From Claims to Compliance: Evaluating Verifiability in AI System Disclosures via Policy-Grounded Analysis

Chandra Mahule [1 2]

Abstract

Model cards and evaluation reports inform AI governance, but the verifiability of their claims remains unclear. We analyze 412 safety, evaluation, and alignment claims from 68 Hugging Face model cards and 24 HELM reports using an OECD-aligned three-point scale. Only 38.1% of claims are well-supported and 20.1% unsubstantiated, with safety claims least verifiable (26.3% well-supported, 28.1% unsubstantiated). Bootstrapped intervals (33.5%–42.8%) reveal persistent gaps between disclosure and verifiability, motivating policy-aligned auditing standards.

1. Introduction and Background

AI governance relies increasingly on documentation artifacts—model cards, system reports, and benchmark evaluations—to inform regulatory and policy decisions (Mitchell et al., 2019; Holstein et al., 2019; Hendrycks et al., 2020). Regulatory frameworks such as the EU AI Act increasingly depend on such documentation for conformity assessment and auditing, implicitly assuming that claims made in these artifacts are independently verifiable (Cantero Gamito & Marsden, 2024). These documents contain statements on robustness, fairness, safety mitigations, and evaluation rigor. A central, underexplored question is whether such statements are *verifiable*: can an independent party confirm their validity from the disclosed information alone? While transparency mechanisms and benchmarking efforts have advanced (Hendrycks et al., 2020; Bender et al., 2021), transparency does not ensure independent confirmation (Pineau et al., 2021; Semmelrock et al., 2025). We define verifiability as distinct from transparency: transparency concerns information availability, whereas verifiability concerns whether that information suffices to independently confirm claims. Despite its importance for auditability and AI assurance, verifiability has rarely been operationalized. Reproducibility studies in machine learning consistently find that experimental details are insufficient to confirm reported results (Pineau et al., 2021; Semmelrock et al., 2025). We introduce a framework that extracts claims and scores them based on reproducible evidence, grounded in OECD AI Principles (AI, 2023) and aligned with emerging regulatory expectations such as the EU AI Act (Cantero Gamito & Marsden, 2024). Our framework is designed to support third-party auditors and regulators by providing a concrete protocol to assess whether documentation meets minimum standards for independent verification. Applying this framework to 92 widely used documents reveals substantial gaps, particularly for safety claims, indicating that current documentation practices may fall short of emerging auditability standards. Our analysis focuses on two influential documentation ecosystems: developer-authored model cards hosted on Hugging Face and independent evaluation reports from HELM. These sources were selected because they are widely referenced in AI governance, benchmarking, and auditing contexts. While they are not intended to be statistically representative of all AI system documentation, they provide a useful basis for examining current documentation practices across both developer-produced and third-party evaluation artifacts. By linking technical evaluation with policy-grounded criteria, our work provides empirical insight and actionable guidance for AI governance, including informing audit criteria, compliance checks, and documentation standards.

2. Methodology

We analyze the verifiability of AI system claims using publicly available documentation from two complementary sources. We randomly sampled 68 model cards from the Hugging Face Hub (Osborne et al., 2024) (top 200 most-downloaded as of January 2026), spanning text, vision, and audio models, and included all 24 evaluation reports from the Stanford HELM benchmark suite (Hendrycks et al., 2020), covering 30 prominent language models. We also reference underlying models including LLaMA-2-7B (Touvron et al., 2023). Together, these 92 documents provide a

[1]University of Oxford [2]Pontifical Catholic University of Chile. Correspondence to: Chandra Mahule <chandramahu18@gmail.com>.

Second Workshop on Technical AI Governance Research (TAIGR) @ ICML 2026, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

modest but diverse corpus for exploratory, policy-relevant analysis. Hugging Face model cards were selected because they constitute one of the largest publicly accessible repositories of AI system documentation and include models from academic, open-weight, and commercial development contexts. HELM reports were included because they provide a contrasting form of independently generated evaluation documentation. Together, these sources allow comparison between developer-provided and benchmark generated disclosures. Claims were extracted using a two-stage process. First, sentences containing keywords such as `evaluated on`, `safe`, `robust`, `fair`, `mitigate`, `bias`, and `alignment` were identified as candidates. Second, a zero-shot LLaMA-2-7B model filtered false positives using the prompt: *"Does the following sentence contain a specific, testable claim about an AI system's safety, evaluation, or alignment? Answer only YES or NO."* Only sentences with confidence above 0.7 were retained. This model was used solely for automated claim selection; all verifiability scoring was performed manually based on predefined criteria. Validation on 100 sentences manually labeled by the author yielded a precision of 0.84 and recall of 0.79, indicating reliable identification of claims for downstream analysis. This process resulted in 412 unique claims (258 from model cards, 154 from HELM reports). Following claim extraction, each claim was manually annotated by the author. Annotation consisted of reviewing the claim together with all supporting evidence available in the source document and any explicitly referenced artifacts (e.g., datasets, evaluation reports, or code repositories). The author assessed the claim against the three rubric dimensions—Evidence Presence, Reproducibility, and Clarity of Limitations—and assigned a verifiability score of 0 (Unsubstantiated), 1 (Partially supported), or 2 (Well-supported). Only information contained in the source document or directly referenced artifacts was considered; external information was not used during annotation. Annotation was completed prior to aggregate statistical analysis, and scores were recorded independently for each claim based solely on evidence available in the source document and any explicitly referenced artifacts. Duplicate claims across documents were removed so that each entry represents a distinct, testable assertion. Among these, 114 safety claims (66 from model cards, 48 from HELM), 228 evaluation claims (142 from model cards, 86 from HELM), and 70 alignment claims (50 from model cards, 20 from HELM) were identified. We define a claim as a declarative statement asserting a property of an AI system that is, in principle, empirically testable. Claims lacking supporting metrics, dataset identifiers, or methodological details were classified as unsubstantiated (Score 0), reflecting gaps in documentation relevant to auditability. Verifiability was operationalized along three policy-grounded dimensions derived through qualitative thematic analysis of OECD AI Principles and 15 national AI strategy documents in the OECD AI Policy Observatory (AI, 2023). The analysis identified recurring audit requirements, consolidated into: (i) **Evidence Presence**, whether metrics, experimental details, or code links are provided; (ii) **Reproducibility**, whether sufficient information is available to replicate claims, including dataset identifiers, evaluation protocols, and model versions; and (iii) **Clarity of Limitations**, whether scope and known caveats are explicitly stated. This mapping ensures alignment between the scoring rubric and policy-defined expectations for auditability rather than ad hoc design. Claims were categorized into three policy-relevant types based on primary focus: **Safety**, referring to harm prevention, robustness, bias mitigation, or risk reduction; **Evaluation**, referring to empirical performance and benchmarking results; and **Alignment**, referring to alignment with human values, intent, or behavioral constraints (including human feedback methods). Each claim was assigned a single dominant category. When a claim could plausibly belong to multiple categories, it was assigned to the category most central to the primary assertion being made. For example, a claim describing bias mitigation through alignment training would be categorized as Safety if the primary emphasis was risk reduction rather than alignment methodology. Each claim was scored on a three-point scale: **0 (Unsubstantiated)**, **1 (Partially supported)**, or **2 (Well-supported)**. Scores were assigned using a rule-based rubric over the three dimensions. Score 2 was assigned when all three dimensions were strongly satisfied, Score 1 when at least one dimension was satisfied but not all, and Score 0 when most or all dimensions were not satisfied. Scoring criteria were predefined using OECD-aligned auditability guidelines and applied consistently across all claims. Scores were aggregated by source and category. To assess robustness, we computed bootstrapped 1,000-resample confidence intervals for the proportion of well-supported claims. An ablation study re-scored 50 claims under stricter reproducibility criteria requiring explicit open-source code or complete hyperparameter specifications, evaluating sensitivity to higher audit standards. These stricter criteria were used only for sensitivity analysis and not in the main results.

2.1. Annotation and Verifiability Scoring Procedure

Following automated claim extraction, all 412 claims were reviewed and scored by the author using the predefined verifiability rubric comprising Evidence Presence, Reproducibility, and Clarity of Limitations. The LLaMA-2-7B model was used solely for identifying candidate claims and filtering false positives and played no role in assigning verifiability scores. Each claim was evaluated using only information contained within the source document and any explicitly referenced artifacts, such as linked datasets, evaluation reports, or code repositories. External information about the model, developer, or organization was not considered when

Table 1. Verifiability scores by source and claim category (N=412 claims). Safety claims are markedly less supported than evaluation claims. (Score 0 = Unsubstantiated, 1 = Partially supported, 2 = Well-supported)

	Score 0	**Score 1**	**Score 2**	**Total**
By Source				
Model Cards	58 (22.5%)	108 (41.9%)	92 (35.7%)	258
HELM Reports	25 (16.2%)	64 (41.6%)	65 (42.2%)	154
By Category				
Safety	32 (28.1%)	52 (45.6%)	30 (26.3%)	114
Evaluation	35 (15.4%)	89 (39.0%)	104 (45.6%)	228
Alignment	16 (22.9%)	31 (44.3%)	23 (32.9%)	70
Total	83 (20.1%)	172 (41.7%)	157 (38.1%)	412

assessing verifiability. Because this study was conducted by a single author, inter-rater reliability could not be assessed. To reduce annotation drift, scoring criteria were specified before annotation began and applied consistently throughout the review process. Example claims corresponding to each score level were developed during rubric design and used as reference points during annotation. To promote consistency, the scoring rubric was defined prior to annotation and applied uniformly across all claims. Future work would evaluate inter-rater agreement across multiple annotators and compare human judgments with automated assessment approaches.

3. Results

Table 1 summarizes verifiability across 412 extracted claims by source and category. Overall, 38.1% of claims are well-supported (**Score 2**), 41.7% partially supported (**Score 1**), and 20.1% unsubstantiated (**Score 0**). Bootstrapped 1,000-resample confidence intervals for well-supported claims (33.5%–42.8%) indicate that the observed gaps are unlikely to be due to sampling variability. These findings suggest that documentation-based governance mechanisms may systematically overestimate the reliability of disclosed claims, creating risks for audit, certification, and regulatory compliance. Consistent with this, a chi-square test ($\chi^2(2) = 3.12$, $p = 0.21$) shows no significant difference between model cards and HELM reports, indicating that structured evaluation documents do not inherently ensure verifiability. By source, model card claims are slightly less verifiable than HELM reports (35.7% vs. 42.2% well-supported), with similar proportions of partially supported claims (42%). By category, safety claims are least verifiable (26.3% well-supported, 28.1% unsubstantiated), evaluation claims are strongest (45.6% well-supported), and alignment claims are intermediate (32.9% well-supported, 22.9% unsubstantiated). Under stricter reproducibility criteria (requiring code or full hyperparameter disclosure), the proportion of well-supported claims decreases from 38.1% to 29.6%, highlighting potential misalignment between current documentation practices and emerging audit standards such as the EU AI Act. Illustrative examples of claims assigned to each score category are provided in the Appendix A.5 to facilitate interpretation of the scoring rubric.

4. Conclusion

This study of 92 AI system documents finds persistent gaps in claim verifiability, with safety claims least supported. Limited documentation for external validation reduces audit usefulness, highlighting the need for stronger requirements for reproducible evidence (datasets, evaluations, and model details) to support auditing and accountable AI deployment. The study is limited by its focus on Hugging Face model cards and HELM reports and by the inherently judgment-based nature of rubric-guided annotation. Future work should examine additional documentation formats and assess annotation reliability across larger reviewer pools.

5. LLM Usage Disclosure

LLMs were used for limited editorial assistance, including grammar correction, wording refinement, and minor programming support during implementation. All study design, data collection, claim scoring, analysis, interpretation of results, and scientific conclusions were conducted and verified by the author. LLMs were not used to generate results, perform verifiability assessments, or draw scientific conclusions.

References

AI, T. Oecd digital economy papers. 2023.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Cantero Gamito, M. and Marsden, C. T. Artificial intelligence co-regulation? the role of standards in the eu ai act. *International journal of law and information technology*, 32(1):eaae011, 2024.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., and Wallach, H. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–16, 2019.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229, 2019.

Osborne, C., Ding, J., and Kirk, H. R. The ai community building the future? a quantitative analysis of development activity on hugging face hub. *Journal of Computational Social Science*, 7(2):2067–2105, 2024.

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Larochelle, H. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of machine learning research*, 22(164):1–20, 2021.

Semmelrock, H., Ross-Hellauer, T., Kopeinik, S., Theiler, D., Haberl, A., Thalmann, S., and Kowald, D. Reproducibility in machine-learning-based research: Overview, barriers, and drivers. *AI Magazine*, 46(2):e70002, 2025.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

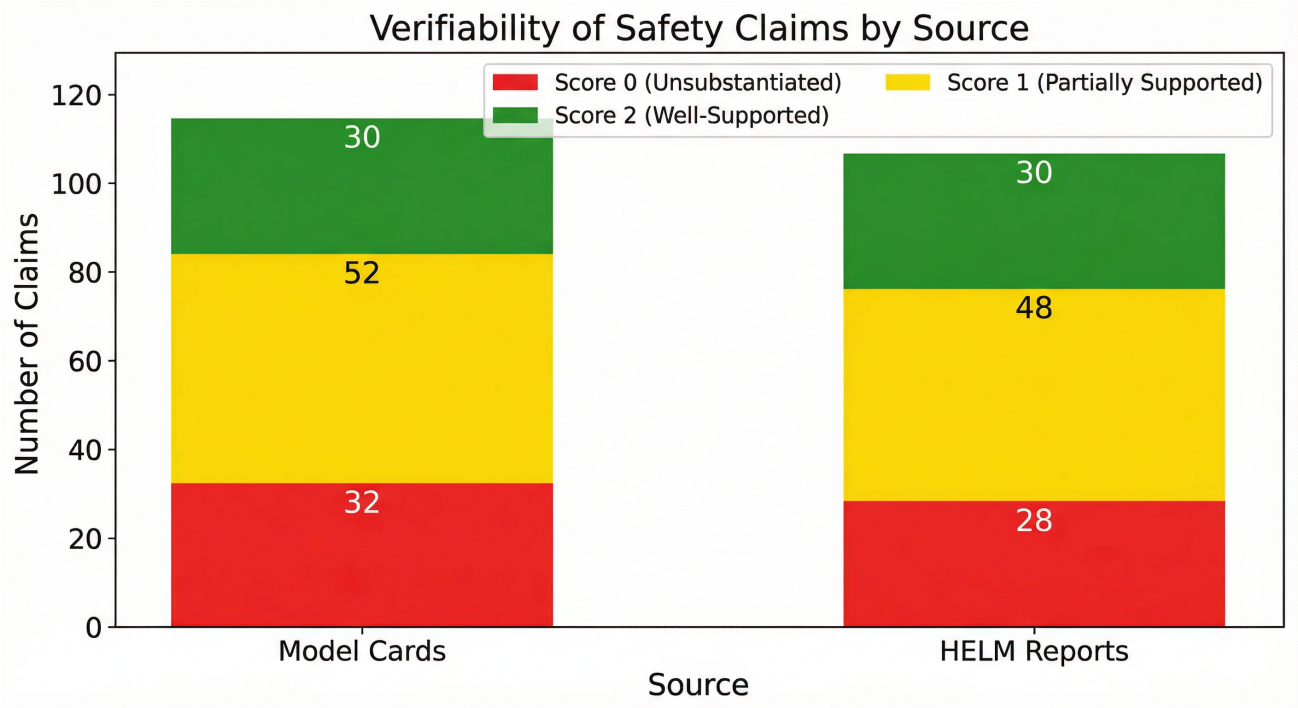

Figure 1. Verifiability of safety-related claims by source. Counts of Unsubstantiated (0), Partially Supported (1), and Well-Supported (2) claims are shown for model cards and HELM reports. Safety claims are the least verifiable.

A. Appendix: Detailed Methodology and Results

A.1. Additional Figures

Safety-related claims are of particular interest in AI governance because they directly relate to risk mitigation, harmful behavior reduction, and deployment suitability. However, such claims are often expressed in qualitative or underspecified terms in model documentation, making them difficult to independently verify without access to detailed evaluation protocols or datasets. To examine this systematically, we isolate safety claims and compare their verifiability distribution across model cards and benchmark reports, as shown in Figure 1, highlighting differences in documentation practices across development and evaluation contexts.

Figure 2 presents the distribution of all 412 claims across categories (Safety, Evaluation, Alignment) and verifiability scores (0 = Unsubstantiated, 1 = Partially supported, 2 = Well-supported). Safety claims exhibit the lowest verifiability, with 32 unsubstantiated and 30 well-supported, frequently lacking sufficient documentation such as dataset identifiers or evaluation protocols. Evaluation claims show higher levels of support, with 104 well-supported, consistent with more structured reporting of quantitative evaluation evidence. Alignment claims fall between these categories, often involving partially specified or non-reproducible evidence. These results highlight systematic differences in verifiability across claim types.

Figure 3 compares verifiability across document sources. Claims from Hugging Face model cards show slightly lower verifiability (35.7% well-supported) compared to HELM reports (42.2%), while partially supported claims remain similar across sources. The number of unsubstantiated claims is higher in model cards (58) than in HELM reports (25), indicating that documentation practices across sources do not consistently meet auditability requirements. This comparison highlights variation in documentation quality across developer-provided and benchmark-generated artifacts.

A.2. Claim Extraction Details

Claims were identified using a two-stage extraction process designed to isolate policy-relevant, testable assertions about AI systems. First, candidate sentences were selected using a keyword-based filter targeting safety-, evaluation-, and alignment-related language: `evaluated on`, `safe`, `robust`, `fair`, `mitigate`, `bias`, `alignment`.

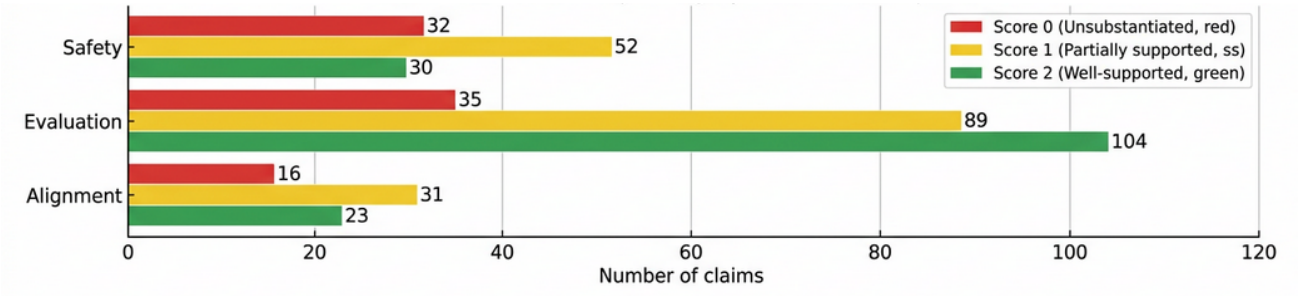

Figure 2. Distribution of claims by category and verifiability score. Claims from 92 AI system documents are categorized into Safety, Evaluation, and Alignment, and scored on a three-point scale. Safety claims exhibit lower verifiability relative to evaluation claims.

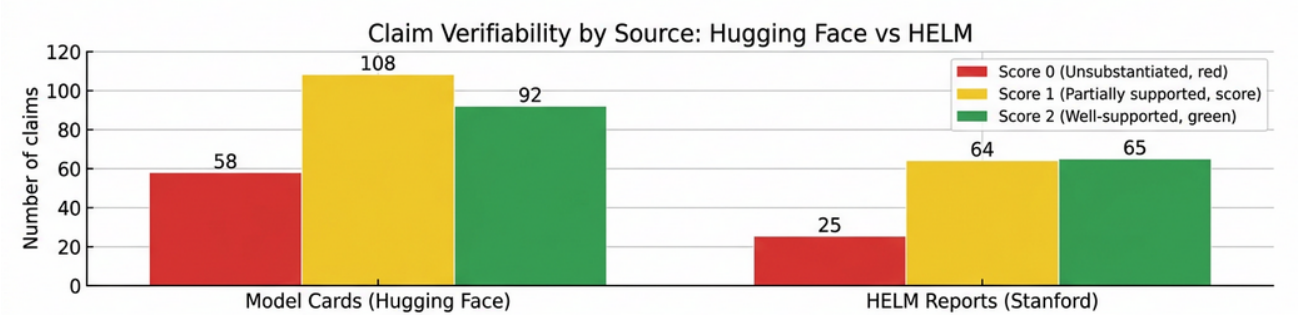

Figure 3. Comparison of claim verifiability across sources. Model cards exhibit a higher proportion of unsubstantiated claims relative to HELM reports, indicating differences in documentation completeness.

Second, a zero-shot language model (LLaMA-2-7B) was used to filter candidate sentences by classifying whether each constitutes a specific, testable claim using the prompt:

> Does the following sentence contain a specific, testable claim about an AI system's safety, evaluation, or alignment? Answer only YES or NO.

Only sentences exceeding a model-assigned confidence threshold of 0.7 were retained. The extracted claims were subsequently reviewed by the author to remove any remaining extraction errors before verifiability assessment. Automated extraction was used only to identify candidate claims and not to determine final verifiability scores. Validation on 100 manually labeled sentences yielded a precision of 0.84 and recall of 0.79, indicating reliable identification of claims suitable for downstream audit-oriented analysis. The resulting dataset consists of 412 unique claims extracted from 92 documents (258 from Hugging Face model cards and 154 from HELM reports). Duplicate claims across documents were removed to ensure each entry corresponds to a distinct, testable assertion relevant to independent verification.

A.3. Document and Model Coverage

The dataset includes both developer-provided documentation and independent benchmark reports, covering text, vision, and audio systems. Hugging Face model cards were selected because they represent one of the most widely used forms of AI system documentation and encompass models from academic, open-weight, and commercial development contexts. HELM reports were included because they provide structured third-party evaluation artifacts generated under a consistent benchmarking methodology. The objective was not to obtain a statistically representative sample of all AI system documentation, but rather to examine verifiability across two influential documentation formats frequently referenced in AI governance, transparency, and auditing discussions. The corpus spans widely deployed models, including systems such as LLaMA-2-7B (Touvron et al., 2023) and GPT-3 (Brown et al., 2020), ensuring that the analysis reflects documentation practices associated with high-impact, real-world AI systems. The sample includes both open-weight and proprietary systems documented through publicly available model cards or evaluation reports, enabling comparison across a range of deployment and disclosure contexts.

A.4. Scoring Criteria

Each claim was evaluated using a three-point rubric grounded in policy-relevant criteria for verifiability, reflecting requirements for auditability in emerging governance frameworks:

- **Score 0 (Unsubstantiated):** The claim lacks sufficient supporting evidence or documentation to enable assessment. Required elements such as dataset identifiers, evaluation procedures, or model specifications are absent, preventing meaningful verification.
- **Score 1 (Partially supported):** The claim includes some supporting information, but key details necessary for independent verification are incomplete, underspecified, or inaccessible. This includes cases where evaluation is referenced but not reproducible due to missing methodological detail.
- **Score 2 (Well-supported):** The claim is accompanied by sufficient, structured documentation to enable independent verification. This includes clear specification of datasets, evaluation protocols, model versions, and, where applicable, links to code or reproducible artifacts.

Scoring was applied directly to claims extracted from real-world documentation using this rubric. To facilitate interpretation of the scoring framework, illustrative examples corresponding to each score category are provided in Table 3.

Table 2. Operational definition of verifiability scores used for claim evaluation.

Score	Operational Criteria
0 (Unsubstantiated)	No verifiable evidence; missing datasets, protocols, or model details required for assessment.
1 (Partially supported)	Incomplete evidence; some evaluation or methodology referenced but insufficient for independent reproduction.
2 (Well-supported)	Complete and structured evidence; datasets, evaluation procedures, and model details disclosed to enable verification.

A.5. Illustrative Scoring Examples

Table 3 provides representative examples illustrating how claims were assigned verifiability scores.

Table 3. Illustrative examples of verifiability scoring. Examples are representative of claims encountered during analysis and are simplified for presentation.

Claim Example	Score	Rationale
"The model has been designed to reduce harmful, biased, and toxic outputs and is suitable for responsible deployment."	0	The claim asserts a safety property but provides no supporting evidence, evaluation methodology, benchmark results, or documentation that would enable independent verification.
"Internal testing indicated that the model generated fewer harmful responses than an earlier version across a set of safety prompts."	1	Some supporting evidence is referenced, but the testing procedure, prompt set, evaluation criteria, model versions, and replication details are not sufficiently disclosed for an independent party to confirm the result.
"The model was evaluated using a documented safety benchmark. The evaluation dataset, benchmark configuration, model version, scoring procedure, and implementation details are disclosed, and the reported results can be independently reproduced using the provided artifacts."	2	Sufficient information is provided for an independent party to reproduce the evaluation and verify the claim.

A.6. Ablation Study

To assess sensitivity to stricter verification standards, a subset of 50 claims was re-evaluated under enhanced reproducibility criteria requiring explicit availability of implementation details, such as open-source code or complete hyperparameter specifications. These stricter conditions reflect higher audit thresholds aligned with emerging regulatory expectations for technical transparency. Under these conditions, the proportion of well-supported claims decreased from 38.1% to 29.6%, indicating that a substantial fraction of claims that appear verifiable under baseline criteria do not meet stricter reproducibility requirements. Category-level percentages are computed within each subset.

A.7. Bootstrapping Procedure

We estimated uncertainty in verifiability proportions using bootstrapping with 1,000 resamples:

- Samples were drawn with replacement from the full set of 412 claims.
- The proportion of well-supported (Score 2) claims was computed for each resample.
- The 2.5th and 97.5th percentiles define the 95% confidence interval (33.5%–42.8%).

This approach provides a distribution-free estimate of variability and supports the statistical stability of the reported verifiability gaps.