Absence of Evidence Is Not Evidence of Absence: Text-Decidability in Automated Open-Science Assessment

Keywords: automated assessment; text-based measurement; measurement validity; open science; reporting bias; computational social science

1.1 Problem

Automated assessment of open-science and reporting standards is increasingly used to describe scientific practice across journals, funders, and fields. These systems often score article text as if it directly measured compliance: found criterion are treated as met; missing criteria as unmet. Studies find that "data available on request" often yields no data when followed up [1,2], yet automated indicators are easily read as compliance measures rather than reported-compliance measures [3]. This conflation is too blunt.

The problem is not only model error. Some criteria are directly decidable from article text, such as whether a sample size is reported. Others are decidable only when the relevant evidence is present or linked, such as allocation concealment. Still others are verification-limited: article text may claim that data availability, protocol existence, or no conflicts, but text alone cannot verify them. Automated assessment can therefore undercount compliance when evidence is expressed non-standardly and overcount it when models fill gaps from background knowledge or treat reported claims as verified conduct.

1.2 Gap and Claim

This problem relates to reporting bias and LLM abstention [4–7], but neither framing fully captures the measurement issue. Reporting-bias research shows that articles are incomplete records of scientific conduct. Abstention research asks whether models should withhold an answer when uncertain. Our claim is narrower: automated criterion-level assessment must preserve the evidential status of each judgment.

Existing automated checklist systems, including CONSORT-TM [8], SPIRIT-CONSORT-TM [9], and recent LLM-on-CONSORT work [3], primarily evaluate whether systems recover checklist-relevant information from text. They do not model whether article text can validly support the scored inference. A model may fail because it missed available evidence, or because it was asked to decide something that article text cannot settle. These failures should not be merged into binary accuracy.

1.3 Construct and Contribution

We introduce criterion text-decidability: whether available article text provides sufficient evidential warrant to score a criterion for a given document. We distinguish three evidential states: (1) text-supported; (2) externally checkable, where the article points

to a source that could settle the criterion; and (3) text-non-decidable. Not applicable is treated separately as an exclusion state.
This taxonomy separates evidential status from substantive judgment. The contribution is a measurement framework estimating reported compliance, text-supported non-compliance, externally checkable cases, and text-non-decidable cases.

1.4 Data, Method, and Pilot

We use SPIRIT-CONSORT-TM [9] as a testbed (200 RCTs, 83 checklist items, full PMC texts) because it contains simple reporting, conditional methodological, and verification-limited open-science items. We supplement the checklist labels with expert-coded evidential-status annotations for criterion–document pairs: evidence in article text, evidence only through article-provided external sources, or not decidable from available text (target $\kappa \geq .70$).

We then evaluate a section-aware retrieval and structured LLM judging pipeline. The model returns both a substantive judgment and an evidential-status judgment, to distinguish supported judgments, missed textual evidence, unsupported inference from insufficient evidence, and appropriate abstention.

A 1,170-judgment pilot across 30 protocol–results pairs, 60 full-text papers, and 20 checklist items suggests that decidability tier predicts error type, not just rate ($\chi^2(10)=489$, $p<.001$, Cramér's $V=0.46$). Conditional items produced more "unclear" classifications (51%) than always-decidable items (31%, Fisher's exact $p < .001$). Always-decidable items more often produced confident false negatives (31% of present cases). Verification-limited items showed a different pattern: rarely reported criteria produced few false positives (5%), while normatively expected criteria such as financial interests and data access produced more unsupported positives (13%), consistent with models filling gaps from background expectations.

1.5 Implication

Automated open-science assessment should not report a single compliance score when the evidence base contains different measurement states. This paper contributes to CSS and meta-science by treating automated assessment as a measurement problem, not just NLP classification. Absence of textual evidence is not evidence of absence.

References

[1] Colavizza et al. 2020. PLOS ONE 15(4). [2] Wicherts et al. 2006. [3] He et al. 2025. arXiv:2511.13107. [4] Rajpurkar et al. 2018. [5] Kirichenko et al. 2025. arXiv:2506.09038. [6] Fu et al. 2025. arXiv:2506.11440. [7] Abdaljalil et al. 2026. arXiv:2602.14189. [8] Kilicoglu et al. 2021. J. Biomed. Informatics 116:103717. [9] Jiang et al. 2025. Scientific Data 12:355.