# Part 3 - Assumptions & Constraints

## Domain assumptions

- **Textual generative outputs.** The evaluation is scoped to text-based outputs from large language models and retrieval-augmented generation systems. Other modalities (images, audio) or purely numeric outputs (e.g., scores) are outside the current scope. Decision or classification outputs may be evaluated similarly but are not addressed here.

- **Task-specific ground truth.** For each evaluated task, we assume there is a clearly defined prompt and a checklist of required elements or facts. Without an explicit ground truth, recall and coverage cannot be computed.

- **Metrics used.** Only the metrics: factual correctness rate, precision, recall, F1-score, hallucination rate, coverage score, RAG completeness, self-consistency, contradiction rate, semantic stability, and reasoning & conversation consistency are considered.

## Human involvement assumptions

- **Human annotation and review.** Some metrics (e.g., factual correctness and contradiction rate) require human annotators or domain experts to label statements as true/false, identify missing elements, or assess coherence across turns. Reasoning & conversation consistency is typically scored by human raters because it involves subjective judgement.

- **Expertise in domain tasks.** Accurate evaluation depends on reviewers having sufficient domain knowledge to judge the correctness and completeness of outputs. In specialised fields (e.g., medical or legal), subject-matter experts must be involved.

## Automation limits

- **Partial automation.** While many metrics can be computed automatically (e.g., precision, recall, F1), automated tools for detecting hallucinations and contradictions are imperfect. These tools may miss subtle errors or misclassify correct statements. Human oversight remains essential.

- **Model stochasticity.** Large language models are probabilistic; repeated runs with identical prompts may produce different outputs. This variability can affect self-consistency and semantic stability metrics. Evaluators must decide how many runs to sample and how to aggregate results.

- **Temporal drift and model updates.** As models evolve, their outputs may change over time. Although temporal stability is not explicitly measured here, evaluators should note that scores may drift if model versions change.

## Data availability limits

- **Reference sources.** Accuracy and hallucination metrics rely on comparison to trusted sources. If reference materials are outdated, incomplete or unavailable (e.g., proprietary data), the evaluation may misclassify statements. Dependence on reference sources is a known constraint.

- **Context access.** For RAG systems, evaluators must have access to the retrieved context to compute RAG completeness. If retrieval results are not accessible, this metric cannot be calculated.

## Typical constraints and risks

According to the roadmap, common constraints include **subjectivity in language evaluation**, **dependence on reference sources**, and **model stochasticity**. Evaluators should document when subjective judgements are made (e.g., judging conversational coherence) and be cautious about the reliability of reference materials. Because the evaluation focuses solely on accuracy, completeness and consistency, it does **not** assess other trustworthiness dimensions such as fairness, bias, privacy, security or ethical impacts.

# Composite Quality Score Model

# 1. Compute dimension scores

Each dimension score aggregates the constituent metric scores (converted to a common 0–100 scale) using weights that reflect their relative importance. Because syntactic accuracy and temporal stability have been removed, they are not included.

## Accuracy dimension (A)

Metrics used: **Factual correctness rate, Precision, Recall, F1‑score, Hallucination rate** (converted to a positive score).

Suggested weights:

| Metric | Weight | Rationale |
|---|---|---|
| Factual correctness rate | 0.30 | Direct measure of correct statements; critical for trust. |
| Precision | 0.15 | Measures false‑positive control; important but overlaps with other metrics. |
| Recall | 0.15 | Reflects how many relevant facts are retrieved; essential for thoroughness. |
| F1‑score | 0.25 | Balances precision and recall; emphasised because it captures trade‑offs. |
| Hallucination rate (inverted) | 0.15 | Penalises unsupported statements; overlaps with factual correctness but highlights hallucinations. |

### Accuracy score (A) = 0.30×(Correctness score) + 0.15×(Precision score) + 0.15×(Recall score) + 0.25×(F1 score) + 0.15×(1 − Hallucination rate)

Scores for each metric should be normalised to 0–1 (or 0–100) according to Phase 3 thresholds before applying these weights.

## Completeness dimension (C)

Metrics used: **Coverage score, Recall (also counted in accuracy but contributes to completeness), RAG completeness**.

Suggested weights:

| Metric | Weight | Rationale |
|---|---|---|
| Coverage score | 0.40 | Direct measure of how thoroughly required elements are covered. |
| Recall | 0.40 | Reflects proportion of relevant facts retrieved; equally important for completeness. |
| RAG completeness | 0.20 | Only applicable to retrieval-augmented systems; captures how fully the model uses the retrieved context. |

For tasks without retrieval, redistribute the RAG completeness weight proportionally to coverage and recall.

### Completeness score (C) = 0.40×(Coverage score) + 0.40×(Recall score) + 0.20×(RAG completeness score)

## Consistency dimension (K)

Metrics used: **Self-consistency, Contradiction rate (inverted), Semantic stability, Reasoning & conversation consistency**.

Suggested weights:

| Metric | Weight | Rationale |
|---|---|---|
| Self-consistency | 0.30 | Captures stability across runs; central to consistency. |
| Contradiction rate (inverted) | 0.30 | Penalises internal contradictions. |
| Semantic stability | 0.20 | Measures semantic variation; less critical than overt contradictions but still important. |
| Reasoning & conversation consistency | 0.20 | Addresses multi-turn coherence; particularly relevant for chatbots. |

**Consistency score (K)** = 0.30×(Self-consistency score) + 0.30×(1 − Contradiction rate) + 0.20×(Semantic stability score) + 0.20×(Conversation consistency score)

## 2. Combine dimension scores into a composite score

Apply the recommended weights from the roadmap—**0.4 for accuracy, 0.3 for consistency and 0.3 for completeness**. Let **A**, **K** and **C** denote the dimension scores defined above.

$$\text{Composite Quality Score } Q = 0.4\,A + 0.3\,K + 0.3\,C$$

Where each dimension score ranges from 0 to 1 (0–100 if using percentages). Higher values indicate better overall quality. This weighting emphasises factual correctness while still giving substantial weight to completeness and consistency. The weights can be adjusted for domain-specific contexts—for instance, safety-critical applications may set 0.50.50.5 or more weight on accuracy and require minimum thresholds on each dimension.

## 3. Interpretation and use

- **Score ranges.** Interpret the composite score according to defined bands (e.g., 90–100 = excellent, 70–89 = good, 50–69 = fair, <50 = poor). These bands should align with the Phase 3 thresholds to maintain consistency across phases.

- **Flexibility.** The composite model is configurable: if a particular metric is not applicable (e.g., RAG completeness for non-RAG tasks), its weight can be redistributed among remaining metrics or the dimension.

- **Transparency.** Provide dimension and metric-level scores alongside the composite score so stakeholders can see where quality deficits arise. This transparency helps target improvements and mitigations.

# Evaluation Workflow

Phase 6: Evaluation Workflow

*Illustrates the pipeline for assessing AI outputs: compute metrics, assign scores, aggregate results and report findings.*