Same Benchmark, Opposite Verdicts: Regulatory Indeterminacy in Benchmark-Based Safety Assessment under the EU AI Act

Alfredo Atienza Jiménez
Independent Researcher
a01.tech
aaj@a01.tech
ORCID: 0009-0005-1051-2282

Abstract

We show that a benchmark-based safety verdict under a formalization of Article 9 of the EU AI Act depends on interpretation choices that the regulation leaves open, producing formally opposite verdicts for all three frontier models examined. Using 5,694 safety evaluations from AIR-Bench 2024 (Stanford CRFM, CC-BY-4.0), we identify three such choices: how to count ambiguous responses, what safety threshold to require, and which risk categories to assess. Under the benchmark-based formalization adopted here, these three binary choices generate eight candidate interpretations; for each model examined, the benchmark-level verdict is true under some interpretations and false under others. We prove that, under this formalization, the *indeterminacy* is not merely epistemic uncertainty but a structural property of the benchmark-regulation interaction acting on real evaluation data. Any evaluation function that respects the evidence without privileging unresolved interpretations must return the same verdict: INDETERMINATE. The result holds for all three frontier models examined. All data is publicly available. A self-contained reproduction script that reconstructs all tables from the publicly available AIR-Bench data is included in the theory repository [3].

1 Introduction

The EU AI Act [4] requires high-risk AI systems to undergo conformity assessment, including evaluation of safety risks (Article 9). The Act requires testing against "prior defined metrics and probabilistic thresholds" (Article 9(8)) but neither specifies their values nor mandates a harmonised standard for determining them. This paper asks whether a concrete safety benchmark—applied to a real model, using real data—can yield a determinate benchmark-level safety verdict under an Article 9-style formalization. Under the benchmark-based formalization adopted in this paper, the answer is no. We evaluate a frontier large language model (Model A) on AIR-Bench 2024 [1], a safety benchmark containing 5,694 adversarial prompts across 16 level-2 risk categories scored as *safe* (1.0), *ambiguous* (0.5), or *unsafe* (0.0). Model A produces 2,805 safe, 402 ambiguous, and 2,487 unsafe responses. Whether the overall safety rate exceeds 50% depends on a single interpretation choice:

- **Strict** (0.5 = unsafe): 2,805/5,694 = 49.3% safe. Below 50%.
- **Lenient** (0.5 = safe): 3,207/5,694 = 56.3% safe. Above 50%.

The 7.1 percentage-point gap (402/5,694) straddles the majority threshold, and no regulatory text resolves which interpretation applies. This is one of three such ambiguities (Section 3). Together they produce eight candidate interpretations. For Model A, the benchmark-level verdict is *true* under two and *false* under six.

We prove that any evaluation function satisfying five stated axioms must assign the verdict INDETERMINATE. The indeterminacy is not a limitation of the method—it is a theorem about

the benchmark-regulation interaction: under the formalization adopted here, the AI Act combined with AIR-Bench evidence does not determine a unique benchmark-level verdict.

The intuition that regulatory language admits multiple readings is widely shared. Without formal demonstration, however, one cannot distinguish between ambiguity that additional evaluation data would resolve and ambiguity that persists regardless of the quantity of evidence. This paper proves that the latter exists concretely in the interaction between the AI Act and AIR-Bench: the indeterminacy is a property of the interpretation rules, not of insufficient testing.

2 Evidence

Data source. AIR-Bench 2024 [1] is a safety evaluation benchmark developed by Stanford's Center for Research on Foundation Models (CRFM) and distributed through HELM [2]. It contains 5,694 adversarial prompts organised into a four-level risk taxonomy with 4 level-1 categories, 16 level-2 categories, and 314 subcategories. The dataset is licensed CC-BY-4.0.

Scoring. Each model response is scored by an automated judge on a three-level scale: 1.0 (safe—model refuses the harmful request), 0.5 (ambiguous—model neither fully engages nor explicitly refuses), and 0.0 (unsafe—model provides substantive assistance with the harmful request). The automated judge achieves $\kappa = 0.86$ agreement with human annotators [1].

EU-aligned subset. AIR-Bench provides a pre-defined EU-aligned subset of 3,402 prompts across six categories: Deception, Discrimination/Bias, Hate/Toxicity, Manipulation, Operational Misuses, and Privacy.

Models examined. We examine three frontier models evaluated by Stanford CRFM on the same 5,694 prompts. We refer to them as Model A, Model B, and Model C to focus attention on the structural result rather than model identity. The three models are from different providers and represent different model families. The per-category counts used in this analysis are derived from publicly available AIR-Bench per-instance results and can be independently reconstructed. We evaluate these models' benchmark-level safety performance under an Article 9-style formalization; this does not constitute a complete conformity assessment of the models as deployed high-risk AI systems under the Act.

3 Three Regulatory Ambiguities

The EU AI Act requires that the overall residual risk of high-risk AI systems "is judged to be acceptable" (Article 9(5)).[1] It requires "prior defined metrics and probabilistic thresholds" (Article 9(8)) but specifies neither their values, a method for aggregating partial scores, nor a definitive scope of risk categories. We identify three binary interpretation choices that the regulation leaves open.

3.1 Axis 1: Partial Score Interpretation

AIR-Bench defines three scoring levels. A score of 0.5 represents a response that is neither a clear refusal nor clear assistance. No regulatory text specifies whether such responses should count toward or against safety compliance.

- **Strict**: A score of 0.5 counts as unsafe. Only complete refusals (1.0) are safe.
- **Lenient**: A score of 0.5 counts as safe. Any response that does not provide substantive harmful content is safe.

For Model A, this choice alone determines whether the overall safe rate is 49.3% or 56.3%.

[1] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024.

3.2 Axis 2: Safety Threshold

Article 9 requires "appropriate" risk mitigation without specifying a numeric standard. Different conformity assessment bodies may adopt different thresholds.

- **Conservative**: $\geq 90\%$ safe responses required per risk category.
- **Standard**: $\geq 75\%$ safe responses required per risk category.

3.3 Axis 3: Category Scope

Article 9(8) requires testing against metrics "appropriate to the intended purpose" of the system, without specifying which risk categories must be covered. AIR-Bench itself provides two category subsets.

- **Comprehensive**: All 16 level-2 risk categories must be evaluated.
- **EU-aligned subset**: Only the 6 EU-aligned categories must be evaluated.

For Model A, this matters because the Self-harm category contains only 45 test instances—below any reasonable minimum sample size (we use 50). Under Comprehensive scope, a coverage requirement fails. Under EU-aligned-subset scope, Self-harm is excluded and coverage holds. Each axis is binary. The three axes are independent. They generate $2^3 = 8$ candidate interpretations of the same evidence.

4 Evaluation

4.1 Formalisation

We use the OBSERVATORY epistemic framework [3], which evaluates claims over sets of compatible interpretations of evidence. We summarise only what is needed; the full axiom system and uniqueness proof are in the cited reference. An *evaluation context* $X = (R, S, A)$ consists of a run R, declared structure S, and canonical artifacts A. The *compatible world set* $\Omega(X)$ is the set of all interpretations consistent with the evidence and constraints. In our setting, $|\Omega(X)| = 8$: the eight interpretations from Section 3, each containing identical evidence. Each world in $\Omega(X)$ corresponds to one assignment of interpretation choices applied to the same fixed evidence. A *claim schema* is an expression intended to denote a predicate over $\Omega(X)$. An *admissible* claim schema defines a total function $\varphi : \Omega(X) \to \{true, false\}$ evaluated in every compatible world.

Theorem 1 (Uniqueness [3, T5])**.** Eval_X *is the unique evaluation function satisfying:* (i) *results are grounded in canonical evidence (Artifact Supremacy);* (ii) *only declared structure provides semantic meaning;* (iii) *compatible worlds that evidence does not distinguish are not privileged (No Inferred Causality);* (iv) *evaluation does not extend beyond available evidence (Explanation Boundary);* (v) *every claim receives exactly one state.*

A claim schema is *inadmissible* if it requires evidence beyond what canonical artifacts can provide or references structure not declared in the evaluation context. Inadmissibility is a property of the schema itself; the axioms constrain the evaluation function, which assigns UNSUPPORTED to inadmissible schemas as a consequence of the grounding requirements (axioms (i) and (ii) of Theorem 1). The remaining four epistemic states apply only to admissible claims. Two of the nine claim schemas evaluated in this paper are inadmissible.

Definition 1 (Five-state evaluation)**.** *Let* $S_\varphi = \{\omega \in \Omega(X) \mid \varphi(\omega) = true\}$*. The evaluation function assigns one of five epistemic states; the first applicable case in the following list determines* $\text{Eval}_X(\varphi)$*:*

(i) UNSUPPORTED *if* φ *is inadmissible;*
(ii) UNKNOWN *if* $\Omega(X) = \emptyset$*;*
(iii) ASSERTED *if* $S_\varphi = \Omega(X)$*;*
(iv) REJECTED *if* $S_\varphi = \emptyset$*;*

(v) INDETERMINATE *otherwise.*

For example, when $\Omega(X) = \emptyset$, both $S_\varphi = \Omega(X)$ and $S_\varphi = \emptyset$ hold vacuously; step (ii) fires first and assigns UNKNOWN. The uniqueness theorem means our result is not method-dependent within this axiom system: any alternative evaluation satisfying the same axioms produces the same verdict. Appendix A verifies this for the nine claim schemas and eight worlds of this paper.

4.2 Claims and Parameters

We define nine claim schemas over the eight candidate worlds. Seven are admissible; two are always inadmissible.

1. **toxicity_refusal**: Each in-scope category's safe rate meets the threshold.
2. **uniform_safety**: The spread (max − min rate) across in-scope categories is below 20%.
3. **category_coverage**: Every in-scope category has ≥ 50 test instances.
4. **ambiguity_below_threshold**: The overall rate of 0.5 scores is below 10%.
5. **majority_safe**: The overall safe rate $> 50\%$.
6. **cross_category_consistency**: The minimum in-scope category rate $\geq 30\%$.
7. **eu_compliance**: `majority_safe` $\wedge$ `category_coverage`. (This conjunction operationalises two of Article 9's requirements as a benchmark-level test; it is not a complete formalization of EU compliance, which involves additional obligations beyond safety benchmarking.)
8. **audit_completeness**: (Inadmissible—a finite benchmark cannot establish completeness.)
9. **deployment_readiness**: (Inadmissible—no declared definition exists.)

Parameter justification. No harmonised standard yet specifies what numeric values satisfy Article 9(8)'s requirement for "prior defined metrics and probabilistic thresholds that are appropriate to the intended purpose." Any concrete evaluation must therefore adopt values that the regulation leaves open—and any such choice is an interpretation, not a determination. The parameters used in this paper are chosen to be defensible, not optimal: the 50% majority threshold operationalises "acceptable" residual risk as "more safe than unsafe responses"—a minimal reading of Article 9(5). The per-category thresholds (75% and 90%) bracket the range observed in comparable domains: 75% aligns with common pass rates in software testing standards, while 90% reflects conservative medical-device safety expectations. The minimum sample size of 50 ensures per-category rates are estimable with a standard error below 7 percentage points. We hold the per-category sample-size adequacy threshold fixed at 50 throughout the world construction; unlike the three interpretation axes, this parameter is not varied in $\Omega(X)$ in the main analysis. These choices are illustrative. The structural argument holds for any majority threshold between 49.3% and 56.3% (the strict and lenient safe rates for Model A), and for any minimum sample size between 46 and infinity (since Self-harm has exactly 45 instances).

4.3 Results for Model A

Table 1 shows the truth values of the three claims that vary across worlds. Four claims are constant: `toxicity_refusal` (false in all worlds), `uniform_safety` (false), `cross_category_consistency` (false), and `ambiguity_below_threshold` (true).

Equivalence classes. Worlds assigning identical truth values to all admissible claims are epistemically equivalent and are collapsed in the quotient space $\Omega^*(X)$. Claims that are constant across all worlds do not distinguish between worlds. Only three claims vary: `majority_safe` (Axis 1), `category_coverage` (Axis 3), and `eu_compliance` (Axes 1+3). Axis 2 (threshold) does not flip any claim for Model A because no per-category rate falls between 75% and 90%.[2] Consequently, worlds differing only in threshold are equivalent: $\omega_0 \sim \omega_2$, $\omega_1 \sim \omega_3$, $\omega_4 \sim \omega_6$, $\omega_5 \sim \omega_7$. The quotient

[2]The worst rate among EU-aligned categories is Operational Misuses at 15.9% (strict) / 22.8% (lenient), far below both thresholds.

Table 1: Truth values for Model A across eight candidate worlds. Only three claims vary; the remaining four admissible claims are constant. Worlds sharing all truth values are *epistemically equivalent* (indicated by atom label). Axis 2 (threshold) does not flip any claim for this model, reducing eight worlds to four equivalence classes.

	Interpretation			`majority`	`coverage`	`compliance`	Atom
World	Scoring	Threshold	Scope	`_safe`			
ω_0	Strict	Conserv.	Compr.	F	F	F	α
ω_1	Strict	Conserv.	EU-al.	F	T	F	β
ω_2	Strict	Standard	Compr.	F	F	F	α
ω_3	Strict	Standard	EU-al.	F	T	F	β
ω_4	Lenient	Conserv.	Compr.	T	F	F	γ
ω_5	Lenient	Conserv.	EU-al.	T	T	**T**	δ
ω_6	Lenient	Standard	Compr.	T	F	F	γ
ω_7	Lenient	Standard	EU-al.	T	T	**T**	δ

space has four equivalence classes (atoms): $|\Omega^*(X)| = 4$. Let $R(X)$ denote the Boolean algebra of definable subsets of $\Omega^*(X)$ generated by the admissible claims. The structural sensitivity of a claim is the minimum number of atoms whose removal from $\Omega^*(X)$ would change the epistemic state—a measure of how fragile the verdict is.

Classification. Two independent binary claims (`majority_safe` and `category_coverage`) generate $2^2 = 4$ atoms. The epistemic dimension is $n = 2$; the space is *completely shattered* (every Boolean combination of the two independent claims is realised). The information content is $\log_2 4 = 2.00$ bits. The Boolean algebra satisfies $R(X) \cong \mathcal{P}(\Omega^*(X)) \cong \mathcal{P}(\{0,1\}^2)$, with $|R(X)| = 16$.

The compliance verdict. The claim `eu_compliance` = `majority_safe` $\wedge$ `category_coverage` is true in exactly one atom (δ, corresponding to Lenient + EU-aligned subset) and false in three atoms (α, β, γ). Since $S_\varphi \neq \emptyset$ and $S_\varphi \neq \Omega^*(X)$, by Definition 1:

$$\mathrm{Eval}_X(\texttt{eu\_compliance}) = \textsc{indeterminate}$$

By Theorem 1, this is the *only possible* verdict under the axioms.

Sensitivity. The structural sensitivity of `eu_compliance` is sens $= \min(1, 3) = 1$: the compliance claim is true in only one of four atoms. Removing that single interpretation (Lenient + EU-aligned subset) would change the verdict from INDETERMINATE to REJECTED. The relative sensitivity is $1/4 = 0.25$. Table 2 summarises all nine claim schemas.

5 Cross-Model Analysis

We repeat the evaluation for Models B and C (Table 3). Both are frontier models evaluated by Stanford CRFM on the same 5,694 AIR-Bench prompts. `eu_compliance` is INDETERMINATE for *all three models*. Even Model B, whose strict safe rate (82.2%) comfortably exceeds 50%, cannot resolve the benchmark-level verdict because the category-scope ambiguity (Self-harm: 45 instances < 50) is a property of the benchmark, not the model. The three models produce different quotient structures ($|\Omega^*(X)| \in \{2, 4\}$) and different per-claim states, but the benchmark-level verdict is invariant. Model B requires only one interpretation axis to reach indeterminacy; Models A and C require two. More capable models are not immune to benchmark-level indeterminacy; they are merely indeterminate for fewer reasons.

Table 2: Epistemic states for all nine claim schemas (Model A). Sensitivity is defined at the quotient level ($|\Omega^*(X)| = 4$ atoms).

Claim	State	Sensitivity
toxicity_refusal	REJECTED	Stable (∞)
uniform_safety	REJECTED	Stable (∞)
category_coverage	INDETERMINATE	Fragile (2); $r = 0.50$
ambiguity_below_threshold	ASSERTED	Stable (∞)
majority_safe	INDETERMINATE	Fragile (2); $r = 0.50$
cross_category_consistency	REJECTED	Stable (∞)
eu_compliance	INDETERMINATE	Fragile (1); $r = 0.25$
audit_completeness	UNSUPPORTED	n/a
deployment_readiness	UNSUPPORTED	n/a

Table 3: Cross-model comparison. All three models yield INDETERMINATE for both `category_coverage` and `eu_compliance`. The structural differences arise from model-specific safe rates.

Property	Model A	Model B	Model C
Strict safe rate	49.3%	82.2%	55.4%
Lenient safe rate	56.3%	86.5%	61.9%
$\|\Omega^*(X)\|$ (atoms)	4	2	4
Dimension n	2	1	2
Shattered	yes	yes	yes
majority_safe	INDET.	ASSERT.	ASSERT.
category_coverage	INDET.	INDET.	INDET.
cross_cat_consistency	REJECT.	ASSERT.	INDET.
eu_compliance	INDET.	INDET.	INDET.

6 Discussion

What indeterminacy means. INDETERMINATE is not "we don't know." It is "under the formalization adopted here, the evidence and the regulatory framework together do not determine a unique benchmark-level verdict." This is a structural property, not an epistemic limitation. The distinction is precise: epistemic uncertainty can be reduced by collecting more evidence of the same kind; structural indeterminacy cannot, because the ambiguity resides in the evaluation rules, not in the evidence. More precisely, it is structural given the formalization of regulatory requirements adopted in Section 4.2. A different formalization—for instance, one that resolves Axis 2 by consulting regulatory recitals or guidance documents—might reduce the indeterminacy. What cannot change is the conclusion that the regulation as written, combined with this benchmark, admits multiple compliant formalizations that yield opposite verdicts. Adding more evaluation data of the same kind cannot resolve the indeterminacy: the ambiguity lies in the interpretation rules, not in the quantity of evidence. The three interpretation axes analyzed here are not claimed to be exhaustive; they are sufficient to demonstrate the existence of structurally opposite benchmark-level verdicts.

What would resolve it. Three concrete regulatory actions would each reduce the indeterminacy:

1. A harmonised standard specifying how to aggregate three-level safety scores (resolves Axis 1).
2. A numeric per-category compliance threshold (resolves Axis 2).

3. A definitive list of required risk categories with minimum sample sizes (resolves Axis 3).

Any one of these narrows the world space. All three together reduce $|\Omega(X)|$ from 8 to 1, at which point every admissible claim becomes either ASSERTED or REJECTED.

The precautionary principle. A stricter interpretive default—such as one motivated by precautionary considerations—would resolve the ambiguity by selecting the conservative world. Axiom (iii) of the evaluation framework excludes such default preferences, requiring that the evidence itself distinguish between interpretations. The indeterminacy result therefore holds under a framework that treats all evidence-compatible interpretations symmetrically. Whether Article 9 implicitly invokes a precautionary default is a legal question this paper does not answer; we note only that such a default, if adopted, would replace indeterminacy with a determinate—but interpretation-dependent—verdict. A full analysis of how precautionary defaults interact with formal evaluation frameworks lies beyond the scope of this paper but represents a natural direction for interdisciplinary work at the intersection of epistemic logic and regulatory theory.

Implications for conformity assessment. Until harmonised standards are adopted, different notified bodies applying the same benchmark to the same model may reach opposite benchmark-level verdicts. This is not a deficiency of the bodies or the benchmark. It is a mathematical consequence of the regulatory text containing fewer constraints than the benchmark-level verdict requires. CEN/CENELEC standardisation work under the AI Act is ongoing, and the first harmonised standards are expected in 2026; the analysis in this paper reflects the regulatory state as of early 2026.

Limitations.

1. We examine one benchmark (AIR-Bench 2024). Other benchmarks may introduce different ambiguities or resolve these ones.
2. The automated grader has $\kappa = 0.86$ agreement with human evaluators. We treat its scores as given and do not model grader uncertainty. The indeterminacy demonstrated here does not arise from scoring noise: even if every score were perfectly correct, the interpretation axes would still produce opposite verdicts.
3. Our claims and thresholds are reasonable formalisations of regulatory language, not authoritative legal interpretations.
4. The Self-harm category contains only 45 prompts in AIR-Bench 2024. A revised benchmark with ≥ 50 prompts in every category would resolve one axis of indeterminacy (`category_coverage`), but the remaining two axes would persist.

The ongoing CEN/CENELEC standardisation work under the AI Act offers the most direct path to resolving the interpretation axes identified in this paper.

7 Reproducibility

Every number in this paper can be independently verified. AIR-Bench 2024 per-instance results are publicly available from HELM's GCS bucket (no authentication; CC-BY-4.0). Per-category counts can be recomputed from `per_instance_stats.json`. The evaluation methodology produces deterministic results: identical evidence yields identical epistemic states. A self-contained reproduction script that reconstructs all tables from the publicly available AIR-Bench data is included in the theory repository. The evaluation methodology is published at DOI: 10.5281/zenodo.18892604.

References

[1] Zeng, Y., et al. *AIR-Bench 2024: A Safety Benchmark Based on Risk Categories from Regulations and Policies*. arXiv:2407.17436, 2024. License: CC-BY-4.0.

[2] Liang, P., et al. *Holistic Evaluation of Language Models*. Transactions on Machine Learning Research, 2023.

[3] Atienza Jiménez, A. *OBSERVATORY: An Axiomatic Epistemic Framework for Computational Evidence — Core Theory (v0.2)*. Zenodo, 2026. DOI: 10.5281/zenodo.18892604.

[4] European Parliament and Council. *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act)*. Official Journal of the European Union, OJ L, 12.7.2024.

A Uniqueness Proof for This Setting

We show that any function $f : \text{Claims} \to \Sigma$ satisfying axioms (i)–(v) from Theorem 1 must agree with Eval_X on the nine claim schemas defined in Section 4.2. The argument instantiates the general uniqueness theorem [3, T5] on the concrete data of this paper: eight worlds, four equivalence classes, nine claim schemas. The evaluation function operates on the quotient space $\Omega^*(X)$ (four atoms). We write $|\Omega(X)| = 8$ when referring to the full world count and $|\Omega^*(X)| = 4$ when referring to the atom count.

Proof. Let φ be any of the nine claim schemas and let f satisfy axioms (i)–(v). By axiom (v), $f(\varphi)$ is exactly one element of $\Sigma = \{\text{ASSERTED}, \text{REJECTED}, \text{UNKNOWN}, \text{INDETERMINATE}, \text{UNSUPPORTED}\}$. We proceed by exhaustive case analysis over the support set S_φ. **Case 1: Inadmissible claim schemas** (`audit_completeness`, `deployment_readiness`). These schemas cannot be grounded in $\Omega(X)$: the first lacks a finite witness condition and the second lacks a declared definition. Since the schemas are inadmissible, the grounding requirements of axioms (i) and (ii) preclude assigning ASSERTED, REJECTED, or INDETERMINATE, each of which would endorse a conclusion not grounded in available evidence. UNKNOWN applies to admissible claims that cannot be evaluated, not to inadmissible schemas. By elimination: $f(\varphi) = \text{UNSUPPORTED} = \text{Eval}_X(\varphi)$. **Case 2:** $S_\varphi = \Omega^*(X)$ (`ambiguity_below_threshold`: true in all four atoms). REJECTED contradicts the evidence (violates (i)). INDETERMINATE asserts disagreement where none exists—all worlds agree, so declaring disagreement contradicts the established evidence, violating (i). UNKNOWN ignores an established, determinate result (violates (i)). UNSUPPORTED misrepresents admissibility. By elimination: $f(\varphi) = \text{ASSERTED} = \text{Eval}_X(\varphi)$. **Case 3:** $S_\varphi = \emptyset$ (`toxicity_refusal`, `uniform_safety`, `cross_category_consistency`: false in all four atoms). Symmetric to Case 2: ASSERTED contradicts the evidence (violates (i)); INDETERMINATE asserts disagreement where all worlds agree on falsehood (violates (i)); UNKNOWN ignores a determinate result (violates (i)); UNSUPPORTED misrepresents admissibility. By elimination: $f(\varphi) = \text{REJECTED} = \text{Eval}_X(\varphi)$. **Case 4:** $\emptyset \subsetneq S_\varphi \subsetneq \Omega^*(X)$ (`eu_compliance`, `majority_safe`, `category_coverage`). We give the full argument for `eu_compliance`; the other two are identical in structure. For this claim, $S_\varphi = \{\delta\}$ and $\Omega^*(X) = \{\alpha, \beta, \gamma, \delta\}$.

(a) ASSERTED requires $S_\varphi = \Omega^*(X)$. Here $S_\varphi = \{\delta\} \neq \Omega^*(X)$. Returning ASSERTED would privilege δ over $\{\alpha, \beta, \gamma\}$—worlds the evidence does not distinguish—violating axiom (iii).

(b) REJECTED requires $S_\varphi = \emptyset$. Here $S_\varphi = \{\delta\} \neq \emptyset$. Returning REJECTED would privilege $\{\alpha, \beta, \gamma\}$ over δ, violating axiom (iii).

(c) UNKNOWN requires $\Omega(X) = \emptyset$. Here $|\Omega(X)| = 8$. The evidence establishes a determinate (mixed) truth distribution; ignoring it violates axiom (i).

(d) UNSUPPORTED requires inadmissibility. The claim is a well-formed conjunction of two admissible predicates over declared structure; it is admissible.

Four states are eliminated. By axiom (v), exactly one state must be assigned. Therefore:

$$f(\texttt{eu\_compliance}) = \text{INDETERMINATE} = \text{Eval}_X(\texttt{eu\_compliance}).$$

The same argument applies to `majority_safe` ($S_\varphi = \{\gamma, \delta\}$; neither empty nor $\Omega^*(X)$) and `category_coverage` ($S_\varphi = \{\beta, \delta\}$; neither empty nor $\Omega^*(X)$) by substituting the appropriate support sets. Since $f(\varphi) = \text{Eval}_X(\varphi)$ for all nine claim schemas, Eval_X is the unique evaluation function for this setting. □