# Toward a Multi-Axis Reliability Benchmark

## 0.1 Toward a *Multi-Axis* Reliability Benchmark

Reliability is multi-faceted: a completion can be lexically fluent yet factually wrong, logically inconsistent, or ethically unsafe. Table 1 groups the most widely used automatic metrics by the failure mode they diagnose and serves as the backbone for our composite benchmark (Section **??**). Below we summarise each family, highlight its blind spots, and supply canonical references.

**Lexical-semantic overlap.** **BLEU papineni-etal-2002-bleu** counts $n$-gram co-occurrences between candidate and reference text and remains the de-facto baseline in MT and summarization. **BERTScore zhang-etal-2020-bertscore** replaces surface $n$-grams with contextual token embeddings from BERT, improving correlation with human adequacy judgments while retaining the reference-based framing. Both metrics assume the reference contains the *only* valid wording, an assumption that breaks down for creative or long-form generation, and they are agnostic to factuality.

**Factual soundness (hallucination focus).** **FactScore min-etal-2023-factscore** decomposes a generated passage into atomic claims and computes the proportion supported by an external knowledge source, achieving $<2\%$ error vs. human annotation on long-form biographies. Because it inspects claim-level entailment, FactScore directly targets hallucination but says nothing about internal coherence or style compliance.

**Logical consistency.** **PARAREL elazar-etal-2021-pararel** probes whether a model answers identical factual queries expressed via 328 paraphrastic patterns for 38 relations; inconsistency reveals brittle internal representations even when each individual answer is factual. Other recent work introduces transitivity and commutativity stress-tests, but PARAREL remains

Table 1: Taxonomy of automatic metrics used in our composite benchmark.

| **Axis** | **Representative metric / dataset** | **Original domain** |
|---|---|---|
| Lexical overlap | BLEU (**papineni-etal-2002-bleu**); BERTScore (**zhang-etal-2020-bertscore**) | MT, summarization |
| Factuality (hallucination) | FactScore (EMNLP 2023) **min-etal-2023-factscore** | Long-form bios |
| Logical consistency | PARAREL (TACL 2021) **elazar-etal-2021-pararel** | Cloze queries |
| Ethical compliance | RealToxicityPrompts (EMNLP 2020) **gehman-etal-2020-realtoxicity** | Toxicity robustness |
| Format adherence | Draft 2020-12 JSON Schema **jsonschema-2020-12** | Code / structured I/O |

the most widely adopted, so we embed its pass-rate as our logic axis.

**Ethical compliance.** **RealToxicityPrompts gehman-etal-2020-realtoxicity** contains 100 k naturally occurring sentence openings of varying toxicity; completions are scored with Perspective API, enabling fine-grained measurement of toxic degeneration. We treat the mean toxicity score over completions as our ethics axis; the prompts can be stratified by baseline toxicity to expose differential robustness.

**Format adherence.** Structured tasks (code generation, JSON APIs) fail if the output violates a target grammar. We implement a *schema-validator axis* based on the draft 2020-12 JSON Schema specification **jsonschema-2020-12**. Any completion that fails `ajv` validation receives zero on this axis, penalising even single-character deviations.

**Design choice for our benchmark.** Each axis yields a normalized subscore in $[0, 1]$; the composite is a weighted geometric mean that penalises a model dropping to zero on *any* dimension. This "no weak links" aggregation ensures progress cannot hide behind one strong metric while neglecting another. Detailed formulas and ablation studies appear in Section **??**. All code, prompts, and schemas will be released for reproducibility.

Takeaway

Single-axis scores mask critical weaknesses. Our multi-axis framework (lexical, factual, logical, ethical, and structural) offers a minimum viable checklist for certifying LLM reliability in scientific and applied settings.