The Eastasouth Journal of Information System and Computer Science
Vol. 2, No. 03, April 2025, pp. 428-438
ISSN: 3025-566X, DOI: 10.58812/esiscs.v2i03

AI Governance: Automatically Checking AI Systems Against the EU AI Act and NIST Rules

Gopinathan Rathinavelu
Data Architect, Arizona, USA

Article Info

Article history:
Received Apr, 2025
Revised Apr, 2025
Accepted Apr, 2025

Keywords:
AI Governance;
Auditability;
EU AI Act;
Evidence-Linked Explanations;
Faithfulness;
Model Risk Management;
NIST AI RMF

ABSTRACT

Financial institutions deploying models for credit evaluation, fraud detection, and anti-money laundering must demonstrate that their internal documentation satisfies frameworks such as the European Union's Artificial Intelligence Act (EU AI Act) and the NIST AI Risk Management Framework (AI RMF). Automating this documentation review is attractive, but a checker that approves a control merely because a document mentions the right topic keywords is dangerous: it converts empty corporate assurance into apparent compliance. We present a system design and evaluation for evidence-linked compliance checking, with three contributions. First, a structured control catalog of twelve documentation-checkable controls mapped to EU AI Act articles and NIST AI RMF functions, each with explicit substance criteria. Second, an evidence-linked checker architecture whose every "satisfied" decision is tied to an exact supporting sentence, supported by a tiered decision pipeline, calibrated abstention, and a reviewer workflow. Third, a controlled synthetic benchmark of financial AI documentation in which each control is substantive, boilerplate-only, or absent, with disjoint training and unseen phrasing sets. In our baseline experiments, keyword rules falsely approve 85% of boilerplate-only controls. Trained sparse lexical classifiers reach high precision but, under unseen phrasing, recall falls to between 0.12 and 0.22. Adding 25 labeled documents in the new phrasing restores recall above 0.99, an estimate we caution is optimistic given the synthetic setting. Removal-based deletion tests show that evidence-faithfulness metrics must be interpreted against the redundancy of evidence within a document. The benchmark is a measurement tool for checker reliability, not a legal determination of conformity.

This is an open access article under the CC BY-SA license.

Corresponding Author:
Name: Gopinathan Rathinavelu
Institution: Data Architect, Arizona, USA
Email: gopinathan.r@gmail.com

1. INTRODUCTION

Regulation (EU) 2024/1689 (the EU AI Act) [1] places obligations on high-risk AI systems for risk management, data governance, technical documentation, record-keeping, transparency, human oversight, accuracy, robustness, cybersecurity, and post-market monitoring. The NIST AI Risk Management Framework (AI RMF) [2] organizes comparable practices under four functions: GOVERN, MAP, MEASURE, and MANAGE. In financial services, credit scoring of natural persons is treated as high-risk under the EU AI Act, while the applicability of fraud-detection systems depends on their stated purpose. These obligations overlap with supervisory expectations such as the

Federal Reserve and OCC guidance on model risk management, SR 11-7 [3].

Reviewing model documentation against these overlapping standards is still largely manual. Language models and retrieval systems promise relief, but evaluating them is hard for three reasons: real documents are confidential, expert annotation is expensive, and automated checkers are easily misled by vacuous assurance language such as "we take security very seriously." A checker that is fooled by such language is worse than no checker, because it produces confident, wrong approvals that reviewers may stop questioning.

This paper therefore treats the central question as one of **measurement and architecture**: how should a compliance checker be built so that its approvals are trustworthy, auditable, and testable, and how can its failure modes be quantified?

Contributions.

(1) **Control catalog.** Twelve documentation-checkable controls mapped to EU AI Act articles and NIST AI RMF functions, with explicit, machine-readable substance criteria (Section 3.1).

(2) **Evidence-linked checker architecture.** A tiered pipeline from document ingestion to reviewer workspace in which every approval cites verifiable evidence, with calibration, abstention, and audit logging (Section 3.2).

(3) **Controlled synthetic benchmark.** Three document states (substantive, boilerplate, absent) with controlled phrasing shifts and ground truth by construction (Section 3.3).

(4) **Baseline evaluation.** Keyword rules and sparse lexical classifiers assessed for performance, false assurance, evidence localization, and removal-based faithfulness (Sections 3.4 and 4).

(5) **A negative result.** Simple lexical methods do not transfer across phrasing shifts without adaptation, which motivates the architectural choices in Section 3.2 (Section 4.4).

2. LITERATURE REVIEW

2.1 Documentation Standards

Model Cards [4] and Datasheets for Datasets [5] standardize what documentation should contain, and internal algorithmic auditing frameworks describe end-to-end review lifecycles [6]. They define the content to look for but not how to verify automatically that a document actually provides it.

2.2 Faithfulness of Rationales

Removal-based tests, including comprehensiveness and sufficiency as in the ERASER benchmark [7], check whether a cited rationale actually drives a prediction, although how faithfulness should be defined remains debated [8]. We adapt sentence-removal tests to evidence sentences and show where they must be corrected for redundancy.

2.3 Frameworks

The NIST Generative AI Profile (NIST AI 600-1) [9] extends the RMF to generative systems. Generative-specific controls are outside the scope of this evaluation.

2.4 Research Gap

Prior work provides content standards and explanation-evaluation methods, but not a control-level, evidence-linked compliance checking design with a benchmark that separates substantive evidence from boilerplate.

3. METHODS

3.1 Control Catalog

a. Substance Criteria

A control is **satisfied** only if the documentation contains verifiable evidence. We define evidence as a sentence that contains at least one **substance attribute**:

1) a specific numeric threshold or metric value,
2) a named operational mechanism or tool,
3) an execution frequency or schedule,
4) a designated role or owner, or

5) a named technical artifact (for example a report, log, or test suite).

Statements of intent or corporate values are classified as **boilerplate** and are non-satisfying. Formally, a sentence s supports control c if it is on-topic for c and carries at least one substance attribute; a document satisfies c if at least one sentence supports it. The checker's job is to find that sentence or to determine that none exists.

b. The Twelve Controls

Table 1. Control Catalog (Documentation-Checkable Subset).

ID	Control	EU AI Act	NIST AI RMF
RISK	Risk management	Art. 9	GOVERN 1.1
DATA	Data governance	Art. 10	MAP
DOC	Technical documentation	Art. 11, Annex IV	GOVERN
LOG	Event logging	Art. 12	MANAGE
TRANS	Transparency	Art. 13	MEASURE
OVERSIGHT	Human oversight	Art. 14	GOVERN 3.2
ACC	Accuracy measurement	Art. 15	MEASURE 2.5
ROBUST	Robustness testing	Art. 15	MEASURE
SEC	Cybersecurity	Art. 15	MEASURE 2.7
BIAS	Bias examination	Art. 10(2)	MEASURE 2.11
EXPL	Explainability	Art. 13	MEASURE 2.9
PMM	Post-market monitoring	Art. 72	MANAGE 4.1

Note. EU AI Act = Regulation (EU) 2024/1689; NIST AI RMF = NIST Artificial Intelligence Risk Management Framework.

c. Catalog as Versioned Data

Each control is stored as a versioned, machine-readable record rather than embedded in code, so that legal and risk reviewers can change it without redeploying the checker:

```
id: LOG
name: Event logging
version: 1.3
mappings:
  eu_ai_act: ["Art. 12"]
  nist_ai_rmf: ["MANAGE"]
  supervisory: ["SR 11-7: documentation"]
substance:
  any_of: [numeric_threshold, named_mechanism, frequency, role, artifact]
  min_supporting_sentences: 1
topic_terms: ["log", "audit trail", "event record", "retention"]
examples:
  substantive: "Inference requests are logged with model version and input hash, retained 2
4 months."
  boilerplate: "We maintain comprehensive logging practices."
source_text_version: "OJ L, 12 July 2024"
review_status: "approved"
reviewed_by: "legal-compliance"
effective_from: "2025-01-01"
```

The catalog carries source_text_version and review_status so that regulatory updates trigger a re-verification workflow, and every decision records the catalog version it used.

3.2 *System Architecture*

a. Overview

The checker is a pipeline with a feedback loop (Figure 1). Boxes marked *(proposed)* were not part of the reported experiments.

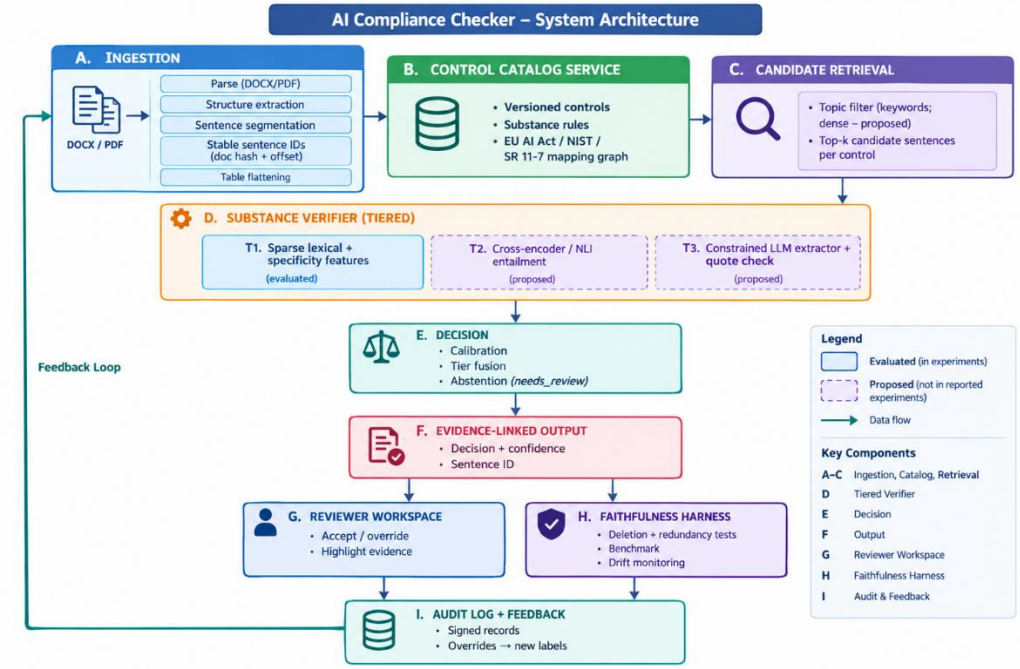


Figure 1. System Architecture of the Evidence-Linked Compliance Checker. Components A–I are described in Sections 3.2.2–3.2.8; dashed boxes denote proposed components that were not part of the reported experiments.

b. Ingestion and Segmentation (A)

Documents arrive as DOCX or PDF. The ingester extracts text with structure (headings, lists, tables) and splits it into sentences. Three design points matter for auditability:

1) **Stable evidence addresses.** Each sentence receives an ID derived from the document hash and its character offsets, so the evidence a decision cites can be located and highlighted exactly, and remains valid if the checker is rerun.
2) **Structure preservation.** Section headings are retained as context features, since a sentence under "Human Oversight" carries different weight than the same sentence in an appendix.
3) **Table flattening.** Table cells are converted to sentence-like units with header context, so evidence in tables is not lost. This is an engineering requirement for real documents even though the synthetic benchmark contains only prose.

c. Candidate Retrieval and Substance Verification (C and D)

Retrieval narrows each document to a small candidate set for each control, and verification decides whether any candidate carries substance. Separating these stages is a deliberate design choice: **topic detection and substance verification are different tasks**, and keyword rules fail precisely because they conflate them. A keyword hit says a sentence is on-topic. It says nothing about whether the sentence says anything.

Tier T1 (evaluated): sparse lexical classifier. For each control, a logistic regression is trained at the sentence level on TF-IDF features (word 1 to 2 grams or character 3 to 5 grams). A variant adds five structural specificity features: digit count, percent-sign count, number count,

acronym count, and sentence length. A document satisfies a control if the maximum sentence probability exceeds 0.5, and the arg-max sentence is the cited evidence.

Tier T2 (proposed): entailment-based verifier. A cross-encoder or natural-language-inference model [10] scores whether a candidate sentence entails a substance hypothesis built from the control's criteria (for example, "This sentence states a specific frequency or mechanism for event logging"). Dense sentence embeddings [11] improve candidate recall under paraphrase, addressing the brittleness we measure in Section 4.

Tier T3 (proposed): constrained extractor with quote verification. A language model is prompted with the control's criteria and the candidate sentences and must return a structured object naming the substance attributes it found. Two safeguards keep this tier auditable. First, the output schema is constrained, so the model cannot return free-form approval. Second, a deterministic **quote check** confirms that every cited span appears verbatim in the document, so a hallucinated evidence sentence is rejected automatically. The model contributes recall under paraphrase, while the verification step prevents it from inventing evidence.

d. Decision, Calibration, and Abstention (E)

The cost of errors is asymmetric: a **false assurance** (approving a boilerplate or absent control) is more harmful than a **false alarm** (sending a valid control to review). The decision layer reflects this:

1) **Calibration.** Per-control scores are calibrated (for example with temperature scaling or isotonic regression, as studied for modern neural networks by Guo et al. [12]) so that a confidence value has a stable meaning.
2) **Threshold selection.** The approval threshold for each control is chosen to satisfy a target false-assurance ceiling on validation data, not to maximize F1.
3) **Tier fusion.** Approval requires agreement across enabled tiers or a verified-evidence condition from T3. Disagreement yields needs_review.
4) **Abstention.** The checker may return needs_review instead of a binary decision, a selective-classification setup [13] in which coverage is traded against error. Our results show why: under unseen phrasing the lexical tier is precise but misses most valid evidence, so its non-approvals should be treated as *unknown*, not as *not satisfied*.

The output label set is therefore satisfied, not_satisfied, and needs_review.

e. Evidence-Linked Output Contract (F)

For every document and control the checker returns a decision, a confidence score, and the best-supporting evidence sentence, so that a reviewer can audit or override the decision by reading a single sentence. A fuller record includes:

```
{
 "doc_hash": "sha256:7ac1...",
 "control_id": "OVERSIGHT",
 "decision": "satisfied",
 "confidence": 0.93,
 "evidence": {
  "sentence_id": "s_0147",
  "text": "The fraud operations lead reviews all model-
blocked payments above 10,000 USD within 4 hours.",
  "section": "Operational Controls",
  "substance": ["role", "numeric_threshold", "frequency"]
 },
 "redundancy": {"other_supporting_sentences": 0},
 "tier_votes": {"T1": "satisfied"},
 "catalog_version": "1.3",
 "model_version": "lex-word-2025-01",
 "mapping": {"eu_ai_act": ["Art. 14"], "nist_ai_rmf": ["GOVERN 3.2"]}
}
```


The redundancy field records how many additional sentences independently support the control. It makes deletion-test results interpretable (Section 3.4) and tells a reviewer how fragile an approval is: a decision resting on a single sentence deserves closer inspection.

f. Faithfulness and Monitoring Harness (H)

The harness runs deletion tests and benchmark evaluations continuously. Because the checker is itself a model used in a governance process, it should be validated, monitored, and change-controlled under the same model-risk expectations it helps document. Concretely:

1) **Re-evaluation on change.** Any change to the catalog, a tier, a threshold, or the training data triggers re-running the benchmark and the deletion tests, with false-assurance rates as gating metrics.
2) **Drift monitoring.** In production, the distributions of confidence scores, needs_review rates, and reviewer override rates are tracked. A rising override rate on a control signals phrasing drift or catalog misalignment.
3) **Held-out phrasing sets.** Evaluation always includes a phrasing set disjoint from training, since in-distribution scores mainly measure memorization (Section 4).

g. Reviewer Workspace and Feedback Loop (G and I)

The reviewer sees each control with the evidence sentence highlighted in context, the confidence, the redundancy count, and the mapped articles. Accepting or overriding a decision requires one click and an optional comment. Overrides are stored as labeled examples, and the adaptation result in Section 4 indicates how quickly they can pay off: a few dozen labeled documents in a new phrasing style substantially restore recall in our synthetic setting. Each decision, override, catalog version, and model version is appended to a signed audit log keyed by document hash, so that a later examiner can reproduce why a control was marked satisfied on a given date.

h. Mapping Graph and Cross-Framework Reuse (B)

Because one piece of evidence often satisfies several frameworks, the catalog stores mappings as a graph from controls to

EU AI Act articles, NIST functions and subcategories, and supervisory guidance such as SR 11-7. A single verified evidence sentence for *Post-market monitoring* can therefore be reported against Art. 72 and MANAGE 4.1 without re-checking. Mapping edges carry the same version and review metadata as controls.

3.3 Synthetic Benchmark Construction

Evaluating compliance checkers on real financial documentation is restricted by confidentiality and the cost of expert labeling. We therefore construct a synthetic benchmark of documents of 20 to 35 sentences describing financial AI systems.

Document states. For each of the twelve controls, a state is drawn independently:

1) **Substantive (50%):** one or two factual evidence sentences.
2) **Boilerplate (22%):** one vacuous sentence that uses topic keywords without substantive backing.
3) **Absent (28%):** no mention of the control.

Each document also contains 12 to 21 distractor sentences describing unrelated business facts.

Templates and phrasing sets. Substantive sentences are generated from four templates per control with variable slots for metrics, roles, and tools. **Phrasing Set A** is used for training and in-distribution evaluation. **Phrasing Set B** introduces distinct wording to test generalization ("unseen phrasing").

Splits. Ground truth is fully defined by construction. The benchmark has 500 training documents, 250 in-distribution test documents, and 250 unseen-phrasing test documents, repeated over five random seeds.

Generator design. The generator is a seeded, template-driven program that writes both the document and its gold labels (state per control and the indices of gold evidence sentences), so labels are exact and the benchmark can be regenerated or extended with new templates, new phrasing sets, or different state probabilities.

3.4 Checkers and Metrics

Keyword rules. Static lists of terms per control. A control is marked satisfied if any keyword appears anywhere in the document.

Lexical classifiers. The T1 models of Section 3.2.3: per-control logistic regression on sentence-level TF-IDF features (word or character n-grams), with and without the five specificity features.

Metrics.

1) **Performance:** F1, precision, and recall for the "satisfied" class.
2) **False-assurance rates:** FA-B, the fraction of boilerplate-only controls incorrectly approved, and FA-A, the fraction of absent controls incorrectly approved.
3) **Evidence localization:** among correctly approved controls, the fraction whose cited sentence matches gold evidence.
4) **Faithfulness (deletion flip rate):** the proportion of approved decisions that flip to rejection when the cited evidence sentence is removed, compared against removing a random distractor sentence.

For the deletion test, **the flip rate depends on redundancy**. When a document has two independent evidence sentences for a control, removing one should not flip a correctly aggregating checker. We therefore interpret flip rates against the redundancy of evidence (Section 4.3).

4. RESULTS AND DISCUSSION

4.1 False Assurance and Phrasing Brittleness

Keyword rules reach an F1 of 0.683 under training-like conditions, but they incorrectly approve **85%** of boilerplate-only controls (FA-B = 0.85) and **29%** of absent controls (FA-A = 0.29). Under unseen phrasing their F1 falls to 0.445.

Sparse lexical classifiers are perfect when test phrasing matches training, which reflects template memorization. Under unseen phrasing, recall collapses to **0.12** (word model), **0.20** (character model), and **0.22** (specificity-enhanced model). Precision stays high at 0.99 to 1.00, which prevents false approvals, but low recall means the checker sends most valid documents to manual review, which limits operational value.

Table 2. Checker Performance Across Training and Unseen Phrasings.

Method	F1	Precision	Recall	FA-B	FA-A
Same phrasing (sanity check)					
Keyword rules	0.683 ± 0.008	0.60	0.80	0.85	0.29
TF-IDF word LR	1.000 ± 0.000	1.00	1.00	0.00	0.00
TF-IDF + spec. LR	1.000 ± 0.000	1.00	1.00	0.00	0.00
Unseen phrasing					
Keyword rules	0.445 ± 0.008	0.46	0.43	0.84	0.21
TF-IDF word LR	0.209 ± 0.020	1.00	0.12	0.00	0.00
TF-IDF + spec. LR	0.365 ± 0.015	0.99	0.22	0.00	0.00

Note. Values are mean ± s.d. over five seeds. LR = logistic regression; spec. = specificity features; FA-B = false-assurance rate on boilerplate-only controls; FA-A = false-assurance rate on absent controls.

4.2 Domain Adaptation

Adding a small number of labeled documents in the new phrasing to the training set restores recall. With just **25 added documents**, recall rises from 0.125 to 0.993 for the word-level model and from 0.223 to 0.991 for the specificity-augmented model. This is an optimistic estimate: the synthetic data has few phrasing templates, and real compliance documents have far higher linguistic variance.

Table 3. Adaptation Results Under Unseen Phrasing Shift.

Model	Added docs	Recall	Precision	FA-B
TF-IDF (word) LR	0	0.125 ± 0.019	1.000	0.000
TF-IDF (word) LR	25	0.993 ± 0.004	1.000	0.000
TF-IDF + spec. LR	0	0.223 ± 0.023	0.984	0.004
TF-IDF + spec. LR	25	0.991 ± 0.004	0.998	0.002

Note. Added docs = number of labeled documents in the new phrasing added to the training set. LR = logistic regression; spec. = specificity features; FA-B = false-assurance rate on boilerplate-only controls.

4.3 Evidence Faithfulness and Redundancy

Removing a random distractor sentence never changes a compliance decision. Removing the cited evidence sentence flips the decision for **49%** of approvals (flip rate 0.49) for lexical classifiers on familiar phrasing. This number is shaped by redundancy: benchmark documents contain two independent evidence sentences 50% of the time, so deleting one does not always flip the outcome. Under unseen phrasing, flip rates rise to 0.80 to 0.94, because the models then tend to recognize only a single supporting sentence. Removal-based tests must therefore be interpreted together with how many independent evidence sentences a document contains.

4.4 Discussion

Keyword matching measures topic, not substance. The 85% boilerplate approval rate shows that topic words appear in empty assurance sentences almost as readily as in substantive ones. Any deployed checker that uses keywords to *approve* will systematically over-assure. Keywords remain useful for *candidate retrieval* (Section 3.2.3), where the downstream verifier carries the approval decision.

Perfect in-distribution scores are a warning, not a result. Perfect scores under matching phrasing indicate template memorization. Without the unseen-phrasing split, this checker would look flawless while being unusable on new text. Held-out phrasing evaluation should be mandatory for any compliance checker.

Precision without recall pushes work back to humans. The lexical models rarely approve wrongly, but they miss most valid evidence under shift. In operation, such a checker degrades into a pre-filter that flags nearly everything. This is the reason the architecture treats non-approval under uncertainty as needs_review and invests in tiers that raise recall under paraphrase.

Cheap adaptation helps, with caution. The jump to recall above 0.99 after 25 labeled documents shows that reviewer overrides can be valuable training signal. Because the synthetic setting has few templates, real-world adaptation will need more data, and it should be measured on genuinely new phrasing.

Faithfulness needs redundancy-aware reading. A deletion flip rate of 0.49 is not evidence of poor faithfulness by itself. It must be read against the share of documents with redundant evidence. We recommend reporting flip rates stratified by the number of independent gold evidence sentences, and adding an all-evidence deletion test (remove every gold evidence sentence, expect a flip rate near 1).

4.5 Limitations

(1) **Synthetic ground truth.** The benchmark measures false assurance and phrasing brittleness in a controlled way, but it does not establish performance on uncurated, confidential corporate documents. Real documents have longer context, tables, cross-references, and ambiguous statements.

(2) **Evaluated scope.** Only keyword rules and sparse lexical models were evaluated. The entailment tier, dense retrieval, and the constrained LLM extractor in Section 3.2 are proposed designs, and their benefits are hypotheses until tested.

(3) **Template diversity.** Four templates per control and two phrasing sets understate natural variation.

(4) **Binary substance definition.** Real evidence is graded, and a sentence may be partly specific. The catalog's criteria would need graded labels and expert agreement studies.

(5) **Regulatory boundaries.** The catalog is an internal documentation-review and pre-filtering aid, not a legal conformity assessment. It omits registration obligations, fundamental-rights impact assessments, and deployer duties. Regulations are subject to legislative updates and must be re-verified against consolidated texts.

(6) **Documentation is not behavior.** A document may describe a control that is not implemented. Checking documentation is necessary for governance but not sufficient to show that the system behaves as described.

4.6 Future Work

The following experiments would test the proposed tiers. No results are reported here.

1) **F1. Dense retrieval and NLI tier.** Compare T2 with T1 on unseen phrasing, reporting recall, FA-B, and FA-A at a fixed false-assurance ceiling.
2) **F2. Constrained LLM extractor with quote verification.** Measure recall, hallucinated-evidence rate before and after quote checking, and cost per document.
3) **F3. Calibration and abstention.** Plot coverage versus false-assurance curves with and without calibration, for each tier and for fusion.
4) **F4. Redundancy-stratified faithfulness.** Report deletion flip rates by gold-evidence count, plus the all-evidence deletion test.
5) **F5. Richer phrasing shifts.** Expand to many phrasing sets, including adversarial boilerplate that imitates

substantive structure with hollow content.

6) **F6. Real redacted documents.** Evaluate on expert-labeled, redacted model documentation, with inter-annotator agreement.
7) **F7. Reviewer study.** Measure review time and error rates with and without evidence links and abstention.

5. CONCLUSION

Automated auditing against the EU AI Act and NIST AI RMF requires checkers that separate substantive engineering evidence from empty assurance language, and benchmarks that expose when they fail. Our experiments show that keyword rules generate unacceptable false assurance, and that sparse lexical models are precise but brittle under phrasing shift. We describe an evidence-linked architecture, with a versioned control catalog, a tiered substance verifier, calibrated abstention, redundancy-aware faithfulness tests, and a reviewer feedback loop, that turns these findings into design requirements. Evidence-linked outputs and sentence-removal tests provide the auditability this domain needs, provided evidence redundancy is accounted for. The benchmark and architecture support measuring and improving checker reliability, and they remain a pre-filtering aid rather than a substitute for legal judgment.

ACKNOWLEDGEMENTS

The author received no financial support for the research, authorship, and/or publication of this article. The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

REFERENCES

[1] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," *Off. J. Eur. Union*, L series, Jul. 12, 2024. [Online]. Available: http://data.europa.eu/eli/reg/2024/1689/oj

[2] National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1. U.S. Dept. of Commerce, 2023, doi: 10.6028/NIST.AI.100-1.

[3] Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, *Supervisory Guidance on Model Risk Management*, SR Letter 11-7, 2011.

[4] M. Mitchell *et al.*, "Model cards for model reporting," in *Proc. Conf. Fairness, Accountability, and Transparency (FAT∗'19)*, 2019, pp. 220–229, doi: 10.1145/3287560.3287596.

[5] T. Gebru *et al.*, "Datasheets for datasets," *Commun. ACM*, vol. 64, no. 12, pp. 86–92, 2021, doi: 10.1145/3458723.

[6] I. D. Raji *et al.*, "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proc. Conf. Fairness, Accountability, and Transparency (FAT∗'20)*, 2020, pp. 33–44, doi: 10.1145/3351095.3372873.

[7] J. DeYoung *et al.*, "ERASER: A benchmark to evaluate rationalized NLP models," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2020, pp. 4443–4458, doi: 10.18653/v1/2020.acl-main.408.

[8] A. Jacovi and Y. Goldberg, "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2020, pp. 4198–4205, doi: 10.18653/v1/2020.acl-main.386.

[9] National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, NIST AI 600-1. U.S. Dept. of Commerce, 2024, doi: 10.6028/NIST.AI.600-1.

[10] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proc. 2018 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT)*, vol. 1, 2018, pp. 1112–1122, doi: 10.18653/v1/N18-1101.

[11] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. 2019 Conf. Empirical Methods Natural Lang. Process. and 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3982–3992, doi: 10.18653/v1/D19-1410.

[12] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, PMLR vol. 70, 2017, pp. 1321–1330.

[13] R. El-Yaniv and Y. Wiener, "On the foundations of noise-free selective classification," *J. Mach. Learn. Res.*, vol. 11, pp. 1605–1641, 2010.