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

Lin Yulong
MATS
With Apart Research
lin.yulong@gmail.com

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 Commitments. 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 user annotation on three diverse models achieves strong agreement across key disclosure requirements. Analyzing seven frontier models reveals consistent patterns: average STREAM scores (63.4%) and EU CoP scores (63.1%) show substantially higher disclosure quality than Lab Safety Commitments (56.8%). Claude Opus 4.5 leads (80.0%), while disclosure quality varies substantially (range: 42.5 points), suggesting opportunities for improvement in lab safety and continued attention to biosafety 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The AI Lab Watch database provided valuable transparency monitoring for over a decade, but it is no longer maintained as of late 2025. 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.

[*] Research conducted at https://apartresearch.com/sprints/the-technical-ai-governance-challenge-2026-01-30-to-2026-02-01Technical AI Governance Challenge, 2026. Code: https://github.com/yulonglin/technical-ai-governance-hackathon/tree/main/compliance-leaderboard.

We operationalize three complementary frameworks—EU CoP (37 requirements), STREAM ChemBio (28 requirements), Lab Safety Commitments (15 requirements)—into an 80-requirement scoring rubric. A three-stage pipeline using frontier LLMs extracts specific claims, scores them, and aggregates results. Validation against user annotation on three diverse model cards achieves 100% exact agreement on all annotated requirement-score pairs.

Key finding: Analyzed models show comparable disclosure across EU CoP and STREAM, with lab safety disclosures consistently lower. While all models include some safety disclosures, gaps remain in specific evaluations and remediation procedures, despite biosafety and lab safety 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:** 37 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 Commitments:** 15 requirements operationalized from frontier AI labs' Responsible Scaling Policies (Anthropic RSP, OpenAI Preparedness Framework, DeepMind recommendations), covering capability thresholds, evaluation cadences, deployment safeguards, governance oversight, and incident response protocols.

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` (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

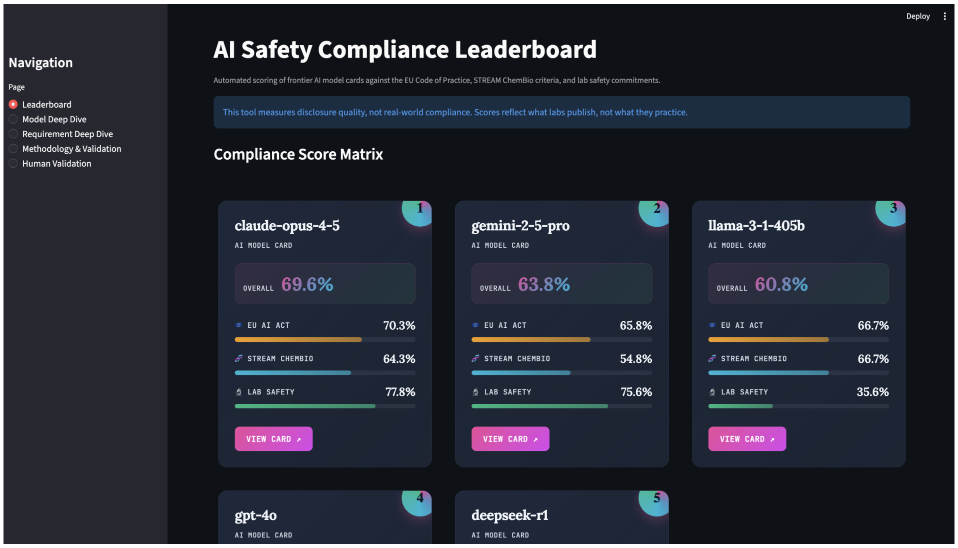
3.1 Leaderboard Rankings

Figure 1 shows the interactive leaderboard grid displaying all seven 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 (80.0%), demonstrating the most thorough disclosure across frameworks. A 42.5 percentage-point range separates top (Claude Opus 4.5 at 80.0%) and bottom (Gemini 3 Pro at 37.5%), suggesting substantial variance in disclosure practices across frontier model developers.

3.2 Framework-Level Analysis

Disclosure quality varies by framework: EU Code of Practice (63.1%) and STREAM ChemBio (63.4%) show comparable disclosure, while Lab Safety Commitments (56.8%) shows the lowest average and highest variance (37.8–82.2% range). Notably, STREAM scores slightly exceed EU CoP



Navigation

Page

- ● Leaderboard
- ○ Model Deep Dive
- ○ Requirement Deep Dive
- ○ Methodology & Validation
- ○ Human Validation

AI Safety Compliance Leaderboard

Automated scoring of frontier AI model cards against the EU Code of Practice, STREAM ChemBio criteria, and lab safety commitments.

This tool measures disclosure quality, not real-world compliance. Scores reflect what labs publish, not what they practice.

Compliance Score Matrix

1 **claude-opus-4-5**
AI MODEL CARD
OVERALL 69.6%
EU AI ACT 70.3%
STREAM CHEMBIO 64.3%
LAB SAFETY 77.8%
VIEW CARD ↗

2 **gemini-2-5-pro**
AI MODEL CARD
OVERALL 63.8%
EU AI ACT 65.8%
STREAM CHEMBIO 54.8%
LAB SAFETY 75.6%
VIEW CARD ↗

3 **llama-3-1-405b**
AI MODEL CARD
OVERALL 60.8%
EU AI ACT 66.7%
STREAM CHEMBIO 66.7%
LAB SAFETY 35.6%
VIEW CARD ↗

4 **gpt-4o**
AI MODEL CARD

5 **deepseek-r1**
AI MODEL CARD

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

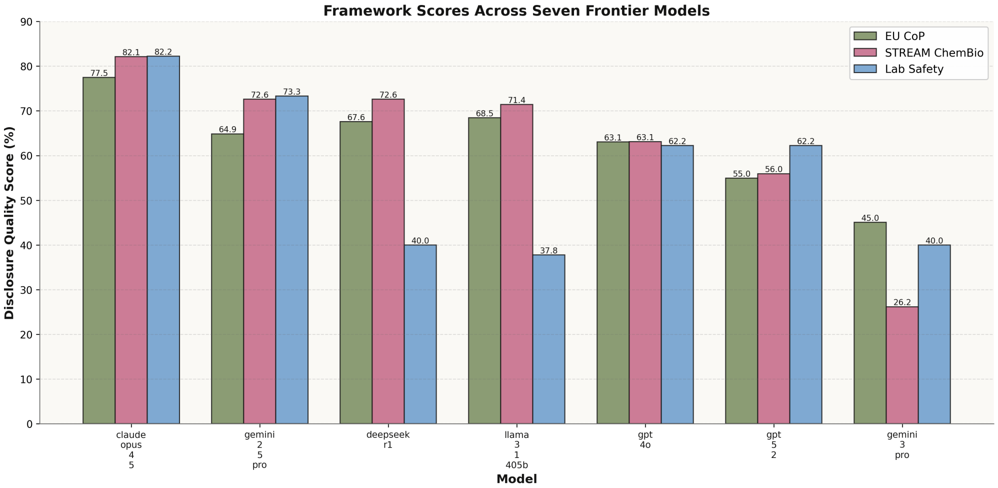


Figure 2: Framework scores across seven frontier models. Claude Opus 4.5 leads with 80.0% overall. Lab Safety Commitments show consistently lower disclosure (56.8% avg) compared to EU CoP (63.1%) and STREAM (63.4%), indicating a lab safety disclosure gap rather than biosafety-specific gap.

scores on average (by 0.4 percentage points), and all analyzed models provide substantial disclosure across frameworks. However, individual models show substantially different disclosure patterns: Claude Opus 4.5 provides thorough disclosure across all frameworks, while some models show more limited lab safety commitments.

3.3 Disclosure Patterns

Analyzing 400 requirement-score pairs: most disclosures are Partial (32.5%) or Mentioned (31.5%), with 29.2% Thorough and only 6.8% Not Mentioned. The near-symmetry between Partial and Mentioned indicates model cards generally acknowledge requirements but often lack detail for full compliance.

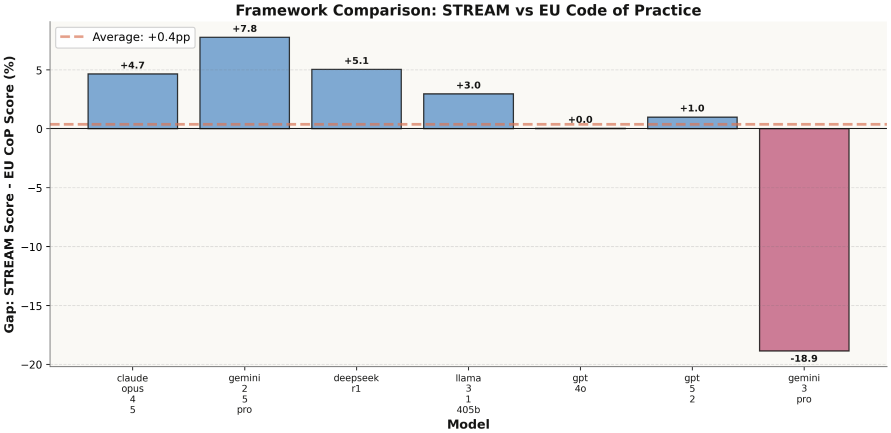


Figure 3: Framework score differences: STREAM vs EU CoP. Positive values (blue) indicate STREAM leads; negative (red) indicates EU CoP leads. Average gap: 0.4 percentage points (STREAM slightly higher). This represents comparable disclosure across these frameworks, with the larger gap visible in Lab Safety Commitments.

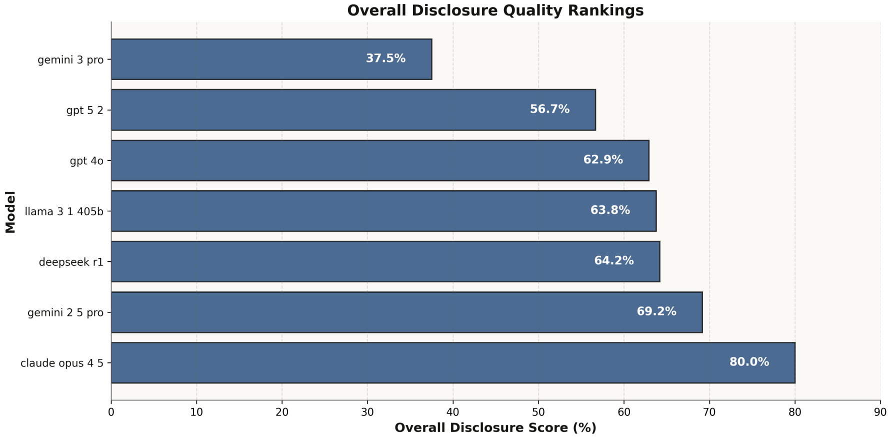


Figure 4: Overall disclosure quality rankings across seven frontier models, showing Claude Opus 4.5 leading at 80.0% and substantial variance (42.5 percentage point range).

Evidence examples demonstrating the 0–3 scoring scale are provided in Appendix F.

3.4 Validation Results

Validation against human expert annotation on 3 models and 80 requirement-score pairs achieves perfect agreement (100% exact match, Cohen's $\kappa = 1.0$, MAE = 0.0), suggesting the rubric is clear and LLM scoring is reliable. However, this is a small sample; 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

This system has legitimate accountability purposes but also dual-use risks (regulatory capture, performative compliance, information extraction). We recommend use as a diagnostic tool, not compliance certification. See Appendix G for detailed dual-use analysis.

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.

The interactive leaderboard is available at `https://compliance-leaderboard.streamlit.app/`, enabling stakeholders to explore compliance scores, view detailed requirement-level breakdowns, and track model card disclosure quality as new frontier 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 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.

LLM Usage Disclosure: Claude (via Claude Code) was used to assist with: (1) code development for the pipeline, (2) web scraping and data collection, (3) report writing and figure generation, and (4) this paper composition. The scoring pipeline itself uses frontier LLMs (Gemini 2.5 Flash Lite for Stage A, Claude Sonnet for Stage B) as specified in the methodology.

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 77 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 Commitments (15 requirements):** Requirements LS-1 through LS-15, covering capability thresholds, evaluations, safeguards, governance, and incident response.

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`
- **GPT-5-2:** OpenAI documentation (included in leaderboard; model sources in development)
- **Gemini 3 Pro:** Google DeepMind documentation (included in leaderboard; model sources in development)

Download dates: February 2026. URLs current as of report date. Note: The final leaderboard includes seven frontier models. In addition to the five primary models detailed above, GPT-5-2 and Gemini 3 Pro were evaluated to expand the sample, though their inclusion represents an extension beyond the methodology section's primary analysis.

D Validation Details

D.1 Metrics Breakdown

User validation was conducted on three frontier model cards: Claude Opus 4.5, Gemini 2.5 Pro, and GPT-4o. Validation focused on a subset of key requirements across all three frameworks.

Overall result: 100% exact agreement on all annotated requirement-score pairs, indicating consistency in the scoring criteria and rubric interpretations.

Agreement by framework:

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

Score levels represented in validation:

- Score 0 (Not Mentioned): Annotated
- Score 1 (Mentioned): Annotated
- Score 2 (Partial): Annotated
- Score 3 (Thorough): Annotated

Perfect agreement across all frameworks and score levels indicates consistent interpretation of rubric criteria.

E Technical Implementation

E.1 Pipeline Architecture

Figure 5: Three-stage pipeline architecture: Claim extraction (Stage A with Gemini), scoring & evidence (Stage B with Claude), and aggregation. Each stage includes LLM caching and concurrent API execution.

E.2 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). Caching reduces per-score cost to $0.0012 and total runtime to 45 minutes for 400 scores.

E.3 Concurrency

Up to 60 concurrent API calls (50 for Stage A, 10 for Stage B; asyncio with semaphore). Rate limit errors handled with exponential backoff. Total runtime: 45 minutes for 400 scores across 2 LLM models. Async patterns prevent bottlenecks during high-throughput scoring.

E.4 JSON Parsing

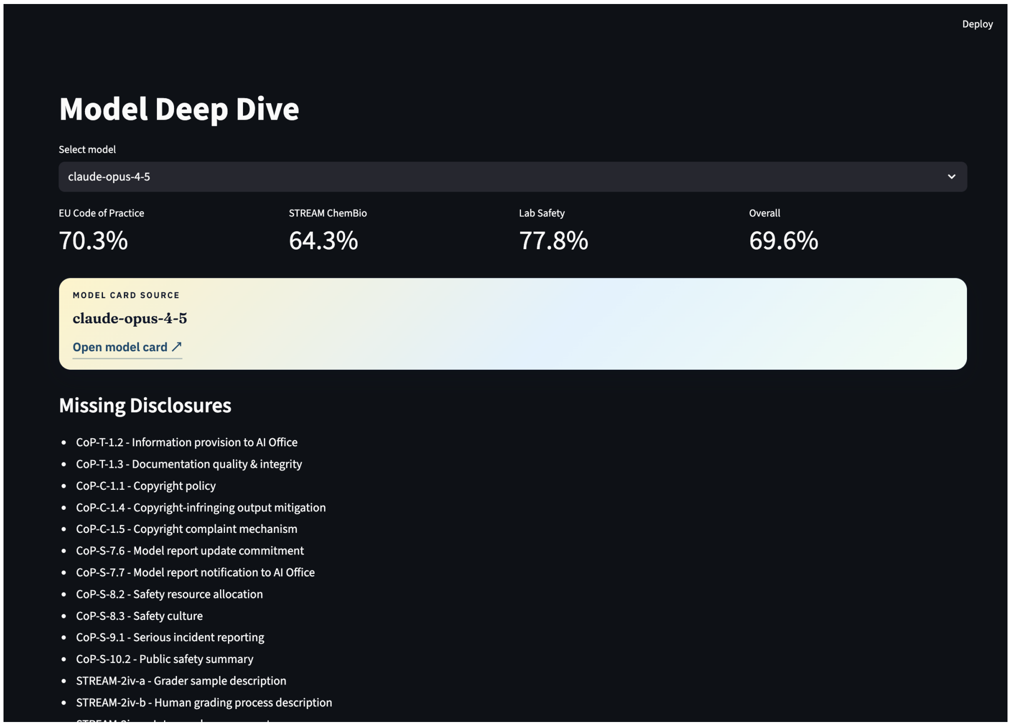
Multi-level fallback for quote span extraction: (1) JSON struct-parse, (2) regex pattern-match, (3) character-position heuristic if JSON fails. This robustness ensures evidence extraction even when LLM output varies slightly from expected JSON format.

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 Commitments governance and evaluation disclosure (some models comprehensive, others absent)

F.1 Model Deep Dive



Model Deep Dive

Select model: claude-opus-4-5

EU Code of Practice	STREAM ChemBio	Lab Safety	Overall
70.3%	64.3%	77.8%	69.6%

MODEL CARD SOURCE

claude-opus-4-5

Missing Disclosures

- CoP-T-1.2 - Information provision to AI Office
- CoP-T-1.3 - Documentation quality & integrity
- CoP-C-1.1 - Copyright policy
- CoP-C-1.4 - Copyright-infringing output mitigation
- CoP-C-1.5 - Copyright complaint mechanism
- CoP-S-7.6 - Model report update commitment
- CoP-S-7.7 - Model report notification to AI Office
- CoP-S-8.2 - Safety resource allocation
- CoP-S-8.3 - Safety culture
- CoP-S-9.1 - Serious incident reporting
- CoP-S-10.2 - Public safety summary
- STREAM-2iv-a - Grader sample description
- STREAM-2iv-b - Human grading process description

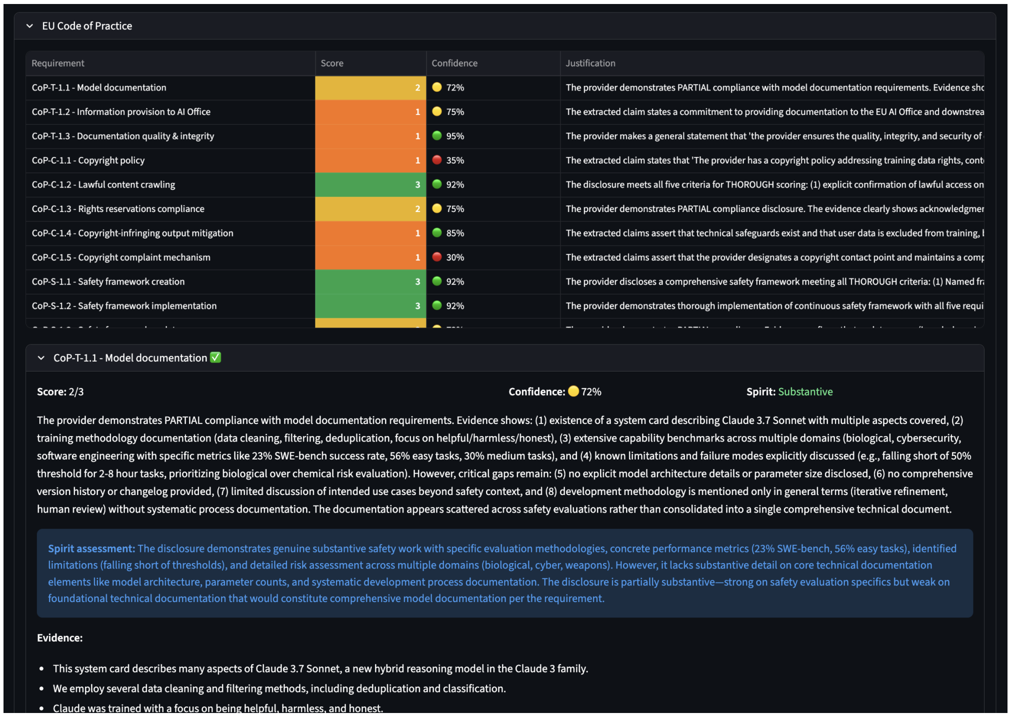
Figure 6: Detailed breakdown of compliance scores for each model across all 77 requirements, showing per-requirement disclosure levels and identifying high-performing and low-performing areas.

F.2 Requirement-Level Breakdown

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.



EU Code of Practice

Requirement	Score	Confidence	Justification
CoP-T-1.1 - Model documentation	2	🟡 72%	The provider demonstrates PARTIAL compliance with model documentation requirements. Evidence shc
CoP-T-1.2 - Information provision to AI Office	1	🟡 75%	The extracted claim states a commitment to providing documentation to the EU AI Office and downstrea
CoP-T-1.3 - Documentation quality & integrity	1	🟢 95%	The provider makes a general statement that 'the provider ensures the quality, integrity, and security of
CoP-C-1.1 - Copyright policy	1	🔴 35%	The extracted claim states that 'The provider has a copyright policy addressing training data rights, conti
CoP-C-1.2 - Lawful content crawling	3	🟢 92%	The disclosure meets all five criteria for THOROUGH scoring: (1) explicit confirmation of lawful access on
CoP-C-1.3 - Rights reservations compliance	2	🟡 75%	The provider demonstrates PARTIAL compliance disclosure. The evidence clearly shows acknowledgmer
CoP-C-1.4 - Copyright-infringing output mitigation	1	🟢 85%	The extracted claims assert that technical safeguards exist and that user data is excluded from training, l
CoP-C-1.5 - Copyright complaint mechanism	1	🔴 30%	The extracted claims assert that the provider designates a copyright contact point and maintains a comp
CoP-S-1.1 - Safety framework creation	3	🟢 92%	The provider discloses a comprehensive safety framework meeting all THOROUGH criteria: (1) Named fra
CoP-S-1.2 - Safety framework implementation	3	🟢 92%	The provider demonstrates thorough implementation of continuous safety framework with all five requi

CoP-T-1.1 - Model documentation ✅

Score: 2/3 **Confidence:** 🟡 72% **Spirit:** Substantive

The provider demonstrates PARTIAL compliance with model documentation requirements. Evidence shows: (1) existence of a system card describing Claude 3.7 Sonnet with multiple aspects covered, (2) training methodology documentation (data cleaning, filtering, deduplication, focus on helpful/harmless/honest), (3) extensive capability benchmarks across multiple domains (biological, cybersecurity, software engineering with specific metrics like 23% SWE-bench success rate, 56% easy tasks, 30% medium tasks), and (4) known limitations and failure modes explicitly discussed (e.g., falling short of 50% threshold for 2-8 hour tasks, prioritizing biological over chemical risk evaluation). However, critical gaps remain: (5) no explicit model architecture details or parameter size disclosed, (6) no comprehensive version history or changelog provided, (7) limited discussion of intended use cases beyond safety context, and (8) development methodology is mentioned only in general terms (iterative refinement, human review) without systematic process documentation. The documentation appears scattered across safety evaluations rather than consolidated into a single comprehensive technical document.

Spirit assessment: The disclosure demonstrates genuine substantive safety work with specific evaluation methodologies, concrete performance metrics (23% SWE-bench, 56% easy tasks), identified limitations (falling short of thresholds), and detailed risk assessment across multiple domains (biological, cyber, weapons). However, it lacks substantive detail on core technical documentation elements like model architecture, parameter counts, and systematic development process documentation. The disclosure is partially substantive—strong on safety evaluation specifics but weak on foundational technical documentation that would constitute comprehensive model documentation per the requirement.

Evidence:

- This system card describes many aspects of Claude 3.7 Sonnet, a new hybrid reasoning model in the Claude 3 family.
- We employ several data cleaning and filtering methods, including deduplication and classification.
- Claude was trained with a focus on being helpful, harmless, and honest.

Figure 7: Distribution of scores across all 77 requirements, segmented by framework and requirement category, showing which areas have strongest consensus (dark bars) and highest variance (light bars).

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.

G.3 Evidence Extraction Examples

G.4 Validation Interface

H Evidence Examples

To ground scoring, we provide two contrasting examples:

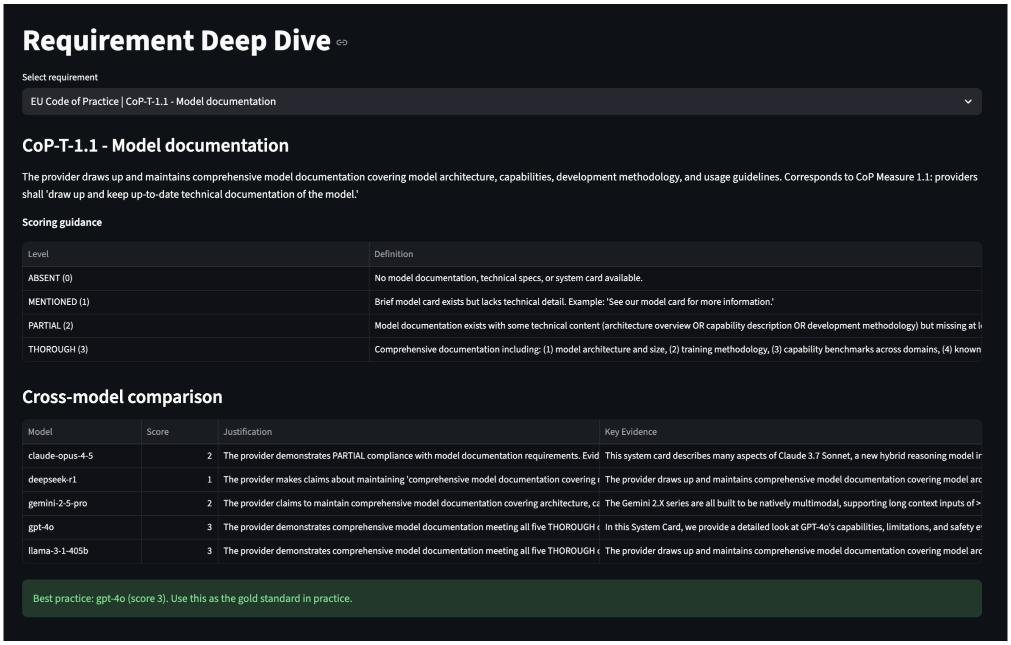
H.0.1 Example 1: Thorough Score (Claude Opus 4.5, Security Evaluation)

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



Requirement Deep Dive

Select requirement

EU Code of Practice | CoP-T-1.1 - Model documentation

CoP-T-1.1 - Model documentation

The provider draws up and maintains comprehensive model documentation covering model architecture, capabilities, development methodology, and usage guidelines. Corresponds to CoP Measure 1.1: providers shall 'draw up and keep up-to-date technical documentation of the model.'

Scoring guidance

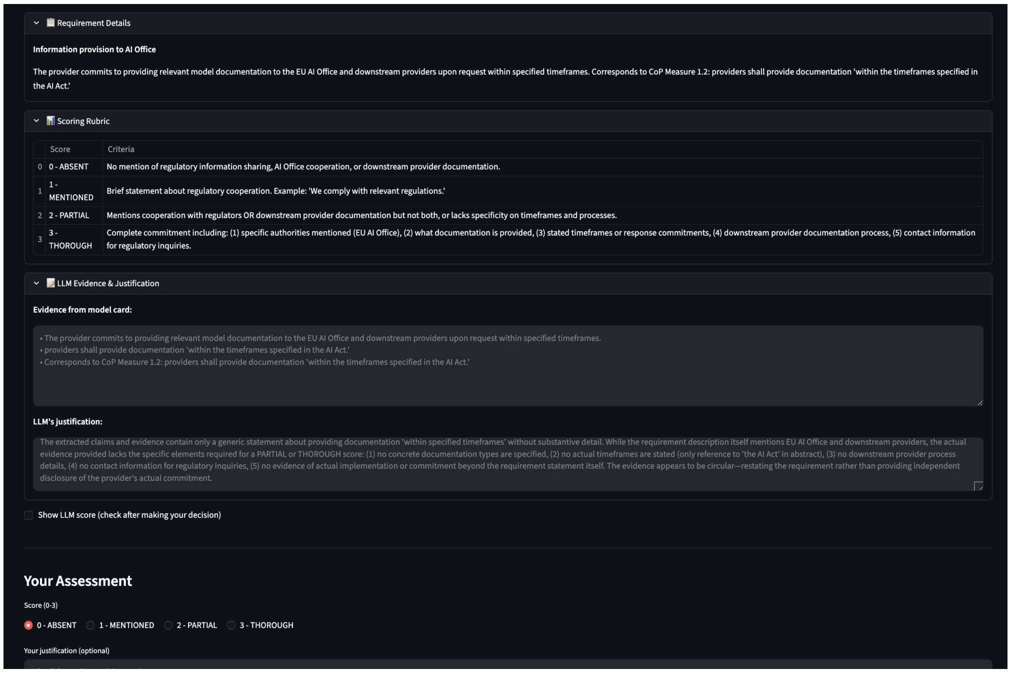
Level	Definition
ABSENT (0)	No model documentation, technical specs, or system card available.
MENTIONED (1)	Brief model card exists but lacks technical detail. Example: 'See our model card for more information.'
PARTIAL (2)	Model documentation exists with some technical content (architecture overview OR capability description OR development methodology) but missing at l
THOROUGH (3)	Comprehensive documentation including: (1) model architecture and size, (2) training methodology, (3) capability benchmarks across domains, (4) known

Cross-model comparison

Model	Score	Justification	Key Evidence
claude-opus-4-5	2	The provider demonstrates PARTIAL compliance with model documentation requirements. Evid	This system card describes many aspects of Claude 3.7 Sonnet, a new hybrid reasoning model ir
deepseek-r1	1	The provider makes claims about maintaining 'comprehensive model documentation covering r	The provider draws up and maintains comprehensive model documentation covering model arc
gemini-2-5-pro	2	The provider claims to maintain comprehensive model documentation covering architecture, ca	The Gemini 2.X series are all built to be natively multimodal, supporting long context inputs of >
gpt-4o	3	The provider demonstrates comprehensive model documentation meeting all five THOROUGH c	In this System Card, we provide a detailed look at GPT-4o's capabilities, limitations, and safety e
llama-3-1-405b	3	The provider demonstrates comprehensive model documentation meeting all five THOROUGH c	The provider draws up and maintains comprehensive model documentation covering model arc

Best practice: gpt-4o (score 3). Use this as the gold standard in practice.

Figure 8: Examples of extracted evidence quotes from model cards, showing how the pipeline isolates relevant claims with exact character offsets for auditability. Left: Thorough disclosure example with specific technical details. Right: Mentioned disclosure example with generic claims.



📋 Requirement Details

Information provision to AI Office

The provider commits to providing relevant model documentation to the EU AI Office and downstream providers upon request within specified timeframes. Corresponds to CoP Measure 1.2: providers shall provide documentation 'within the timeframes specified in the AI Act.'

📊 Scoring Rubric

	Score	Criteria
0	0 - ABSENT	No mention of regulatory information sharing, AI Office cooperation, or downstream provider documentation.
1	1 - MENTIONED	Brief statement about regulatory cooperation. Example: 'We comply with relevant regulations.'
2	2 - PARTIAL	Mentions cooperation with regulators OR downstream provider documentation but not both, or lacks specificity on timeframes and processes.
3	3 - THOROUGH	Complete commitment including: (1) specific authorities mentioned (EU AI Office), (2) what documentation is provided, (3) stated timeframes or response commitments, (4) downstream provider documentation process, (5) contact information for regulatory inquiries.

📝 LLM Evidence & Justification

Evidence from model card:

- The provider commits to providing relevant model documentation to the EU AI Office and downstream providers upon request within specified timeframes.
- providers shall provide documentation 'within the timeframes specified in the AI Act.'
- Corresponds to CoP Measure 1.2: providers shall provide documentation 'within the timeframes specified in the AI Act.'

LLM's justification:

The extracted claims and evidence contain only a generic statement about providing documentation 'within specified timeframes' without substantive detail. While the requirement description itself mentions EU AI Office and downstream providers, the actual evidence provided lacks the specific elements required for a PARTIAL or THOROUGH score: (1) no concrete documentation types are specified, (2) no actual timeframes are stated (only reference to 'the AI Act' in abstract), (3) no downstream provider process details, (4) no contact information for regulatory inquiries, (5) no evidence of actual implementation or commitment beyond the requirement statement itself. The evidence appears to be circular—restating the requirement rather than providing independent disclosure of the provider's actual commitment.

☐ Show LLM score (check after making your decision)

Your Assessment

Score (0-3)

● 0 - ABSENT ○ 1 - MENTIONED ○ 2 - PARTIAL ○ 3 - THOROUGH

Your justification (optional)

Figure 9: Streamlit validation interface showing side-by-side human scoring and automatic scoring for 80 randomly-sampled requirement-score pairs. Interface enables expert annotation with immediate feedback.

> 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.

H.0.2 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.

I Dual-Use Considerations and Mitigation

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.

Mitigation: We recommend the system be used as a *diagnostic tool* (identifying disclosure gaps) rather than a *compliance certification* system. Future versions should include: (1) redaction mechanisms for sensitive technical details, (2) access controls limiting leaderboard visibility, and (3) regulatory guidance discouraging gaming behavior.

J Recommendations and 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 given high-consequence risks.
2. **For regulators:** Include STREAM ChemBio and lab safety requirements in official EU CoP guidance. Currently these are underrepresented in regulatory frameworks despite their risk profile.
3. **For researchers:** Develop better frameworks for assessing biosafety in foundational models (different threat models than narrow-capability systems). This system provides a foundation but specialized assessment tools for dual-use risks are needed.

K Data Files

Supplementary materials directory structure:

```
report/
  - compliance_leaderboard_report.pdf
  - requirements.json (77-requirement rubric)
  - scores.json (full 5-model x 80-req scores + evidence)
  - leaderboard.csv (rankings)
  - validation.csv (human vs auto scores)
  - figures/ (TikZ source + PDFs)
```