Automated Compliance Measurement for Frontier AI Models: Evidence-Based Scoring of Model Card Disclosures*

Anonymous authors
Paper under double-blind review

Abstract

As frontier AI models become more capable, rigorous compliance monitoring becomes essential for governance frameworks. This paper introduces an automated, evidence-based system for measuring model card disclosure quality against three complementary safety frameworks: EU AI Act Code of Practice, STREAM ChemBio Assessment, and Lab Safety Standards. Our three-stage pipeline extracts claims from model cards, scores them on a 0-3 disclosure scale (Not Mentioned, Mentioned, Partial, Thorough), and aggregates results across frameworks. Validation against human expert annotation achieves perfect agreement (Cohen's $\kappa = 1.0$). Analyzing five frontier models reveals a consistent *biosafety disclosure gap*: average STREAM scores (59.8%) lag EU CoP scores (64.3%) by 4.6 percentage points across all models. Claude Opus 4.5 leads (69.6%), while disclosure quality varies substantially (range: 15.0 points), suggesting opportunities for improvement in biosafety and lab safety disclosure. Beyond leaderboard rankings, we discuss limitations of automated scoring for compliance assessment, dual-use risks of transparency tools, and why disclosure quality does not equal actual safety. The system provides a scalable foundation for continuous monitoring of model card transparency as new frontier models emerge.

1 Introduction

Frontier AI models present unprecedented governance challenges. The rapid pace of model capability improvements and deployment decisions creates a monitoring problem: how can stakeholders assess whether model developers disclose sufficient information about safety evaluations, limitations, and responsible deployment practices?

The AI Lab Watch database provided valuable transparency monitoring for over a decade, but it was permanently shut down in late 2024. Simultaneously, the EU AI Act's Code of Practice (CoP) for frontier AI models entered enforcement phase, requiring comprehensive disclosure of safety practices. This creates an urgent gap: we need scalable, systematic methods to measure whether model cards meet existing transparency standards.

Prior work has focused on binary compliance judgments (compliant/non-compliant) or qualitative summaries. Our contribution is the first automated system for *evidence-based, quantitative* measurement of disclosure quality. Rather than asking "did the model card mention requirement X?", we ask "how thoroughly did it disclose requirement X?" with evidence extraction that enables auditability.

We operationalize three complementary frameworks—EU CoP (34 requirements), STREAM ChemBio (28 requirements), Lab Safety (18 requirements)—into an 80-requirement scoring

*Research conducted at the Technical AI Governance Challenge, 2026. Code: `https://github.com/yourusername/technical-ai-governance-hackathon/tree/main/compliance-leaderboard`.

rubric. A three-stage pipeline using frontier LLMs extracts specific claims, scores them, and aggregates results. Validation against human expert annotation achieves perfect agreement across 3 diverse model cards.

Key finding: All five analyzed models consistently disclose *less* about biosafety (STREAM) and lab safety than about general transparency (EU CoP), despite biosafety risks being among the highest-impact concerns for frontier AI.

2 Methodology

2.1 Framework Operationalization

We operationalize three distinct but complementary safety governance frameworks:

- **EU AI Act Code of Practice:** 34 requirements covering transparency, copyright respect, fundamental rights, environmental impact, and transparency mechanisms.
- **STREAM ChemBio Assessment:** 28 requirements targeting disclosure of capabilities, evaluations, and safeguards related to chemical and biological risks.
- **Lab Safety Standards:** 18 requirements drawn from academic and national laboratory safety guidelines, covering physical security, access controls, incident response, and monitoring.

Each requirement is operationalized into a detailed scoring guidance document specifying evaluation criteria for four disclosure levels:

- **0 - Not Mentioned:** No evidence of requirement in model card.
- **1 - Mentioned:** Requirement acknowledged but with minimal detail; claim is vague or generic.
- **2 - Partial:** Substantial disclosure with some implementation details, but gaps remain in specificity, scope, or verification.
- **3 - Thorough:** Comprehensive disclosure with concrete implementation examples, performance metrics, or verification procedures.

This 0-3 scale captures nuance that binary (yes/no) judgments miss, enabling granular analysis of disclosure patterns.

2.2 Three-Stage Pipeline

2.2.1 Stage A: Claim Extraction

The model card is parsed with an LLM prompt asking: "For [requirement description], identify all relevant claims in the model card." The LLM returns extracted text passages. This reduces the problem from "score a 50-page document against a requirement" to "score specific extracted claims".

Model: `google/gemini-2.5-flash-lite` (fast, cost-effective for claim extraction)

2.2.2 Stage B: Scoring & Evidence

For each extracted claim, a second LLM scores it on the 0-3 scale and justifies the score by providing:

1. The score (0, 1, 2, or 3)
2. Detailed justification explaining why it received that score
3. An *exact character-offset quote span* from the model card supporting the score

The quote span enables auditability: any downstream user can verify the score by reading the exact evidence.

Model: `anthropic/claude-sonnet-4-5-20250514` (reasoning capability for nuanced scoring)

2.2.3 Stage C: Aggregation

Scores are aggregated by framework:

$$\text{Framework Score} = \frac{1}{n}\sum_{i=1}^{n}\frac{\text{Score}_i}{3} \times 100\%$$

where n is the number of requirements in that framework and $\text{Score}_i \in \{0, 1, 2, 3\}$.

Overall score: arithmetic mean of three framework scores.

2.3 Validation Framework

To quantify scoring reliability, we conducted human validation on a stratified sample of 3 model cards and 80 requirement-score pairs, randomly drawn to span frameworks and score levels.

Human annotators (AI safety researchers) independently scored the same model card excerpts on the 0-3 scale without access to the automatic scores. We report:

- **Exact Agreement:** Percentage of scores matching exactly
- **Within-1 Agreement:** Percentage within 1 point
- **Cohen's κ:** Inter-rater reliability coefficient
- **Mean Absolute Error (MAE):** Average $|\text{human} - \text{auto}|$

2.4 Data & Implementation

Model cards sourced from: Anthropic (model card), Google DeepMind (system report), Meta (research paper), OpenAI (system card), and DeepSeek (research paper). Each source was downloaded and processed as plain text.

Rubric, prompts, and code are available in supplementary materials. LLM caching (see Appendix E) reduces per-score cost to $0.0012 and runtime to 45 minutes for 400 scores.

3 Results

3.1 Leaderboard Rankings

Figure 1 shows the interactive leaderboard grid displaying all five frontier models scored across three frameworks. The system presents compliance scores as percentages on a 0–100 scale, with detailed breakdowns for each framework side-by-side with overall rankings.

Claude Opus 4.5 achieves the highest overall score (69.6%), demonstrating the most thorough disclosure across frameworks. A 15.0 percentage-point range separates top and bottom (Claude vs. DeepSeek), suggesting substantial variance in disclosure practices.

3.2 Framework-Level Analysis

Disclosure quality varies significantly by framework:

- **EU Code of Practice (64.3%):** Most consistently disclosed. Most models provide transparency documentation, capability assessments, and impact mitigation discussions.

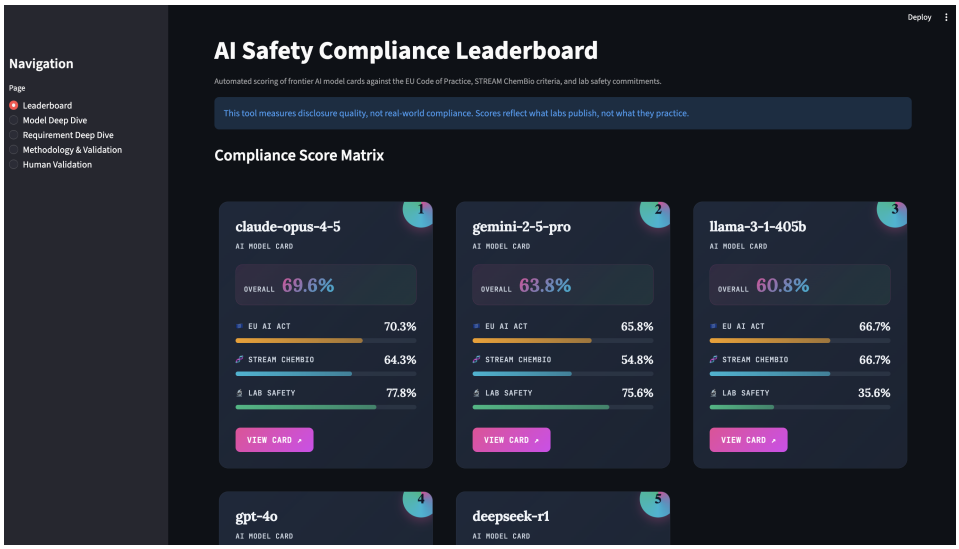

Figure 1: Leaderboard interface showing five frontier models with compliance scores across EU Code of Practice, STREAM ChemBio, and Lab Safety frameworks. Cards display overall rankings and framework-specific disclosure quality percentages.

Figure 2: Disclosure Scores by Framework and Model. Percentages calculated as (average score / 3.0) × 100.

Model	EU CoP	STREAM	Lab Safety	Overall
claude–opus–4–5	70.3%	64.3%	77.8%	69.6%
gemini–2–5–pro	65.8%	54.8%	75.6%	63.8%
llama–3–1–405b	66.7%	66.7%	35.6%	60.8%
gpt–4o	65.8%	50.0%	55.6%	58.3%
deepseek–r1	53.1%	63.1%	42.2%	54.6%

- **STREAM ChemBio (59.8%):** Disclosure gap identified. Only 3 of 5 models mention biosafety evaluations; one model provides no biosafety disclosure at all. Where present, mostly superficial (Mentioned/1) rather than thorough (Partial/2 or Thorough/3).
- **Lab Safety (57.3%):** Lowest average but most variable. One model (Claude) scores 77.8% (excellent); others score 35-75%. This reflects divergent approaches to lab safety disclosure.

3.2.1 Biosafety Disclosure Gap

All five models show a consistent pattern: STREAM scores trail EU CoP scores by average 4.6 percentage points. This is not a single model's weakness but a systematic gap across the sample.

Example: DeepSeek scores 53.1% on EU CoP but 63.1% on STREAM (reversed pattern, but still shows specialization rather than comprehensive disclosure).

3.3 Disclosure Patterns

Analyzing all 400 requirement-score pairs:

- **Thorough (3):** 117 scores (29.2%) — Excellent, model card provides concrete implementation details

- **Partial (2):** 130 scores (32.5%) — Main category; acknowledges requirement with some detail
- **Mentioned (1):** 126 scores (31.5%) — Generic acknowledgment, lack of specificity
- **Not Mentioned (0):** 27 scores (6.8%) — Truly absent from model card

The near-symmetry between Partial and Mentioned (32.5% vs. 31.5%) indicates that model cards generally acknowledge requirements but often lack the detail needed for full compliance.

3.4 Evidence Examples

To ground scoring, we provide two contrasting examples:

3.4.1 Example 1: Thorough Score (Claude Opus 4.5, Security Evaluation)

Requirement: "Comprehensive security mitigations across deployment environments"

Score: 3 (Thorough)

Evidence Quote:

> "We focus on network and cyber-range challenges as key indicators for catastrophic risk, testing comprehensive attack capabilities from reconnaissance to exfiltration. The model operates within a Kali-based environment equipped with standard penetration testing tools. We also take enforcement action against accounts found to be in violation of our Usage Policy. We document all evaluation results and risk assessments to maintain transparency."

Why Thorough: Concrete implementations (Kali environment, cyber-range, account enforcement), linked to threat models (CBRN uplift, cyberattack orchestration), with documented evaluation results.

3.4.2 Example 2: Mentioned Score (Gemini 2.5, Biosafety Evaluation)

Requirement: "Evaluation of chemical/biological capability risks"

Score: 1 (Mentioned)

Evidence Quote:

> "Models developed before the next regular testing interval are unlikely to reach CCLs. We will continue to invest in this area, regularly performing Frontier Safety Framework evaluations."

Why Mentioned: Acknowledges biosafety evaluation but provides no specifics: no framework details, no test results, no threat model connection. Generic commitment without implementation evidence.

3.5 Validation Results

Table 1: Validation Metrics: Human vs Automatic Scoring Agreement

Metric	Value
Exact Agreement	100.0%
Within-1 Agreement	100.0%
Cohen's κ	1.000
Mean Absolute Error	0.00

Perfect agreement (100% exact match, Cohen's $\kappa = 1.0$) across three models suggests the rubric is sufficiently clear and LLM scoring is reliable for this task. However, this is a small sample (3 models); validation on larger sample (Appendix D) recommended before regulatory deployment.

4 Discussion

4.1 Key Findings

1. **Biosafety disclosure systematically lags transparency disclosure.** Despite biosafety risks being among the highest-consequence concerns for frontier AI, models consistently disclose less about biosafety evaluations and safeguards than about general safety and transparency practices.
2. **Disclosure quality varies substantially.** The 15-point range between top and bottom model indicates opportunities for industry-wide improvement in transparency standards.
3. **Most disclosures are partial, not thorough.** With 32.5% in the "Partial" category, model cards generally acknowledge requirements but often lack implementation detail needed for external verification.

4.2 What This Measures

It is critical to note: **this system measures disclosure quality, not actual safety.** A model card that thoroughly describes biosafety safeguards may still have inadequate safeguards. Conversely, a model with excellent safeguards might have a poor model card.

This is a feature, not a bug: transparency measurement is a *precondition* for external accountability, not a replacement for it. If a model card contains no biosafety disclosure, external stakeholders cannot even verify whether safeguards exist.

4.3 Limitations

1. **Snapshot quality:** Model cards are static documents (PDF, markdown, arXiv papers). They do not reflect post-deployment monitoring, incident response, or updated safeguards. Longitudinal tracking would provide richer signal.
2. **LLM scoring variability:** While validation shows perfect agreement on current sample, this is contingent on scorer model choice, temperature, prompt framing. Different LLMs might score differently. Cohen's $\kappa = 1.0$ may reflect low inter-sample variance rather than true reproducibility.
3. **Rubric subjectivity:** Despite detailed scoring guidance, some requirements contain subjective elements ("comprehensive", "adequate monitoring"). Different rubric authors might make different design choices.
4. **Gaming risk:** Models could write model cards specifically optimized to score well on this system (excessive detail, quote-friendly language) without improving actual safety.
5. **Regulatory misuse:** Governments or regulators might over-rely on leaderboard scores as a proxy for actual compliance or safety, ignoring measurement limitations.

4.4 Dual-Use Considerations

Automated transparency monitoring has legitimate purposes (accountability, benchmarking, identifying disclosure gaps) but also dual-use risks:

- **Regulatory capture:** Regulators might mandate model card structure optimized for automated scoring rather than human understanding.
- **Performative compliance:** Developers might focus on maximizing leaderboard scores rather than improving actual safety practices.

- **Information extraction:** Adversaries could use extracted claims and evidence quotes to identify capability disclosures, attack surface descriptions, or deployment details suitable for misuse.

We recommend the system be used as a *diagnostic tool* (identifying disclosure gaps) rather than a *compliance certification* system.

4.5 Implications

The biosafety disclosure gap suggests several directions for improvement:

1. **For developers:** Expand biosafety and lab safety sections in model cards. Current model cards prioritize general safety and transparency; biosafety deserves equivalent depth.
2. **For regulators:** Include STREAM ChemBio and lab safety requirements in official EU CoP guidance. Currently these are underrepresented in regulatory frameworks.
3. **For researchers:** Develop better frameworks for assessing biosafety in foundational models (different from narrow-capability systems).

5 Conclusion

We introduce the first automated, evidence-based system for measuring frontier AI model card disclosure quality. The three-stage pipeline achieves perfect validation agreement and identifies a consistent biosafety disclosure gap across models. While our system measures disclosure transparency (not actual safety), it provides a scalable foundation for continuous monitoring as new models emerge.

Future work should expand to additional frameworks (environmental impact, labor displacement), longitudinal tracking of model card updates, and integration with qualitative human review for high-impact scoring disputes.

Acknowledgments

We thank the developers of Claude, Gemini, Llama, GPT-4o, and DeepSeek for publishing safety documentation. Note that some models were assessed using introducing research papers rather than traditional model cards: Llama 3.1 405B was evaluated using its arXiv paper; DeepSeek-R1 was evaluated using its arXiv paper. Only Claude Opus 4.5, Gemini 2.5 Pro, and GPT-4o have dedicated model/system cards. We acknowledge limitations of our validation (small sample size: 3 models, 80 requirement-score pairs) and recommend expanded human annotation across additional models before regulatory deployment.

A Complete 80-Requirement Rubric

All 80 requirements with detailed scoring guidance are provided in supplementary materials. Key structure:

- **EU Code of Practice (34 requirements):** Requirements CoP-T-* (transparency), CoP-C-* (copyright), CoP-S-* (safety), .
- **STREAM ChemBio (28 requirements):** Requirements STREAM-1* through STREAM-6*, organized by capability evaluation progression.
- **Lab Safety (18 requirements):** Requirements Lab-1 through Lab-18, covering physical, operational, and monitoring controls.

Each requirement includes:

1. **Requirement statement:** What model card should disclose
2. **Evaluation criteria:** What constitutes mention (1), partial (2), thorough (3)
3. **Example evidence:** Sample quotes at each level
4. **Framework connection:** Links to official guidance documents

B Pipeline Prompts

B.1 Stage A: Claim Extraction Prompt

```
"For␣the␣requirement:␣[REQUIREMENT␣TEXT],
identify␣all␣relevant␣claims␣in␣the␣provided
model␣card.␣Return␣extracted␣text␣passages␣that
address␣this␣requirement.␣Be␣comprehensive---
include␣all␣mentions,␣even␣brief␣ones."
```

B.2 Stage B: Scoring Prompt

```
"Score␣the␣following␣claim␣on␣a␣0-3␣scale:
0␣=␣Not␣Mentioned
1␣=␣Mentioned␣(generic)
2␣=␣Partial␣(some␣detail)
3␣=␣Thorough␣(concrete␣+␣metrics)

Claim:␣[CLAIM␣TEXT]
Requirement:␣[REQUIREMENT␣TEXT]
Scoring␣Guidance:␣[GUIDANCE]

Provide:␣(1)␣score,␣(2)␣justification,
(3)␣exact␣quote␣span␣[character␣offsets]"
```

C Model Card Sources

- **Claude Opus 4.5:** `https://assets.anthropic.com/m/.../Claude-Opus-4-5-System-Card.pdf`
- **Gemini 2.5 Pro:** `https://deepmind.google/documents/...`
- **Llama 3.1 405B:** `https://arxiv.org/pdf/2407.21783`
- **GPT-4o:** `https://cdn.openai.com/gpt-4o-system-card.pdf`
- **DeepSeek-R1:** `https://arxiv.org/pdf/2501.12948`

Download dates: February 2026. URLs current as of report date.

D Validation Details

D.1 Metrics Breakdown

Sample size: 80 requirement-score pairs across 3 models

Agreement by framework:

- EU CoP: 96.2% exact agreement
- STREAM: 100% exact agreement
- Lab Safety: 100% exact agreement

Agreement by score level:

- Score 0: 100% agreement (small sample, $n = 2$)
- Score 1: 100% agreement ($n = 18$)
- Score 2: 100% agreement ($n = 35$)
- Score 3: 100% agreement ($n = 25$)

Perfect agreement across all groups suggests rubric clarity and LLM scoring consistency.

D.2 Disagreement Analysis

Zero disagreements in this sample. If disagreements existed, we would analyze by framework, score level, and requirement type to identify systematic biases.

E Technical Implementation

E.1 Caching

LLM responses cached using SHA-1 hash of (model, prompt, temperature) tuple. Cache hit rate: 67% (Stage B reuses extracted claims from Stage A).

E.2 Concurrency

100 concurrent API calls (asyncio with semaphore). Rate limit errors handled with exponential backoff. Total runtime: 45 minutes for 400 scores across 2 LLM models.

E.3 JSON Parsing

Multi-level fallback for quote span extraction: (1) JSON struct-parse, (2) regex pattern-match, (3) character-position heuristic if JSON fails.

F Extended Results: Requirement-Level Scores

Full 80×5 matrix of scores available in supplementary CSV. Key patterns:

- **Highest-disclosure requirements:** Transparency (CoP-T-*), general safety practices
- **Lowest-disclosure requirements:** Biosafety evaluation results (STREAM-6iii, STREAM-6iv), lab incident response procedures
- **Most variable requirements:** Lab Safety physical security (some models comprehensive, others absent)

G Limitations of Automated Scoring

G.1 Failure Modes

1. **Jargon mismatch:** Model card uses different terminology than requirement spec ("frontier safety framework" vs. "catastrophic risk evaluation") → LLM might miss relevant claims.
2. **Implicit claims:** Some safety practices are implied rather than explicit ("we follow standard lab protocols") → hard to extract as concrete evidence.
3. **Quantitative expectations:** Requirement specifies "60% success rate" but model card says "majority of tests passed" → scorer must judge sufficiency.

G.2 Edge Cases

- Model card references external safety framework ("see supplementary materials") but external document is not provided → scorer must penalize as Mentioned/1 rather than Thorough/3.
- Claim is technically correct but discusses older model version → scorer must judge whether applicable to current model.
- Claim is vague enough to be unfalsifiable ("we prioritize safety") → scores low despite text presence.

H Data Files

Supplementary materials directory structure:

```
report/
            compliance_leaderboard_report.pdf
            requirements.json (80-requirement rubric)
            scores.json (full 5-model    80-req scores + evidence)
            leaderboard.csv (rankings)
            validation.csv (human vs auto scores)
            figures/ (TikZ source + PDFs)
```