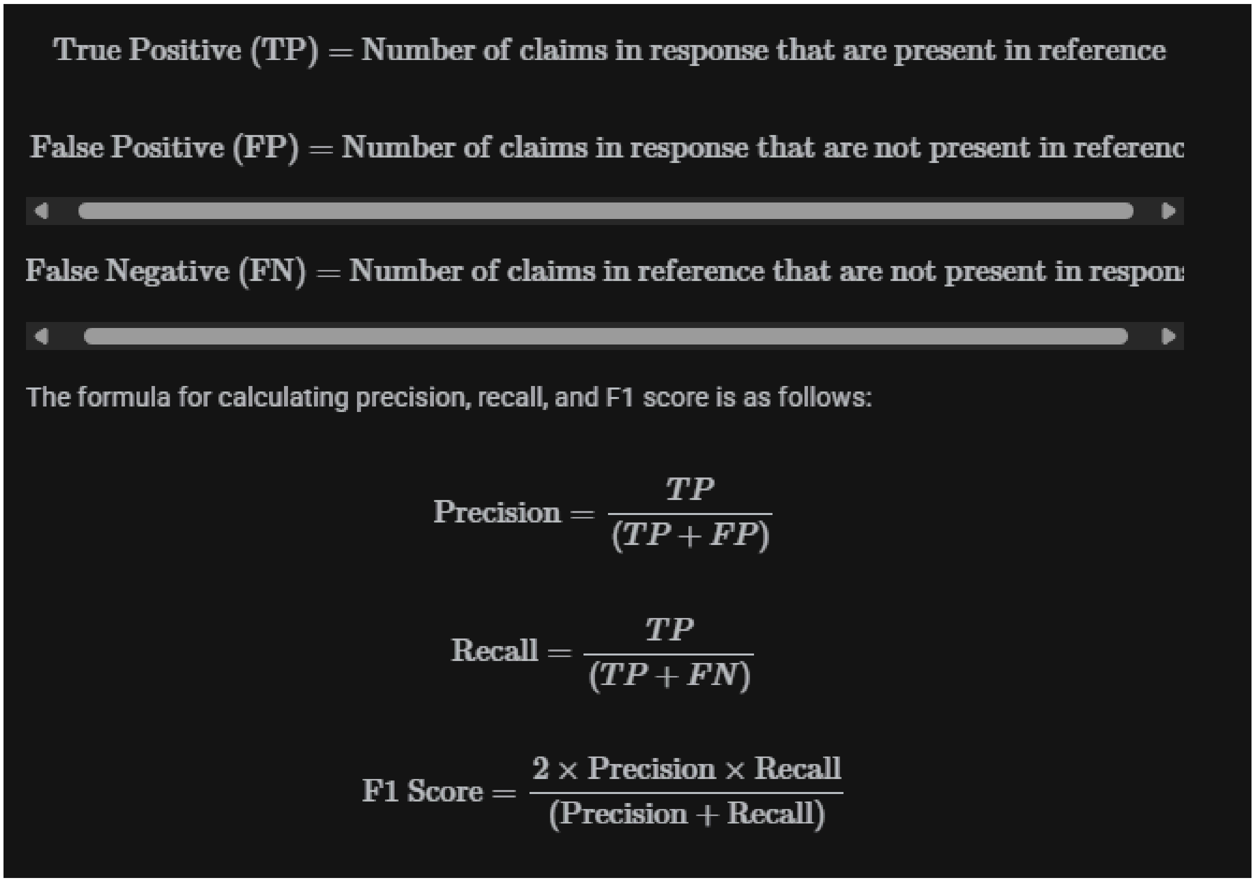
Self-consistency	Consistency	Agreement across multiple model outputs	Compare outputs for same prompt	Value in [0, 1]
Contradiction rate	Consistency	Frequency of logical contradictions	# contradictions / # statements	Percentage
Semantic stability	Consistency	Similarity of outputs across runs	Measure cosine similarity of embeddings	Score
Reasoning & conversation consistency	Consistency	Ability to maintain logical coherence across turns	Evaluate multi-turn dialogues for coherence	Rating or percentage

Metric explanations and sources

Factual correctness rate and hallucination rate (accuracy) – Accuracy is defined by ISO/IEC 25012 as the degree to which data values correctly represent the true value of the intended attribute. The NIST AI RMF describes accuracy as the closeness of results of observations or computations to true values and stresses that accuracy measurements should use clearly defined test sets and document the methodology. For AI-generated text, *factual correctness rate* counts the number of statements that can be verified as true (true positives) versus incorrect or fabricated statements (false positives) and reports the proportion of correct statements. The complementary *hallucination rate* measures the share of statements that are unsupported by the context or external sources—mitigating hallucinations is essential because fabricated content undermines reliability.

$$\text{Hallucination} = \frac{\text{Number of Contradicted Contexts}}{\text{Total Number of Contexts}}$$

Precision, recall and F1 (accuracy & completeness) – In classification tasks, **precision** is the proportion of positive classifications that are actually positive. Here, a “positive” represents a factual claim in the model output. **Recall** (also called true-positive rate) measures the proportion of actual positives that are correctly identified and is relevant when a task requires covering all relevant facts. **F1** is the harmonic mean of precision and recall and balances the trade-off between being accurate and being thorough. These metrics are computed by counting true positives (correct claims), false positives (incorrect claims), and false negatives (omitted necessary claims).



$$\text{True Positive (TP)} = \text{Number of claims in response that are present in reference}$$

$$\text{False Positive (FP)} = \text{Number of claims in response that are not present in referenc}$$

$$\text{False Negative (FN)} = \text{Number of claims in reference that are not present in respon}$$

The formula for calculating precision, recall, and F1 score is as follows:

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

$$\text{Recall} = \frac{TP}{(TP + FN)}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})}$$

Coverage score and RAG completeness (completeness) – ISO/IEC 25012 defines completeness as the degree to which subject data has values for all expected attributes. In the quality-evaluation framework, a *coverage score* is obtained by creating a checklist of required elements for the task and computing the ratio of elements covered by the response. For retrieval-augmented generation (RAG) systems, Galileo's documentation notes that *completeness* measures how thoroughly the model's response reflects the relevant information in the provided context and differentiates between context adherence (precision) and completeness (recall). Evaluators compare the answer to the retrieved context or ground truth to compute a RAG completeness score.

Self-consistency and semantic stability (consistency) – ISO/IEC 25012 defines consistency as data being free from contradiction and coherent. Giskard's glossary explains that the **self-consistency evaluation metric** compares multiple outputs generated by the AI model to assess its reliability and consistency; agreement across outputs indicates stability. *Semantic stability* extends this concept by embedding outputs (e.g. using sentence embeddings) and computing cosine similarity to quantify variation across runs.

Contradiction rate – Since consistency requires outputs to be free from internal contradictions, the *contradiction rate* counts instances where an output contains conflicting statements or contradicts previously provided information. Automated contradiction detection models or human annotators can identify contradictions, and the rate is reported as a percentage of statements.

Reasoning & conversation consistency – In multi‑turn interactions, coherence across turns is critical. A reasoning‑and‑consistency metric measures whether the AI's responses maintain logical coherence and do not provide contradictory advice across multiple turns. Evaluation involves multi‑turn testing, context management and manual audits.

Reliability and robust measurement considerations – The NIST AI RMF notes that reliability is the ability of an AI system to perform as required over time and under expected conditions. While reliability is broader than output quality, continuous testing and monitoring of metrics such as consistency and accuracy support reliable operation. Accuracy measurements should include clearly defined and realistic test sets, detailed methodology and may require disaggregation by data segment.

Using the metrics

The metrics above provide a toolkit for evaluating AI‑generated outputs under the framework introduced in **Part 1**. Organisations should select metrics appropriate to the task and domain. For example, high‑risk applications may prioritise low hallucination rates and high coverage, while conversational agents should emphasise reasoning‑and‑consistency scores. When reporting results, it is best practice to compute multiple metrics and present them together; a response may achieve high precision but low recall, and an aggregate score such as **F1** or a custom weighted sum can capture this trade‑off. Human review remains important—metrics provide quantitative indicators, but domain experts should verify that outputs are contextually appropriate and ethically acceptable.

Scoring & Evaluation Criteria

The **plan** identifies Phase 3 as converting the metrics defined in Phase 2 into numerical scores that can be compared across AI outputs. The phase requires deciding appropriate scoring scales, specifying thresholds and interpreting the scores. To implement this, each metric needs a scoring rubric that is easy to apply and yields repeatable judgments.

Principles for scoring

- **Choose appropriate scales.** Binary scales (0/1) work for metrics such as the presence or absence of contradictions. Percentage or ratio scales (0 – 1) are suitable for factual correctness rates, precision, recall, coverage and self-consistency. Likert-type scales (e.g., 1 – 5) can be used when human raters assess subjective qualities such as conversational coherence.

- **Define thresholds.** Thresholds translate continuous measures into discrete ratings. The ranges below are examples; high-risk applications may require stricter thresholds.

- **Interpretation.** Each score category should be linked to a qualitative description (e.g., “fully correct” vs “major factual errors”) so evaluators and stakeholders can understand its significance.

Scoring table

Metric	Scale	Suggested thresholds (for 5-point ratings)	Interpretation
Factual correctness rate (Accuracy)	0 – 1 ratio; discretize to 1–5	≥ 0.90 → 5; 0.80–0.89 → 4; 0.60–0.79 → 3; 0.40–0.59 → 2; < 0.40 → 1	Proportion of statements that are factually correct. High scores mean the output is almost entirely accurate; low scores indicate pervasive inaccuracies.
Precision (Accuracy)	0 – 1 ratio	≥ 0.90 → 5; 0.80–0.89 → 4; 0.60–0.79 → 3; 0.40–0.59 → 2; < 0.40 → 1	Measures the fraction of positive statements that are actually correct. High precision means few false positives.
Recall (Accuracy / Completeness)	0 – 1 ratio	≥ 0.90 → 5; 0.80–0.89 → 4; 0.60–0.79 → 3; 0.40–0.59 → 2; < 0.40 → 1	Measures how many relevant facts are retrieved. High recall means the output covers most required facts.

F1‑score (Accuracy)	0 – 1 ratio	≥ 0.90 → 5; 0.80–0.89 → 4; 0.60–0.79 → 3; 0.40–0.59 → 2; < 0.40 → 1	Harmonic mean of precision and recall. Balances correctness and coverage; useful when neither should dominate.
Hallucination rate (Accuracy)	0 – 1 ratio (lower is better)	≤ 0.05 → 5; 0.06–0.10 → 4; 0.11–0.25 → 3; 0.26–0.40 → 2; > 0.40 → 1	Proportion of unsupported statements. Low hallucination rates signal faithful answers; high rates indicate many fabricated claims.
Coverage score (Completeness)	0 – 1 ratio	≥ 0.90 → 5; 0.80–0.89 → 4; 0.60–0.79 → 3; 0.40–0.59 → 2; < 0.40 → 1	Percentage of required elements present. High scores indicate the response covers nearly all required aspects.
RAG completeness (Completeness)	0 – 1 ratio	≥ 0.90 → 5; 0.80–0.89 → 4; 0.60–0.79 → 3; 0.40–0.59 → 2; < 0.40 → 1	Measures how fully a retrieval‑augmented answer reflects relevant context. High scores mean the model used the retrieved information extensively and accurately.
Self‑consistency (Consistency)	0 – 1 ratio	≥ 0.90 → 5; 0.80–0.89 → 4; 0.60–0.79 → 3; 0.40–0.59 → 2; < 0.40 → 1	Agreement across multiple runs or reasoning chains. A high score implies deterministic behaviour; a low score suggests variability or instability.
Contradiction rate (Consistency)	0 – 1 ratio (lower is better)	≤ 0.01 → 5; 0.02–0.05 → 4; 0.06–0.15 → 3; 0.16–0.25 → 2; > 0.25 → 1	Fraction of statements that conflict with each other. High rates indicate internal inconsistency.
Semantic stability (Consistency)	Cosine similarity (0 – 1); discretize	≥ 0.95 → 5; 0.90–0.94 → 4; 0.75–0.89 → 3; 0.60–0.74 → 2; < 0.60 → 1	Measures how similar outputs are across runs using embedding similarity. High similarity reflects stable semantic content.
Reasoning & conversation consistency (Consistency)	Likert 1–5 (human‑rated)	5 = logical and coherent across turns; 4 = minor inconsistencies; 3 = some contradictions; 2 = frequent contradictions; 1 = completely incoherent	Evaluates logical coherence over multi‑turn interactions. Requires human evaluation; can also be approximated by dialogue‑coherence models.

Using the scoring criteria

1. **Calculate raw metric values.** For each output, compute the metrics defined in Phase 2 (e.g., factual correctness rate, coverage score, contradiction rate). Some metrics, such as reasoning & conversation consistency, require human raters.

2. **Apply thresholds.** Map the raw value to the corresponding score (1–5). For metrics where lower values are better (hallucination rate, contradiction rate, syntactic error rate), invert the thresholds so that 5 represents the most desirable performance.

3. **Interpret scores.** Scores of 5 indicate excellent quality (e.g., fully correct and complete, no hallucinations or contradictions), while scores of 1 signal serious quality problems. Intermediate scores denote varying levels of deficiency, with qualitative descriptions to guide reviewers.

4. **Adjust for domain risk.** High‑risk domains (medicine, legal, safety‑critical systems) may require setting higher thresholds (e.g., > 0.95 for accuracy metrics) and treating scores below 4 as unacceptable. Low‑risk applications may accept lower thresholds.

5. **Prepare for aggregation.** In later phases, these discrete scores can be combined into composite quality scores or risk maps. Maintaining consistent scales across metrics simplifies aggregation.