Output Classification Method

Context

The study tests whether different AI models can recognize as false a set of statements that are deliberately incorrect. The model makes an error when it judges the statement as correct, and it does not make an error when it judges the statement as incorrect.

Interpretive model

Red (wrong): the model fully endorses the incorrect statement, labeling it with strong, unambiguous, and conclusive terms such as "correct", "true", "right", "perfect", or "approved". Strong expressions with secondary qualifications (qualifications in parentheses) that indicate a broad range of acceptability, such as "correct (mostly)" or "true (in the majority of contexts)", fall into this category. Expressions that suggest universal correctness in use, such as "correct in practice", also fall into this category.

Orange (misleading): the model largely endorses the incorrect statement, using less conclusive but still moderately favorable labels such as "mostly correct", "generally correct", or "often true". This category also includes expressions that confine any context limitations to parentheses - such as "correct (partially)" - or that suggest usefulness or moderate but not universal acceptance - such as "convenient", "conventional", "conventionally accepted", or "true in most occasions".

Yellow (ambiguous): the model labels the incorrect statement with terms such as "mostly incorrect" or "incorrect (partially)". This category also includes "partially true", "ambiguous", "imprecise", and all phrasings that explicitly indicate a non-ideal but unclear status. This category remains agnostic about correctness: it simply flags outputs that require a more in-depth lexical analysis.

Green (correct): the model correctly rejects the statement, labeling it with terms such as "incorrect", "false", "wrong", "inaccurate", "misleading", or "disapproved".

Useful heuristics

1. If the phrase starts with a strong approval term - e.g., "correct" - and then adds a qualifier that frames the statement as valid in use - e.g., “in practice” or “in real-world” - then treat it as red.

2. If the phrase starts with a strong approval term - e.g., "correct" - and then includes a parenthetical qualifier indicating a generally favorable or above-average condition, then treat it as red.

3. If the phrase contains a strong approval term - e.g., "correct" - and a **non**-parenthetical qualifier indicating a generally favorable or above-average condition, then treat it as orange.

4. If the phrase contains a strong approval term - e.g., "correct" - and then a parenthetical qualifier that weakens its universality - e.g., "with caution" - then treat it as orange.

5. If the phrase indicates, overall, a below-average level of approval - e.g., "partially correct" - or explicitly states any disapproval condition - e.g., "partially incorrect" - then treat it as yellow.

6. If the phrase explicitly states any negative condition - e.g., not quite correct - then treat it as yellow.

Some references and examples

RED	ORANGE	YELLOW	GREEN
Correct	Mostly correct	Mostly incorrect	Incorrect
True	Generally correct	Incorrect (partially)	False
Right	Often true	Partially true	Wrong
Perfect	Correct (partially)	Ambiguous	Inaccurate
Approved	Convenient	Imprecise	Invalid
Accurate	Conventional	Partially correct, but misleading	Misleading
Valid	Conventionally accepted	Partially correct, but not always preferred	Disapproved
Correct (mostly)	True in most occasions	Partially correct but imprecise	Misconception
True (in the majority of contexts)	Essentially correct (with an important caveat)	Not quite correct	Inappropriate
Correct in practice	Acceptable	Not entirely correct	
Correct (approximately)	Correct (with caution)		
Correct (general recommendation)	Correct (with nuance)		
Correct (common guideline)	Correct (with clarification)		
	Correct (with a caveat)		
	Largely accepted		

Inter-rater reliability assessment

Inter-rater reliability was assessed in three complementary settings. A binary “direction” endpoint was prespecified as the primary summary, operationalized as endorsement versus rejection (RED + ORANGE vs YELLOW + GREEN), while the four-level scheme (RED, ORANGE, YELLOW, GREEN) was retained for more granular description.

First, under the prespecified two-level direction coding with three raters, estimated reliability was high and consistent across metrics (Fleiss’ kappa = 0.723; Krippendorff’s alpha = 0.726). When agreement was evaluated at the level of individual scale items (distinct lexical expressions) using the four-level coding, estimated reliability was moderate according to Fleiss’ kappa (0.467) and Krippendorff’s alpha (0.419).

Second, following a revision of the mapping table, we conducted a targeted stress test on 11 new items using the two reviewers who had previously shown the greatest disagreement. Under the two-level direction coding, agreement was perfect (100%). Under four-level coding, Cohen’s kappa was 0.532 (unweighted) and 0.864 (squared-weighted), consistent with the interpretation that the remaining discrepancies were primarily small ordinal shifts rather than extreme category reversals.

Third, prior to the revision of the mapping table, we conducted a separate experiment in which 75 AI-generated outputs were independently coded by two raters. Agreement was high under the four-level scheme (Cohen’s kappa = 0.824, null $p < 0.001$, unweighted; Cohen’s kappa = 0.978, null $p < 0.001$, squared-weighted) and perfect under the prespecified two-level direction coding (100% agreement). The high level of agreement plausibly reflects the fact that the distribution of outputs was largely dominated by response labels corresponding to clear-cut endorsement or rejection, for which rater consensus is straightforward.