Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Risk-constrained human-in-the-loop repair of anti-money laundering policies: Temporal legal graphs, minimal revision, and finite-sample assurance

Joscha Krause

j.krause@cure-intelligence.com

CURE S.A.

Cathrine Aeckerle-Willems

CURE S.A.

Gerti Sana

CURE S.A.

Stéphane Cortina

Luxembourg Institute of Science and Technology

Yuejun Guo

Luxembourg Institute of Science and Technology

Orlando Amaral Cejas

Luxembourg Institute of Science and Technology

Djamel Khadraoui

Luxembourg Institute of Science and Technology

Moritz von Kunow

Governance Tailor SARL

Research Article

Keywords:

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11119310/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Risk-constrained human-in-the-loop repair of anti-money laundering policies: Temporal legal graphs, minimal revision, and finite-sample assurance

Anonymous author(s) *

Abstract

This paper addresses human-in-the-loop compliance automation for obliged entities when the anti-money laundering legal corpus is amended. We treat the updating of an entity's policy as a minimal repair problem over two graphs, a versioned legal graph and a clause-and-control policy graph. The semantic difference between legal versions yields an impact cone of affected obligations, and we search for a minimally disruptive set of policy edits that satisfies them. An agentic workflow drafts, criticizes, and verifies these edits, while a learned orchestration policy decides how much machine computation and which human expert to use on each, subject to a chosen ceiling on the probability of missing a material obligation. We show that under explicit delta completeness and locality properties of the encoding, the impact cone contains every conservatively applicable obligation whose encoded legal evaluation can change across the amendment. We separately quantify the failure probability when dependency edges are missing. Moreover, minimal repair is NP-hard, and its pure coverage restriction has a logarithmic inapproximability threshold unless P = NP. A greedy method attains a logarithmic guarantee for integer-valued submodular coverage with positive additive costs. Conflicts and unresolved feasibility require separate handling and possible escalation. Finally, we show that machine acceptance can be calibrated in a distribution-free, finite-sample way, and the component guarantees compose into a single end-to-end bound on the probability of a material miss.

Keywords: temporal knowledge graphs, learning to defer, conformal risk control, submodular optimization, multi-agent systems, distribution-free calibration

Funding: No comment.

*Contact: anonymous contact details

1 Introduction

Anti-money laundering and countering the financing of terrorism (AML/CFT) is an umbrella term for a set of national and international laws and rules (legal corpus) that are designed to stop criminals from hiding illegal money as well as to block funds from reaching terrorist groups. Within the European Union (EU), AML/CFT requirements apply to financial institutions and specified non-financial businesses, including certain art market participants, subject to the relevant scope conditions and thresholds. Among other duties, an obliged entity must maintain an AML policy that states and specifies how the entity complies with these regulations. Depending on applicability, an amendment of the legal corpus requires the timely updating of the entity's policy to ensure compliance.

In practice, policy maintenance is a change propagation problem under legal and epistemic uncertainty. An obliged entity does not respond to a regulatory amendment simply by writing new words in the policy document. Rather, it must determine which provisions are applicable, when they take effect, which adjacent entities or activities they govern, which existing clauses and controls depend on them, and which policy changes are sufficient without being overly disruptive. The current regulatory landscape of the EU makes this particularly visible. Regulation (EU) 2024/1624 introduces a new AML/CFT rulebook that is designed to harmonize the European landscape. The Regulation entered into force on 9 July 2024 and generally applies from 10 July 2027. Its application to football agents and professional football clubs is deferred to 10 July 2029, subject to the scope and exemptions in Articles 3 and 5. Directive (EU) 2024/1640 has a general transposition deadline of 10 July 2027, with different deadlines for specified provisions. The resulting framework combines directly applicable EU rules, national implementing legislation, supervisory instruments, and transitional provisions (European Parliament and Council of the European Union, 2024b,a). The relevant legal corpus therefore spans multiple sources and changes over time.

Technology has long supported AML/CFT compliance, e.g., via the detection of suspicious transactions (Chen et al., 2018). Likewise, the Financial Action Task Force has identified substantial opportunities for new technologies to improve both efficiency and effectiveness of AML/CFT (Financial Action Task Force, 2021). However, policy maintenance poses a different problem. Its key risk is not a false transaction alert, but the omission, mistranslation, or false timing of a legal obligation in the entity's policy and control framework. A system may retrieve a relevant passage and still fail because it applies the wrong version, overlooks an exception, loses a dependency, edits an unrelated clause, or accepts a draft without sufficient evidence. Hence, retrieval accuracy alone does not constitute an assurance argument for automated policy change.

This paper develops a human-in-the-loop framework for the AI-supported updating and repairing of AML policies. The legal corpus is represented as a versioned, typed property graph, while the entity's policy is treated as a clause-and-control graph. Knowledge time (point in time when the system registers a source) is separated from legal evaluation time (point in time when a norm is applicable). Applicability and policy compliance are three-valued, such that missing entity facts produce an unresolved value rather than a silent exclusion. A semantic graph delta seeds a dependency closure (impact cone) that identifies obligations whose legal evaluation may have changed. The resulting obligations define a constrained minimal repair problem over a finite library of admissible policy edits.

Within the system, the role of machine learning is limited by design. Authenticated sources, adjudicated interpretations, validity intervals, and hard compliance constraints form a normative layer that operational learning cannot alter directly. Learning occurs only on the operational layer, that is, for retrieval effort, verification depth, drafting strategy, agent selection, expert routing, and acceptance thresholds. This separation prevents the model optimization from converting a legal proposition into a popularity-weighted estimate. Further, it facilitates the division

between propositions depending on the correctness of the legal encoding on the one hand, and statistical guarantees depending on calibration data on the other hand. The repair process is then treated as a sequential decision problem. A workflow may retrieve further authority, expand the legal neighborhood, require an independent critique, revise a proposed edit, consult a specialist, or accept only after the relevant hard and risk gates pass. Corresponding choices trade expected compliance loss against human cost, machine cost, delay, and documentation disruption. As the legally preferable interpretation and the practical adequacy of a draft are only partially observed, the orchestration problem is formulated as a finite-horizon constrained partially observable Markov decision process (POMDP). A chance constraint controls the probability of a material miss, while a conditional value-at-risk (CVaR) constraint protects against low-frequency, high-severity failures. Human review remains an action available to the policy, rather than an informal safeguard applied after optimization. Therefore, the paper makes five contributions.

1. It gives a formal decision object for AML/CFT policy maintenance: a versioned legal graph, an entity-specific three-valued applicability semantics, a clause-and-control policy graph, and a separation principle that limits learning to operational choices.

2. It defines amendment impact as dependency reachability from a semantic delta. Under explicit delta completeness and evaluation locality conditions, the impact cone contains every conservatively applicable norm whose legal evaluation can change. A separate probabilistic result quantifies the chance of omission when dependency edges are imperfectly captured.

3. It formulates policy updating as minimal constrained repair. A minimum exists for every finite edit library with a feasible repair and a real-valued objective. Even the pure coverage decision problem is NP-complete, and its optimization version has a logarithmic inapproximability threshold unless P = NP. A greedy submodular cover attains the corresponding logarithmic guarantee. The analysis also isolates the additional effects of mutually exclusive edits and prerequisite closures.

4. It formalizes the agentic and human workflow as risk-constrained orchestration. The framework includes hard legal gates, an attainment result for finite horizon randomized policies, identifiable reviewer-specific preference models under randomized presentation order, and learning to defer rules that price expert error, delay, as well as consultation cost.

5. It connects structural, algorithmic, and statistical assurance. Machine acceptance is calibrated on held-out episodes with finite-sample risk control. Expert routing guarantees remain distinct from machine guarantees, and the component failure probabilities are combined into an explicit end-to-end bound on the probability of a material miss.

It is important to note that these contributions do not depend on an LLM being a reliable source of law. Generative models may propose graph elements, candidate repairs, critiques, or search actions, but every such output remains subordinate to provenance, typed dependencies, feasibility checks, and calibrated routing. The framework is therefore model-agnostic. Improvements in extraction, retrieval, and generation can reduce cost or enlarge the automatically acceptable region without altering the normative invariants.

The remainder of the paper is organized as follows. Section 2 summarizes related work in legal knowledge representation, regulatory compliance automation, change impact analysis, human and AI decision systems, and distribution-free risk control. Section 3 introduces the formal decision objects, the separation principle, the temporal legal graph, entity context, and conservative compliance semantics. Section 4 develops the semantic graph delta and the corresponding impact closure results. Section 5 studies minimal policy repair. Section 6 formalizes risk-constrained human and AI orchestration, structured feedback, and expert allocation. Section 7 develops safe adaptive learning, off-policy evaluation, calibrated acceptance, and finite-sample assurance.

Section 8 combines these components in the end-to-end updating algorithm, states the composed assurance result, and describes the audit record required for its application. Section 9 provides a practical framework for the future empirical evaluation of a complete prototype. Section 10 discusses the interpretation and limitations of the proposed system and the analysis performed in this paper. Section 11 concludes. Appendix A states the assumptions used by the mathematical results. Appendix B contains further theoretical results and the formal evaluation model. Appendix C contains the proofs.

2 Related work

In what follows, we summarize related work in the key research areas of the paper. The order reflects the system developed below. We begin with the allocation of machine and human work, then turn to the representation and propagation of legal change, and finally address the control of operational and statistical risk. Thereafter, we position the present paper in that landscape.

Agentic AI and human-AI allocation
Multi-agent LLM frameworks demonstrate how complex tasks can be decomposed among agents with different prompts, tools, and interaction patterns (Wu et al., 2024). Learning to defer research, on the other hand, elaborates on when a predictor should act and when it should pass a decision to a human expert. Consistent objectives and sample-efficient complementary predictors have been developed for that setting (Mozannar and Sontag, 2020; Charusaie et al., 2022). For repeated allocation among a fixed set of admissible review routes, our learning component is also related to stochastic multi-armed bandits. In particular, upper-confidence-bound algorithms admit logarithmic, gap-dependent finite-time regret under stationary bounded rewards (Auer et al., 2002). In our framework, this principle is applied only within the hard-gate-admissible route set. We combine these perspectives, but treat orchestration as a governed decision process. Retrieval, critique, revision, specialist consultation, and acceptance are actions with different information and cost consequences. Hard legal gates are not reward terms that can be offset by drafting quality, and expert routing is optimized against reviewer-specific accuracy and cost rather than invoked only when a model reports low confidence.

Legal NLP, knowledge graphs, and regulatory compliance
Legal NLP has progressed from document classification and information extraction to evaluations of legal reasoning, e.g., as in the LegalBench suite (Guha et al., 2023). Retrieval-augmented generation (RAG) grounds model output in an external corpus, and GraphRAG augments retrieval with graph-based indexing and aggregation for corpus-level questions (Edge et al., 2024). Formal business process compliance represents deontic requirements against possible execution traces (Governatori, 2013). Closer to our setting, regulatory knowledge graphs have been constructed from annotated rules and LLM outputs (Ershov, 2023). Recent requirements engineering work compares graphs derived from regulation with graphs derived from internal specifications (Falcarin et al., 2025). A closely related approach is RC4PC, which extracts atomic regulatory changes and assesses their effects on business-process compliance through a hybrid LLM and algorithmic pipeline (Barrientos et al., 2026). These contributions supply extraction, retrieval, formalization, and alignment mechanisms. However, our focus is complementary. The graph is not only a retrieval index, but the object on which temporal applicability, dependency closure, policy traceability, and repair guarantees are defined. Relative to change analysis for process compliance, we also study a conservative entity-specific obligation cone, minimal policy repair, risk-constrained human orchestration, and finite-sample acceptance. The question is not simply whether a system can retrieve or summarize an amended provision, but whether it can delimit the obligations that may have changed and account for those it cannot resolve.

Change impact and minimal repair
The impact cone is structurally related to program slicing and change impact analysis, where de-

pendency reachability delimits which computations may be affected by a change (Weiser, 1984). The repair problem also draws on classical weighted set cover and submodular cover. Greedy selection has a logarithmic approximation guarantee under diminishing returns (Chvátal, 1979; Wolsey, 1982), and the corresponding logarithmic hardness threshold limits what polynomial-time algorithms can guarantee in general (Dinur and Steurer, 2014). The contribution here is not a new set cover bound, but the reduction of legally constrained policy maintenance to a repair problem for which those bounds become interpretable. Obligations form coverage requirements, edits carry implementation and review cost, conflicts represent mutually exclusive drafting choices, prerequisites generate edit closures, and escalation preserves safe disposition when compliant automatic repair is unavailable.

Risk-sensitive control and finite-sample assurance
Constrained Markov decision processes provide the general language for optimizing expected performance subject to cost or safety constraints (Altman, 1999). Risk-constrained reinforcement learning has studied both chance constraints and CVaR constraints for rare, severe outcomes (Chow et al., 2018). Separately, risk-controlling prediction sets and Learn-then-Test calibration provide distribution-free finite-sample control for user-specified losses (Bates et al., 2021; Angelopoulos et al., 2025), while conformal risk control extends the conformal construction to general monotone losses (Angelopoulos et al., 2024). We use these tools at the acceptance boundary of a compliance workflow rather than as a substitute for legal semantics. Calibration controls the empirical decision to accept, defer, or escalate under stated exchangeability and shift assumptions. It cannot compensate for an omitted source or dependency. The end-to-end assurance argument therefore composes statistical risk with separate bounds for graph incompleteness and human review error.

Paper position
The systems most similar to ours automate one or more of regulatory extraction, graph construction, retrieval, requirements alignment, drafting, or human referral. This paper studies the connection among them. Our central object is an amendment-to-policy pipeline in which the normative state is versioned, the potentially affected set has a stated completeness condition, repair has a defined feasibility predicate and approximation behavior, orchestration is risk-constrained, and acceptance is calibrated. The intended result is not autonomous legal judgment, but a formally delimited division of labor. Machines search and propose within an auditable structure, statistical learning allocates effort, and unresolved or insufficiently assured decisions remain with accountable human experts.

3 Formal framework

This section introduces the formal objects on which the subsequent analysis operates. Section 3.1 establishes the separation principle that fixes legal content and hard feasibility as a controlled normative state while confining learning to operational choices. Section 3.2 defines the time-varying typed graph and provides the formal object in which sources, norms, interpretations, dependencies, and amendments are represented. Section 3.3 adds entity specificity, three-valued applicability, and conservative compliance semantics for evaluation. It ensures that incomplete entity context causes retention, investigation, or escalation rather than silent exclusion of potentially applicable obligations. Regulatory updates and their propagation through the graph are treated separately in Section 4.

3.1 Decision object and separation principle

In general, the system receives 1) the context of an obliged entity, 2) an internal AML policy and its associated controls, 3) a time-indexed legal graph, and 4) a legal graph update. Its output

is a set of proposed policy edits. Every edit is linked to the legal propositions and sources that motivate it, together with an assurance decision: A) machine acceptance, B) additional machine verification, C) specialist routing, or D) mandatory human approval.

Observe that two categorically different kinds of state coexist in the system. The first is normative content, such as legal propositions, legislative source hierarchy, effective and transitional dates, as well as the non-negotiable compliance constraints. The second is the operational configuration, that is, data retrieval strategies, agent prompts, tool budgets, escalation thresholds, routing tables, or any other parameter that determines how the system spends computation and expert attention. Note that descriptions and developments related to learning only concern the second content type throughout the paper. We make this separation explicit in the subsequent definition.

Definition 1 (separation principle). *Denote the system state as a pair $(\mathcal{N}, \omega)$, where $\mathcal{N}$ is the normative layer and ω is the operational layer. Two update operators act on the state. For every ingestion event r, the ingestion operator $\mathcal{I}_r$ maps $(\mathcal{N}, \omega)$ to $(\iota_r(\mathcal{N}), \omega)$, where the normative update map ι_r is independent of ω, and is driven exclusively by authenticated legal sources and adjudicated interpretations. The learning operator $\mathcal{L}$ maps $(\mathcal{N}, \omega)$ to $(\mathcal{N}, \omega')$ and is driven by feedback, logged trajectories, and reward signals. The system respects the separation principle if every state transition factors as a finite composition of $\mathcal{I}_r$ and $\mathcal{L}$ steps.*

Note that $\mathcal{N}$ contains the legal graph G_s to be defined in Definition 2 with its attribute maps, and the hard feasibility predicate Φ. In light of this definition, we state the following elementary proposition.

Proposition 1 (normative invariance). *In a system respecting Definition 1, the normative layer after any finite sequence of update steps equals the result of applying the ingestion steps alone, in order. In particular, no sequence of feedback events, however long and unanimous, alters a legal proposition, a source ranking, a validity interval, or the predicate Φ.*

Proposition 1 prevents the reward model introduced in Section 6.2 from turning legal validity into a popularity score in the sense of preference-based reward learning (Christiano et al., 2017). The human feedback may inform applicability judgments, interpretive uncertainty, drafting quality, operational proportionality, or the allocation of review effort. However, it cannot replace binding regulatory sources or erase conflicting interpretations. The only path into $\mathcal{N}$ runs through $\mathcal{I}$, which demands provenance. When a human reviewer's feedback reveals that an encoded legal proposition is wrong, the correction enters as an ingestion event with its own provenance record.

3.2 Time-varying graph and norm representation

Next, we introduce the formalization of a legal state and the representations of norms, edge taxonomy, as well as dependency semantics.

Definition 2 (legal state). *For every knowledge time s in a discrete index set $\mathcal{S}$, the corpus known to the system is a typed, directed property graph*

$$G_s = (V_s, E_s, \tau, \rho, \nu, \kappa), \tag{1}$$

where V_s contains the sources, provisions, interpreted norms, definitions, exceptions, concepts, entity classes, duties, prohibitions, permissions, controls, and evidentiary propositions. $E_s \subseteq V_s \times V_s$ contains typed relations among them, while τ is the type map assigning each node and edge a legal or semantic type. The authority map ρ represents the jurisdiction, source hierarchy, issuing authority, and precedence. The validity map ν contains intervals on legal effectivity, amendments, repeals, and transitions. The provenance κ links every legal proposition to a source text, version, location, as well as an interpretive status. Persistent legal objects have stable identities across graph versions. The attribute maps are version-specific, although their time subscripts are suppressed.

Graph-based retrieval is developed in GraphRAG (Edge et al., 2024) and LegalGraphRAG (Chen et al., 2026). Related compliance work combines retrieval-augmented generation with structured regulatory and process knowledge (Sun et al., 2025). Observe that the process $s \mapsto G_s$ is piecewise constant, that is, it only changes at ingestion events. Thus, between consecutive versions of the operative corpus, the corpus is a fixed object against which claims can be checked. Two temporal axes must be distinguished here. On the one hand, we have the validity time t, which is recorded in ν and displays when a norm produces a legal effect. On the other hand, we have knowledge time s, which is recorded in κ and displays when the system learned of it.

Note that this distinction is important. The EU AML package, in particular Regulation (EU) 2024/1624, entered the regulatory corpus of interest in 2024. However, the Regulation generally applies from 10 July 2027. The graph must distinguish this future application from the legal requirements applicable before that date and from national measures and supervisory instruments relevant to the entity. Publication, entry into force, application, and national transposition are distinct events (European Parliament and Council of the European Union, 2024b,a). Accordingly, a graph without both temporal axes cannot represent an obligation that is already known but not yet operative. This is precisely the object a forward-looking policy update must reason about.

An interpreted norm n is represented as a structured object according to

$$n = (\mathrm{id}, j, \rho, [b, e), q, m, s_n, a, o, d, X, w, \kappa), \tag{2}$$

where id is the identifier, j is the jurisdiction, ρ is the regulatory authority, $[b, e)$ is the validity interval, q is the applicability condition, $m \in \{\mathbf{O}, \mathbf{F}, \mathbf{P}\}$ (obligation, prohibition, permission; see von Wright, 1951) is the modality, and s_n is the regulated subject. Moreover, a denotes the act or state, o is an object, d is a deadline, X is an exception set, w is a materiality weight, and κ is again the provenance. We distinguish between a provision as text on the one hand, and a norm as an interpretation of that text on the other hand (Prakken and Sartor, 2015; Constant, 2024). Multiple norm nodes may point to the same provision node when interpretation is contested, where the interpretive status of each reading lives in κ. The temporal edge types (amends, repeals, transitional to) follow the formal treatments of legal change and abrogation (Governatori and Rotolo, 2010). Materiality weights w_n are set through a documented legal risk taxonomy and are not learned from the reviewer preferences. Note that this is an instance of the separation principle in Definition 1.

The typed edges of E_s fall into six families: structural (contains, paragraph-of, annex-of, implements), normative (imposes, prohibits, permits, exempts, overrides, qualifies), semantic (defines, narrows, broadens, refers-to, presupposes), temporal (amends, repeals, replaces, applies-from, transitional-to), operational (implemented-by-control, evidenced-by, documented-in-clause), and interpretive (supports-reading, conflicts-with, uncertainty-about). A designated subset D of edge types carries dependency semantics. The direction convention is thereby fixed. A dependency edge $u \to v$ means that a change at u may alter the truth, scope, priority, applicability, or operational realization of v. Section 4 turns this informal reading into a precise property of evaluation semantics, which will be important for the impact closure theorem provided there.

3.3 Entity context and compliance evaluation

Next, we elaborate on an obliged entity's relevant context and address compliance evaluation, thereby following Kleene (1952).

Definition 3 (entity context). *Let $\mathcal{A}$ be a countable set of ground atoms describing legally relevant entity characteristics: entity type, licences, products, customer and distribution channels, geography, outsourcing arrangements, risk appetite, and governance facts. An entity context is a partial assignment $x : \mathcal{A} \rightharpoonup \{0, 1\}$. Contexts are ordered by information $x \sqsubseteq x'$ if x' extends x,*

that is, x' agrees with x wherever x is defined. A context x' is a completion of x if $x \sqsubseteq x'$ and x' has full domain.

Applicability conditions q_n are therefore formulas over $\mathcal{A}$ that are created from atoms via negation, conjunction, disjunction, and bounded quantification over finite domains. Since contexts are partial, compliance evaluation must include three values. The strong Kleene valuation $[\![q]\!]_x \in \{1, 0, \bot\}$ evaluates an atom to its assigned value where x is defined, to $\bot$ elsewhere, and extends through the connectives in the standard way. The negation $\neg$ swaps 1 and 0 while fixing $\bot$. A conjunction is 1 if both conjuncts are 1, 0 if either is 0, and $\bot$ else. Likewise, the extension is dual for disjunction. From here, we can state monotonicity of Kleene's logic in the knowledge order (see Kleene, 1952; Fitting, 1985).

Lemma 1 (Kleene monotonicity). *Order the truth values by knowledge, $\bot \sqsubseteq 0$ and $\bot \sqsubseteq 1$, with $0, 1$ incomparable. Then, for every q over $\mathcal{A}$ and all contexts $x \sqsubseteq x'$, it holds that*

$$[\![q]\!]_x \sqsubseteq [\![q]\!]_{x'}. \tag{3}$$

A definite value never flips as facts are added, that is, if $[\![q]\!]_x \in \{0, 1\}$, then $[\![q]\!]_x = [\![q]\!]_{x'}$.

Lemma 1 marks the first basis for a conservative applicability screen. The context x of an entity may be incomplete, thus there might be aspects of it that are unknown. Assume that

$$q = \text{credit institution } \wedge \text{ cross-border service.}$$

If an entity is known to be a credit institution but its cross-border status is unknown, then $[\![q]\!]_x = \bot$. After fact gathering, if cross-border activity is confirmed, then q becomes 1. If it is ruled out, q becomes 0. However, if both facts are already known and $q = 1$, adding consistent facts cannot make $q = 0$.

Definition 4 (applicability and conservative candidates). *Let $N_s \subseteq V_s$ be the norm nodes known at knowledge time s. The definitely applicable set and the conservative candidate set at legal evaluation time t are*

$$\mathrm{App}_{s,t}(x) = \{n \in N_s : t \in \nu(n), [\![q_n]\!]_x = 1\}, \quad \mathrm{App}^{\dagger}_{s,t}(x) = \{n \in N_s : t \in \nu(n), [\![q_n]\!]_x \in \{1, \bot\}\}. \tag{4}$$

In Definition 4, $\mathrm{App}^{\dagger}_{s,t}(x)$ denotes the conservative applicability screen. From here, and on the basis of Lemma 1, we can state the following proposition.

Proposition 2 (soundness and monotone refinement of conservative screening). *(i) For every completion x' of x, it holds that $App_{s,t}(x') \subseteq App^{\dagger}_{s,t}(x)$. A represented norm excluded by the conservative screen at the fixed legal evaluation time is inapplicable under every completion of the recorded facts. (ii) If $x \sqsubseteq x'$, then $App^{\dagger}_{s,t}(x') \subseteq App^{\dagger}_{s,t}(x)$ and $App_{s,t}(x) \subseteq App_{s,t}(x')$.*

The proposition states that gathering facts only shrinks the candidate set, and only grows the definite set. Neither operation ever silently drops an applicable norm. Proposition 2 converts Lemma 1 into a safety guarantee for determining which legal norms might apply to an entity. For a policy P, entity context x, and norm n, satisfaction is likewise three-valued, $\mathrm{Sat}(P, x, n) \in \{1, 0, \bot\}$. Here, 1 means supported compliance, 0 marks supported non-compliance, and $\bot$ represents unresolved. As an AML policy alone may not establish operational compliance, Sat factors through four conjuncts under the Kleene conjunction:

$$\mathrm{Sat} = \mathrm{Sat}^{\mathrm{doc}} \wedge \mathrm{Sat}^{\mathrm{design}} \wedge \mathrm{Sat}^{\mathrm{impl}} \wedge \mathrm{Sat}^{\mathrm{effective}}. \tag{5}$$

Here, we distinguish between documentary coverage, control design, implementation evidence, and operating effectiveness. The framework repairs the first two conjuncts, and records the latter

two as out-of-document dimensions. Against this background, we define the compliance of the pair (P, x) at time t as

$$\begin{aligned} \mathrm{Comp}_{s,t}(P,x) = 1 \iff & \mathrm{App}^{\dagger}_{s,t}(x) = \mathrm{App}_{s,t}(x) \\ & \text{and } \ \mathrm{Sat}(P,x,n) = 1 \text{ for every } n \in \mathrm{App}_{s,t}(x). \end{aligned} \tag{6}$$

This is a conservative rule for asserting compliance: every applicability question must be resolved and every applicable norm must have supported satisfaction. Membership in the conservative candidate set alone would not make unresolved applicability block a compliance claim, which is why the equality of the two applicability sets is explicit.

The formal framework therefore provides the state against which a policy can be evaluated at a given knowledge time and legal evaluation time. A regulatory amendment changes that state. The next question is consequently not yet how the policy should be rewritten, but which legal evaluations may have changed and must enter the repair stage.

4 Regulatory change and impact analysis

This section formalizes regulatory change and its propagation through the legal graph. It first defines the semantic graph delta and dependency-local evaluation, then constructs the impact cone and establishes its completeness. It closes by identifying missing dependency edges as the principal empirical qualification of that guarantee.

Definition 5 (semantic graph delta). *Under the stable cross-version identity alignment of Definition 2, the update from knowledge time s to $s+1$ is the semantic difference*

$$\Delta_s = G_{s+1} \ominus G_s = (V^+, V^-, E^+, E^-, A_V^{\Delta}, A_E^{\Delta}), \tag{7}$$

containing added and removed nodes and edges, as well as the persistent nodes and edges in A_V^{Δ} and A_E^{Δ} whose attributes changed. A change to the evaluation rule associated with a norm is recorded as an attribute change at that norm. The seed set is

$$\mathrm{Seed}(\Delta_s) = V^+ \cup V^- \cup A_V^{\Delta} \cup \mathrm{ends}(E^+ \cup E^- \cup A_E^{\Delta}), \tag{8}$$

where ends$(\cdot)$ collects the endpoints of the listed edges. Changes to edge types, including membership in D, are included in A_E^{Δ}. Purely administrative provenance updates may be retained separately from this semantic delta.

It is important to note that Δ_s is a semantic difference, not simply a change in text. Under stable identity alignment, a renumbering that changes no evaluation-relevant attribute or dependency produces no substantive change seed, although provenance and citation locations must still be updated, while a single word addition that narrows a definition produces a seeded node. In light of Definition 5, the impact cone induced by a regulatory update is structurally a dependency slice of the legal graph (Weiser, 1984). We must therefore establish that every obligation whose evaluation can change across the update is reachable from the seeds. For this, the informal reading of the dependency edges is replaced by a locality property of the evaluation semantics.

Definition 6 (evaluation locality). *Let every node and edge carry its attribute data and let $Anc_D(n; G)$ denote the D-ancestors of n in G, which is all u with a directed D-path $u \to \ldots \to n$, including n itself. Write $G \restriction_{Anc_D(n;G)}$ for the rooted induced subgraph with its node, edge, and attribute data. An evaluation semantics is a family of maps assigning to each norm n its legal evaluation*

$$\mathrm{Eval}_n(G, x, P) = F_n\left(G \restriction_{\mathrm{Anc}_D(n;G)}, x, P\right), \tag{9}$$

returning the triple of applicability status, satisfaction status, and scope of n at a fixed legal evaluation time t, suppressed in the notation. Satisfaction notation also suppresses its dependence on the legal graph and t. The functional F_n is fixed across graph versions unless its change is recorded and seeded under Definition 5. If an object is not a norm in a given version, set its evaluation in that version to the distinguished value absent. *A repeal may instead be encoded by a changed validity interval while retaining the historical node. Either representation must enter the semantic delta. The semantics is D-local if* (9) *holds for every present n.*

Differently put, the evaluation of a norm is a function of the attributed subgraph induced by its D-ancestors, the entity context, as well as the policy, and nothing else in the graph. Thus, Definition 6 may be seen as a dependency completeness assumption with respect to the encoding. If a definition, exception, override, or a transitional provision can alter what norm n requires, then the encoding must contain a D-path from that node to n. Whether a given encoding satisfies it is an empirical question about the graph construction pipeline. We address this topic in more detail in Section 9.

Definition 7 (impact cone). *Compute reachability in the union graph whose D-edges are $E_D(G_s) \cup E_D(G_{s+1})$, so that both severed and newly created dependencies are traversed. The impact cone at the update, for legal evaluation time t, is*

$$I_{s,t} = \mathrm{Reach}_D\left(\mathrm{Seed}(\Delta_s)\right) \cap \left(\mathrm{App}^{\dagger}_{s,t}(x) \cup \mathrm{App}^{\dagger}_{s+1,t}(x)\right), \tag{10}$$

where $Reach_D(S)$ is the set of nodes reachable from S along D-edges of the union graph, which is a set that contains S itself.

Here, the applicability union is deliberate. Intersecting with $\mathrm{App}^{\dagger}_{s+1,t}(x)$ alone would cover current obligations but silently drop repealed or newly inapplicable norms. Policy maintenance must also surface the clauses those norms motivated so that obsolete text can be removed, revised, or retained with a documented justification. The post-update task for a repealed norm is to resolve its policy consequences, not to continue satisfying a repealed requirement. Next, we can state the main result of the section in Theorem 1.

Theorem 1 (impact closure completeness). *Let the evaluation semantics be D-local in both G_s and G_{s+1}, let Δ_s record every semantic change between versions under the stable identity alignment, and fix the context x, policy P, and legal evaluation time t. With the absent convention in Definition 6, every norm $n \in App^{\dagger}_{s,t}(x) \cup App^{\dagger}_{s+1,t}(x)$ with*

$$Eval_n(G_s, x, P) \neq Eval_n(G_{s+1}, x, P), \tag{11}$$

belongs to $I_{s,t}$.

Equivalently stated, a norm in the applicability union but outside the impact cone has identical applicability status, satisfaction status, and scope before as well as after the update. In practice, the theorem converts the question *Did we miss an affected obligation?* into two measurable encoding properties, namely semantic delta recall and D-locality. The impact cone represents a hard candidate set. A ranking may order the cone, but a retained candidate requires a documented disposition. Further evidence may resolve uncertain applicability, justify no policy change, support a repair, or require escalation.

Missing dependency edges are a realistic failure mode. Proposition 3 in Appendix B bounds omission risk when seeds and applicability are correct and edge-capture events are independent. Redundant paths can reduce this bound under those assumptions. Lemma 2 provides a ranking computation, not a completeness or probabilistic calibration guarantee.

The theorem compares graph versions at fixed x, P, and t. The passage of time across an application or expiry date, or a change in entity facts, requires a separate evaluation trigger and corresponding impact analysis even when no new source is ingested.

The impact cone is thus the interface between legal change detection and policy revision. It removes obligations whose evaluation is provably unchanged under the stated assumptions, while retaining every obligation that may require a new clause, a changed control, the removal of obsolete text, or an escalation. The repair problem in the next section takes precisely this retained set as its coverage target.

5 Minimal policy repair

In what follows, we address minimal revision of a policy given a regulatory amendment as constrained combinatorial optimization over a finite edit library. The section receives the impact cone from Section 4 and returns a candidate repair together with every obligation that still requires escalation. A feasible minimum exists when the finite library admits a feasible repair (Proposition 4). Even the pure coverage decision problem is NP-complete, and the optimization problem has a logarithmic inapproximability threshold unless P = NP (Theorem 2). For integer-valued submodular coverage with positive additive costs and no other constraints, greedy selection achieves a factor $1 + \ln c_{\max}$ (Theorem 3). This bound does not extend automatically to arbitrary conflicts or to the full objective J. Conflicts can make feasibility NP-complete (Proposition 5). Laminar prerequisite closures preserve the coverage guarantee under the conditions of Proposition 6.

We first formalize the policy state and the edit library.

Definition 8 (policy state and edit library). *The policy is a typed clause-and-control graph $P_s = (C_s, K_s, L_s, M_s)$ with clause nodes C_s, control nodes K_s, links L_s among clauses, controls, owners, and evidence, as well as a traceability mapping M_s from policy elements to legal norms. A finite edit library $\mathcal{E}(P_s)$ contains the admissible operations:* add, delete, replace, split, merge, re-reference, reassign control owner, modify threshold, add evidence requirement, *or* escalate an unresolved interpretation. *A policy repair is a set $R \subseteq \mathcal{E}(P_s)$ together with the canonical application order fixed by the library, and $P_s \oplus R$ denotes the repaired policy. The canonical order makes $\oplus$ well-defined on sets. Edits whose effects do not commute under the canonical order are declared conflicting and handled by the conflict relation of Proposition 5.*

For the minimal revision, the repair objective $J(R)$ combines implementation cost, residual compliance loss, documentary disruption, and review burden:

$$J(R) = C^{\text{edit}}(R) + \lambda \mathbb{E}\left[L^{\text{res}}\left(P_s \oplus R, G_{s+1}, x, t\right)\right] + \mu D(P_s, P_s \oplus R) + \xi C^{\text{review}}(R). \tag{12}$$

The four terms correspond to different practical consequences of a repair. They measure the work required to implement it, the compliance exposure that remains, the extent to which the existing policy is disturbed, and the expert attention needed before release. These quantities guide the choice among admissible repairs. They do not replace the feasibility predicate, which remains part of the normative layer. This cost is minimized over a set of feasible repairs, that is,

$$R^* \in \operatorname{argmin}\left\{J(R) : R \subseteq \mathcal{E}(P_s), \Phi(P_s \oplus R, G_{s+1}, x, t) = 1\right\}. \tag{13}$$

The feasibility predicate Φ encodes the non-negotiable requirements, namely disposition of every requirement arising from the impact cone $I_{s,t}$ under the updated legal state, source and date validity, internal consistency, permitted drafting operations, traceability, and resolution or escalation of every material $\perp$ value. Observe that Φ belongs to the normative layer of Definition 1. When escalation discharges Φ, feasibility means a documented referral for adjudication and does not assert that the policy is compliant or that regulatory deadlines have been suspended. Minimal revision in this context may imply cost minimality in (13), set-inclusion minimality,

semantic distance minimality, or Pareto efficiency in the components of J. The subsequent results are stated exclusively for cost minimality, and Section 9 reports the semantic edit distance separately. For clarity of the subsequent developments, we define the two key problems.

Definition 9 (min-repair and repair decision). *An instance consists of a policy state P_s, a finite edit library $\mathcal{E}(P_s)$, an impact cone $I_{s,t}$, an updated legal graph G_{s+1}, context x together with a feasibility predicate Φ, and an objective J that are evaluable in time polynomial in the instance size. All numerical inputs have polynomial-size rational encodings. Min-repair is the optimization problem* (13) *that returns a feasible repair of minimum cost, or reports that none exists. Repair decision takes in addition a budget $B \in \mathbb{Q}_{\geq 0}$ and is the language of instances admitting some $R \subseteq \mathcal{E}(P_s)$ with $\Phi(P_s \oplus R, G_{s+1}, x, t) = 1$ and $J(R) \leq B$. An instance is called pure coverage if $\lambda = \mu = \xi = 0$, so that $J(R) = \sum_{e \in R} w_e$ with non-negative rational costs w_e, where each edit e carries a known coverage set $S_e \subseteq I_{s,t}$, edits neither conflict nor interact, and $\Phi(P_s \oplus R, G_{s+1}, x, t) = 1$ exactly when $\bigcup_{e \in R} S_e = I_{s,t}$.*

Proposition 4 (existence). *If $\mathcal{E}(P_s)$ is finite, if at least one feasible repair exists, and if J is real-valued on feasible repairs, then the minimum in* (13) *is attained.*

Proposition 4 is elementary, but it isolates a genuine design obligation, namely that the edit library must admit a jointly feasible disposition of all requirements the impact cone may surface. Separate fallbacks for individual obligations need not be jointly feasible if they conflict. A documented escalation can discharge a disposition requirement when Φ permits it, but does not establish legal compliance or resolve an outstanding deadline. The hardness results concern general libraries without a guaranteed unrestricted escalation fallback. If such a fallback always satisfies the disposition predicate, existence is trivial, although optimization can remain difficult. In light of Definition 9 and Proposition 4, we now state NP-hardness of (13).

Theorem 2 (hardness and inapproximability of minimal repair). *(i) The repair decision in Definition 9 is NP-complete, and remains NP-complete when restricted to pure coverage instances. (ii) Fix any α with $0 < \alpha < 1$. It is NP-hard to approximate pure coverage within a factor $(1 - \alpha) \ln |I_{s,t}|$ for all sufficiently large $|I_{s,t}|$.*

The proof embeds weighted set cover, whose unweighted special case is a classical NP-complete problem (Karp, 1972). Obligations are the universe, edits are the sets, costs are the weights, and Φ is exactly coverage. The reduction preserves costs, so both hardness and the logarithmic approximation threshold transfer verbatim (Dinur and Steurer, 2014, Feige, 1998). The reduction uses the generic situation after an amendment, namely the coverage of impacted obligations by candidate edits. The other objective terms only refine the choice among covering repairs. The hardness therefore attaches to the repair task itself. While the hardness of the minimization problem calibrates expectations, it does not preclude guarantees. Model coverage by a function $c : 2^{\mathcal{E}(P_s)} \to \mathbb{Z}_{\geq 0}$ where $c(R)$ counts the impacted obligation requirements discharged by R, and assume that $c(\mathcal{E}(P_s))$ equals the full required coverage, not merely the maximum coverage attainable by an incomplete library. For the coverage-only results, feasibility is exactly $c(R) = c_{\max} := c(\mathcal{E}(P_s))$. Additional legal checks, conflicts, and interaction constraints must already have been resolved or must be handled separately.

Assumption (submodular coverage). *c is normalized ($c(\emptyset) = 0$), monotone ($A \subseteq B \to c(A) \leq c(B)$). Moreover, c is submodular, that is, for all $A \subseteq B$ and $e \notin B$, it holds that $c(A \cup \{e\}) - c(A) \geq c(B \cup \{e\}) - c(B)$.*

Submodularity is the statement that edits exhibit diminishing coverage returns and holds, in particular, when requirements are covered disjunctively by edits (Nemhauser et al., 1978). Genuine synergies between edits may violate it. After bundling synergistic edits into composite library entries, submodularity must be checked again. This is the same device for the precedence below.

Theorem 3 (greedy repair guarantee; Chvátal, 1979; Wolsey, 1982) *Let c be integer-valued, normalized, monotone and submodular, with $c_{max} \geq 1$. Assume coverage is evaluable in polynomial time and feasibility is exactly $c(R) = c_{\max}$. Let every edit have cost $w_e > 0$, let total cost be additive, and let W^* be the minimum cost of a covering repair. Starting from $A_0 = \emptyset$, the greedy algorithm that repeatedly adds the edit maximizing marginal coverage per unit cost,*

$$e_{k+1} \in \underset{e \in \mathcal{E}(P_s) \setminus A_k}{argmax} \left(c(A_k \cup \{e\}) - c(A_k) \right) / w_e,$$

until $c(A_k) = c_{max}$, terminates with a feasible repair of cost at most

$$W^{greedy} \leq (1 + lnc_{max}) W^*. \tag{14}$$

By Theorem 2, this logarithmic factor is essentially the best any polynomial-time method can promise in the regime where the two logarithms are comparable, as the lower bound is $(1 - \alpha) \ln |I_{s,t}|$ while the greedy factor is $1 + \ln c_{\max}$. With heavily aggregated or materiality-weighted coverage units the two scales can diverge and the claim is only near matching. The near-matching claim concerns the pure coverage model, not arbitrary conflict-constrained repair or the full objective J. In practice, the greedy pass proposes, and the orchestration layer in Section 6 spends its verification as well as escalation budget on exactly the steps whose marginal coverage claims are least certain.

Real edit libraries carry mutual exclusions, e.g., two alternative redraftings of the same clause, or a threshold change incompatible with a clause deletion. Therefore, conflict feasibility within the system must be addressed.

Proposition 5 (conflict feasibility is NP-complete). *Let a symmetric conflict relation $\mathcal{X} \subseteq \mathcal{E} \times \mathcal{E}$ forbid repairs containing a pair of conflicting edits. Deciding whether any conflict-free feasible repair exists is NP-complete.*

The proof represents each Boolean variable by a conflicting pair of edits, one asserting true and one asserting false. It represents each clause by an obligation covered by the edits associated with its literals. The consequence is that with conflicts, already feasibility, rather than only optimality, requires search. The algorithm of Section 8 treats an unsatisfiable conflict cluster as an escalation event with the cluster as evidence, and not as an empty result.

Proposition 6 (precedence through closures). *Under the coverage and additive-cost assumptions of Theorem 3, with no constraints other than coverage and the stated prerequisites, let $\preceq$ be a partial order on $\mathcal{E}(P_s)$, require repairs to be down-closed, and give each closure $\bar{e} = \downarrow e$, including e itself, the full closure cost $w(\bar{e}) = \sum_{f \in \bar{e}} w_f$. If the closure family $\{\bar{e} : e \in \mathcal{E}(P_s)\}$ is laminar, that is, any two closures are nested or disjoint, define the coverage of a family F of closures by $g(F) = c(\bigcup_{\bar{e} \in F} \bar{e})$. Then greedy selection on the derived ground set of closures with full closure costs is feasible for the precedence-constrained problem and inherits the bound (14), with c_{max} unchanged. The actual cost of the union of the selected closures is no greater than its full closure accounting cost.*

Non-laminar overlapping prerequisite closures break the comparison because shared prerequisites are paid once but credited multiple times. Greedy with marginal closure costs remains well-defined and feasible, but no general approximation factor is claimed here. The characterization of the exact degradation is a question for future research. From here, the residual risk can be stated as asymmetric loss

$$L^{\text{res}} = \sum_{n \in \text{App}^{\dagger}_{s+1,t}(x)} w_n \left[c_{\text{FN}} \cdot \mathbf{1}(\text{FN}_n) + c_{\text{FP}} \cdot \mathbf{1}(\text{FP}_n) + c_{\text{U}} \cdot \mathbf{1}(\text{Sat}(P, x, n) = \perp) \right], \tag{15}$$

with $c_{\mathrm{FN}} \gg c_{\mathrm{FP}}$, since omitting a material obligation is more serious than proposing an unnecessary clause, and c_{U} calibrated to the escalation mechanism. The penalty for unresolved uncertainty must reflect the risk and cost of its eventual disposition, including reviewer error and delay. Escalation alone does not make that risk negligible.

This residual loss connects the static repair problem to the sequential workflow. A repair engine can identify low cost candidate edits, but it cannot by itself decide whether additional evidence, another critique, or a human decision is worth the associated delay and cost. Those decisions form the orchestration problem considered next.

6 Risk-constrained human-AI orchestration

This section receives the candidate repairs, unresolved obligations, and residual losses produced above and determines how they should be processed. It formalizes the governed interaction between machine agents and human reviewers. Section 6.1 introduces the constrained partially observable workflow and its hard legal gates, while Section 6.2 develops the feedback model and one-step expert allocation.

6.1 Workflow and constrained POMDP

Hereafter, we model the multi-agent review workflow through six classes of action, namely retrieval, specialist consultation, critique, revision, expert routing, and acceptance. We formalize it as a constrained partially observable Markov decision process whose objective prices errors, human and machine cost, delay, and edit drift. It has two constraints for distinct purposes. A chance constraint caps the probability of any material omission, while a CVaR constraint bounds the severe loss tail that average-case performance would hide (Definition 10). Under bounded losses and a finite horizon, an optimal orchestration policy exists and is anchored by a feasible mandatory review baseline (Proposition 7). Finally, the hard gate on source fidelity, dates, and authority is genuinely lexicographic. Optimizing under the gate as a constraint is equivalent to the $-\infty$ reward formulation, and no finite penalty reproduces it uniformly over all instances (Proposition 8).

The multi-agent workflow is a controlled stochastic process. At workflow step k, the state summarizes the updated graph, the current draft, remaining impacted obligations, retrieved evidence, agent disagreements, uncertainty estimates, resource use, and history:

$$S_k = (G_{s+1}, \Delta_s, P_k, Q_k, E_k, U_k, B_k, H_k). \tag{16}$$

The finite action set $\mathcal{A}$ allows the system to retrieve additional authority, expand the graph neighborhood, consult a specialist agent, request an independent critique, revise a proposed edit, route the case to a human expert, or accept the repair after the hard gate and risk gate both pass. The specialist roles cover applicability, drafting, temporal validity, and contradiction analysis. The revision action may split, merge, replace, or reject a proposed edit. Action admissibility is determined by the observable history. The transitions follow $S_{k+1} \sim T_k(\cdot|S_k, A_k)$.

The environment is partially observed, that is, the legally preferable interpretation and the practical adequacy of a repair are latent. Moreover, it is non-stationary, since the law, the obliged entity, the expert pool, and the model stack all drift. The general formulation is therefore a constrained partially observable Markov decision process (Kaelbling et al., 1998; Altman, 1999).

Definition 10 (constrained orchestration problem). *With finite action and observation sets, a finite horizon K per amendment episode, and a discount factor $0 < \gamma \leq 1$, a policy π maps*

observable histories to distributions over actions. The orchestration problem is given by

$$\min_{\pi} \mathbb{E}\left[\sum_{k=0}^{K-1} \gamma^k \ell(S_k, A_k)\right] \quad \textit{subject to } \mathbb{P}_\pi(M=1) \le \varepsilon,\ \mathit{CVaR}_\eta(L^{\mathit{severe}}) \le \zeta, \tag{17}$$

where M indicates a material miss, L^{severe} is the severe loss component, and, with $0 < \eta < 1$ denoting the upper-tail mass, $\mathit{CVaR}_\eta(L) = \min_c\{c + \eta^{-1}\mathbb{E}\max(L-c,0)\}$ is the CVaR in its variational form. The per-step loss is

$$\ell = \lambda_1 FN + \lambda_2 FP + \lambda_3 \mathcal{C}^{human} + \lambda_4 \mathcal{C}^{machine} + \lambda_5 Delay + \lambda_6 EditDrift. \tag{18}$$

Observe that the constraints in (17) serve two purposes. The chance constraint (Charnes and Cooper, 1959) caps the probability of any material omission, while the tail constraint protects against policies that perform well on average, but fail on rare, complex amendments, such as the union graph of a transposition deadline colliding with a national transitional regime.

Proposition 7 (attainment). *For fixed initial-state, transition, and observation laws within the episode, assume the per-step loss ℓ is bounded and the severe loss satisfies $0 \le L^{severe} \le \bar{L} < \infty$. With finite action and observation sets as well as a finite horizon K, the set of history-dependent randomized policies is a compact polytope in finite dimension, and the objective as well as the constraint functionals in* (17) *are continuous on it. Assume the feasible set is non-empty. The natural anchor is the mandatory-review baseline π_0, which escalates every material repair, provided its residual human error miss probability is at most ε and its severe loss tail risk is at most ζ by assumption. In that case, an optimal policy exists, and it may be taken randomized. Throughout, η with $0 < \eta < 1$ in CVaR_η denotes the tail mass, so the coefficient η^{-1} in Definition 10 is consistent with that convention. Under the confidence-level convention, the coefficient would read $(1-\eta)^{-1}$.*

This proposition ensures that optimization returns an attainable policy rather than an infimum approached by ever more aggressive acceptance policies. Moreover, it names the natural feasible anchor π_0, which also constitutes the baseline of the self-improvement algorithm in Section 8.

Before feedback can be used to improve the workflow, the system must determine which aspects of a review may enter the reward at all. In the setting of compliance automation, a hard gate assigns reward $-\infty$ to an acceptance that violates the gate. This corresponds to an infinite penalty for non-compliance. The following proposition shows that this representation implements constrained optimization in our setting.

Proposition 8 (hard gate as constraint). *Define $V(\pi) = -\infty$ if $\mathbb{P}_\pi(\text{acceptance violates gate}) > 0$ and $V(\pi) = \mathbb{E}_\pi[u-c]$ otherwise. If at least one policy respecting the gate exists, its expected utility is finite, all gate-respecting policies have finite expected utility, and the maxima are attained, then*

$$\operatorname*{argmax}_{\pi} V(\pi) = \operatorname*{argmax}_{\pi:\, \mathbb{P}_\pi(\textit{gate violation})=0} \mathbb{E}_\pi[u-c].$$

For every finite penalty there is an instance in which the penalized optimum violates the gate with positive probability. Hence, no single finite penalty reproduces the hard constraint uniformly over all instances.

The consequence for the remainder of the system is direct. Feedback scores may rank admissible drafts and may help allocate review effort, but they cannot compensate for a failed gate. The feedback model below therefore records normative defects separately from the dimensions used for operational learning.

6.2 Structured human feedback and expert allocation

This subsection replaces the undifferentiated approval score with a multidimensional feedback record. Five dimensions are hard gate constraints, including applicability. The remaining dimensions feed a reviewer-specific Bradley and Terry model. Proposition 9 gives conditions under which reviewer skill, presentation bias, and repair quality can be separated from pairwise judgements. Propositions 10 and 11 then address one-step expert allocation through learning to defer. Repeated and sequential route learning are treated separately in Section 7.

6.2.1 Structured human feedback and reward construction

For a proposed repair y_i and a reviewer h, we define a multidimensional feedback as follows:

$$F_{ih} = (z_{\text{app}}, z_{\text{legal}}, z_{\text{source}}, z_{\text{time}}, z_{\text{complete}}, z_{\min}, z_{\text{ops}}, z_{\text{clarity}}, \text{sev}, \text{comment}) . \tag{19}$$

Here z_{app} records whether applicability is resolved correctly, z_{legal} legal sufficiency, z_{source} source fidelity, z_{time} temporal validity, and z_{complete} completeness. These five dimensions are mandatory checks. The remaining scores assess minimal disruption ($z_{\min}$), operational suitability (z_{ops}), and drafting clarity (z_{clarity}). The fields sev and comment record defect severity and the reviewer's explanation. A material unresolved result on a mandatory check blocks machine acceptance. They act as constraints in the sense that by Proposition 8, they cannot be offset by stylistic changes or efficiency gains. The remaining scored dimensions describe the practical quality of an otherwise admissible repair and feed a learned utility. Severity and comments retain information needed for escalation and audit. Where pairwise feedback is available, a reviewer-specific Bradley-Terry model (Bradley and Terry, 1952) with a within-pair order effect in the sense of Davidson and Beaver (1977) estimates latent utility while acknowledging heterogeneous reviewer precision and presentation bias. With $\delta_{ij} = u_\theta(y_i) - u_\theta(y_j)$ and $s_{ij} = 1$ when y_i is presented first and $s_{ij} = -1$ when it is presented second, we have

$$\mathbb{P}(y_i \succ y_j | h, s_{ij}) = \sigma(\beta_h \delta_{ij} + s_{ij} b_h), \qquad \beta_h > 0, \tag{20}$$

where $\sigma(z) = (1 + e^{-z})^{-1}$, β_h captures reviewer discrimination in preferences, and b_h captures systematic first-position bias. Preference discrimination is not an estimate of legal accuracy. Without design care, these parameters are confounded with the utilities. The following proposition states exactly what ensures identifiability.

Proposition 9 (identifiability of the feedback model). *Suppose presentation order is randomized so that each compared pair is shown to reviewer h in both orders with positive probability. Then, from the comparison of the probabilities, it follows:*

(i) If $p^+_{h,ij}$ and $p^-_{h,ij}$ denote the probabilities that y_i is preferred when it is presented first and second, respectively, then

$$\beta_h \delta_{ij} = \frac{\mathit{logit}(p^+_{h,ij}) + \mathit{logit}(p^-_{h,ij})}{2}, \qquad b_h = \frac{\mathit{logit}(p^+_{h,ij}) - \mathit{logit}(p^-_{h,ij})}{2}.$$

(ii) if additionally the comparison graph on repairs is connected (Ford, 1957), and every other reviewer shares with a reference reviewer h_0 at least one compared pair with $\delta_{ij} \neq 0$, then under the normalization $\beta_{h_0} = 1$ and $u(y_{i_0}) = 0$, all observed-repair utilities $u(y_i)$, discriminations β_h, and biases b_h are identified. Identification of a parameter θ in a representation u_θ additionally requires that the parameterization be injective on the observed design.

(iii) without order randomization, joint identifiability of (β_h, b_h, u) can fail in the sense that there exist natural deterministic presentation designs admitting distinct parameter values with identical observable comparison probabilities. Randomization is thus a sufficient design guarantee.

Note that part (iii) is in fact a design mandate. Randomized order provides the variation required by the stated identification argument. A fixed-order design need not identify these parameters without additional restrictions. Reviewer identity, seniority, specialization, and access to auxiliary information can then be layered hierarchically on (20). Throughout, the adjudicated legal label remains distinct from the preference model. Preferences allocate attention and tune drafting, while adjudications provide governed legal labels subject to the human-error assumptions below. This preserves the separation principle when feedback is converted into operational decisions.

We first address the one-step allocation problem. Repeated and sequential allocation are treated in Section 7.

6.2.2 Learning to defer

Learning to defer is required when one principal decision is to be made. In that case, the machine handles a case or selects a human expert. The one-step formulation is the learning to defer problem (Madras et al., 2018; Mozannar and Sontag, 2020), extended to multiple experts as in Mao et al. (2025), and to complementarity as in Charusaie et al. (2022). Let $\rho_M(x)$ be the expected loss if the machine finalizes case x, and for each expert $h \in \mathcal{H}$, let $\rho_h(x) = \mathbb{E}[L^h|x] + c_h$ be that expert's expected error loss plus consultation cost and delay. The one-step router chooses from a non-empty admissible subset $\mathcal{A}_{\text{safe}}(x) \subseteq \{M\} \cup \mathcal{H}$. Mandatory review and acceptance requirements are enforced before this minimization. The rule below minimizes one-step expected loss within that set; it does not, on its own, solve the global chance- or CVaR-constrained problem.

Proposition 10 (Bayes deferral rule). *Under correctly specified conditional risks and additive review cost, the policy $a^*(x) \in \operatorname{argmin}_{a \in \mathcal{A}_{\text{safe}}(x)} \rho_a(x)$ minimizes expected total loss among all routing rules. Deferring to an admissible expert h^* is optimal exactly when $\rho_{h^*}(x) \leq \rho_a(x)$ for every admissible alternative a. The rule need not remain sufficient when a review action generates information that alters later workflow decisions.*

Note that the conditional risks are never known, but estimated in practice. The next result bounds the price of estimation error and thereby tells the calibration effort where to go.

Proposition 11 (plug-in deferral excess risk). *On the same admissible action set, let $\hat{\rho}_a$ be estimates with $|\hat{\rho}_a(x) - \rho_a(x)| \leq \varepsilon_a(x)$ for all a and x, and let $\hat{a}(x) \in \operatorname{argmin}_a \hat{\rho}_a(x)$ be the plug-in router. Then, for every x,*

$$\rho_{\hat{a}(x)}(x) - \rho_{a^*(x)}(x) \leq \varepsilon_{\hat{a}(x)}(x) + \varepsilon_{a^*(x)}(x) \leq 2 \max_a \varepsilon_a(x), \tag{21}$$

and the same bound holds in expectation over x.

Proposition 11 states that routing quality degrades no faster than twice the worst calibration error among the actions actually in contention. The argument is the classical plug-in excess risk bound of pattern recognition (Devroye et al., 1996) applied to routing. This result concerns a single routing decision. The next section treats repeated choices, logged outcomes, and complete workflow trajectories while retaining the same hard admissibility boundary.

7 Safe adaptive learning and finite-sample assurance

This section studies learning procedures that operate only within the action set admissible under the hard gate. The legal state and feasibility conditions remain fixed in accordance with Proposition 1. Learning uses logged operational outcomes to compare review routes, allocate future cases, and calibrate machine acceptance. Section 7.1 develops logged data estimators for contextual routing, Section 7.2 gives safe route regret guarantees under stationarity and detected drift,

and Section 7.3 treats sequential policy selection from logged trajectories. Section 7.4 connects these learning results to calibrated acceptance, expert review risk, and the end-to-end assurance argument. Section 7.5 then states how correlated expert errors, detected drift, and uncertainty about policy performance affect operational deployment.

7.1 Off-policy evaluation for contextual routing

For context X, the orchestrator chooses action A from a finite set and observes reward, defined as negative loss, only for the chosen action. The practical question is whether a candidate routing policy would have performed better than the policy that generated the log, without deploying the candidate first. Throughout this subsection, the evaluation observations are i.i.d., the target policy π is fixed independently of them, and the conditional reward law given (X, A) is invariant between behavior and target policies. Thus the observed context must contain the information needed to remove confounding of action and reward. Assume the relevant expectations are finite, and write $\mu(x, a) = \mathbb{E}[R \mid X = x, A = a]$ and $V(\pi) = \mathbb{E}_X[\sum_a \pi(a \mid X)\mu(X, a)]$. Logged data must record the behavior policy propensity $b(A|X)$. This logging discipline makes the counterfactual estimators in this subsection available later.

Lemma 3 (inverse-probability unbiasedness; Horvitz and Thompson, 1952; Li et al., 2012). *Suppose that*

$$\pi(a \mid x) > 0 \quad \Longrightarrow \quad b(a \mid x) > 0.$$

Terms corresponding to $b(a \mid x) = \pi(a \mid x) = 0$ are omitted from propensity-weighted sums. If the logged propensities are correct, then the inverse-propensity estimator

$$\widehat{V}_{\mathrm{IPS}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(A_i \mid X_i)}{b(A_i \mid X_i)} R_i \tag{22}$$

satisfies

$$\mathbb{E}_b[\widehat{V}_{\mathrm{IPS}}(\pi)] = V(\pi).$$

Lemma 4 (double robustness; Robins et al., 1994; Dudík et al., 2011) *Let $\widehat{\mu}(x, a)$ and $\widehat{b}(a \mid x) > 0$ be nuisance functions fitted on a sample independent of each observation being evaluated; the same condition may be enforced fold by fold by cross-fitting. Put*

$$\widehat{\mu}_\pi(x) = \sum_{a \in \mathcal{A}_{\mathrm{safe}}(x)} \pi(a \mid x) \widehat{\mu}(x, a)$$

and define

$$\widehat{V}_{\mathrm{DR}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \left[\widehat{\mu}_\pi(X_i) + \frac{\pi(A_i \mid X_i)}{\widehat{b}(A_i \mid X_i)} \big(R_i - \widehat{\mu}(X_i, A_i)\big) \right]. \tag{23}$$

Under the positivity condition of Lemma 3, for an independent nuisance-training sample, $\widehat{V}_{\mathrm{DR}}(\pi)$ is conditionally unbiased for $V(\pi)$ if either $\widehat{b} = b$ on the support of π or $\widehat{\mu} = \mu$, provided the weighted residual is integrable. With cross-fitting, this argument applies to each held-out fold conditional on its own training sample. Averaging gives unconditional unbiasedness when one of the two correctness conditions holds in every fold; conditioning jointly on all fitted folds need not preserve unbiasedness.

If nuisance functions are fitted on the evaluation observations themselves without sample splitting, exact finite-sample unbiasedness need not hold. Asymptotic validity then requires additional consistency, rate, and empirical-process conditions.

Lemma 5 (clipping bias; Ionides, 2008). *Assume $|R_i| \le R_{\max} < \infty$ and the positivity condition of Lemma 3. Let*

$$W_i = \frac{\pi(A_i \mid X_i)}{b(A_i \mid X_i)}$$

and, for $M > 0$, define the weight-clipped estimator

$$\widehat{V}_M(\pi) = \frac{1}{n} \sum_{i=1}^{n} \min\{W_i, M\} R_i. \tag{24}$$

Then

$$\left| \mathbb{E}_b[\widehat{V}_M(\pi)] - V(\pi) \right| \le R_{\max}\, \mathbb{E}_b[(W - M)_+], \tag{25}$$

where $W = \pi(A \mid X)/b(A \mid X)$ and $(u)_+ = \max\{u, 0\}$. The right-hand side converges to zero as $M \to \infty$.

Together, Lemmas 3 to 5 identify the operational requirements for credible reuse of workflow logs. The system must preserve overlap between the logged and candidate policies, record the propensities, separate nuisance fitting from evaluation, and report the effect of any clipping. The first two lemmas provide unbiasedness under complementary conditions. The third makes the variance reduction obtained by clipping visible as an explicit bias term. These requirements motivate controlled data collection within the safe route set.

7.2 Safe route selection under stationarity and drift

Propensities close to zero create unstable weights and weak evidence for a live compliance system. The system therefore collects information by randomizing only among review routes that already satisfy the mandatory review rules and the hard gate. The exploration changes how admissible work is allocated, but it does not change which actions are admissible. Against this background, the classical optimism principle applies with its full guarantee.

Theorem 4 (regret of cost sensitive route selection; Auer et al., 2002). *Fix a context stratum and let its safe slate $\mathcal{A}_{\text{safe}}$ contain K routes. For each route $a \in \mathcal{A}_{\text{safe}}$, suppose that its successive per-episode total losses are independent and identically distributed in $[0, 1]$, are pre-generated independently across routes, and are observed by a non-anticipating router that uses only past observations, and have mean μ_a. Define*

$$\mu^* = \min_{a \in \mathcal{A}_{\text{safe}}} \mu_a, \qquad \Delta_a = \mu_a - \mu^*.$$

The optimistic rule first selects every route once, insofar as the horizon allows, and thereafter, at round t, selects

$$A_t \in \operatorname*{argmin}_{a \in \mathcal{A}_{\text{safe}}} \left\{ \widehat{\mu}_a(t-1) - \sqrt{\frac{2 \log t}{N_a(t-1)}} \right\}, \tag{26}$$

where $N_a(t-1)$ and $\widehat{\mu}_a(t-1)$ are respectively the number and the empirical mean of the losses observed from route a before round t. For $T \ge 2$, define the pseudo-regret

$$\mathcal{R}_T = \sum_{t=1}^{T} \big(\mu_{A_t} - \mu^*\big).$$

Then

$$\mathbb{E}[\mathcal{R}_T] \le \sum_{a:\, \Delta_a > 0} \left\{ \frac{8 \log T}{\Delta_a} + \left(1 + \frac{\pi^2}{3}\right) \Delta_a \right\}. \tag{27}$$

In particular, the expected pseudo-regret is logarithmic in T, and every exploratory action remains inside the safe slate. Under the standard regularity and uniform-efficiency conditions for

stochastic-bandit lower bounds, logarithmic instance-dependent regret cannot generally be improved in order (Lai and Robbins, 1985). In system terms, the cost of learning which approved route works best grows more slowly than the number of amendment episodes, while no episode is assigned to a route that the normative layer has excluded.

Corollary 1 (piecewise stationary regret under drift alarms; Garivier and Moulines, 2011). *In the same fixed context stratum, suppose the horizon is partitioned into Φ deterministic stationary amendment phases $\mathcal{I}_1, \dots, \mathcal{I}_\Phi$ of lengths $T_1, \dots, T_\Phi$. Within phase φ, let $\mathcal{A}_{\varphi,\mathrm{safe}}$ be the current safe slate. Each route in that slate has an i.i.d. loss stream in $[0,1]$, with mean $\mu_{\varphi,a}$, optimum $\mu^*_\varphi = \min_{a \in \mathcal{A}_{\varphi,\mathrm{safe}}} \mu_{\varphi,a}$, and gap $\Delta_{\varphi,a} = \mu_{\varphi,a} - \mu^*_\varphi$. Define*

$$B_\varphi(T_\varphi) = \sum_{\substack{a \in \mathcal{A}_{\varphi,\mathrm{safe}}: \\ \Delta_{\varphi,a} > 0}} \left\{ \frac{8 \log(\max\{2, T_\varphi\})}{\Delta_{\varphi,a}} + \left(1 + \frac{\pi^2}{3}\right) \Delta_{\varphi,a} \right\}.$$

Assume the route loss streams satisfy the independence and non-anticipation conditions of Theorem 4 within each phase. Alarms must be stopping times for the observed history, and unobserved post-alarm losses must retain the stated conditional i.i.d. laws. Assume that each phase change is signalled before the next phase change and within at most a fixed Δ_{det} episodes, that there are no false alarms, and that the rule is restarted at every alarm. If an amendment changes the safe slate, assume additionally that the hard gate is updated when the amendment takes effect. The possibly delayed alarm concerns only the restart of the statistical estimates. Then the expected dynamic pseudo-regret against the best safe route in each phase, defined by

$$\mathcal{R}_T^{\mathrm{dyn}} = \sum_{\varphi=1}^{\Phi} \sum_{t \in \mathcal{I}_\varphi} \left(\mu_{\varphi,A_t} - \mu^*_\varphi\right),$$

satisfies

$$\mathbb{E}[\mathcal{R}_T^{\mathrm{dyn}}] \le \sum_{\varphi=1}^{\Phi} B_\varphi(T_\varphi) + (\Phi - 1)\Delta_{\mathrm{det}} \le \sum_{\varphi=1}^{\Phi} B_\varphi(T_\varphi) + \Phi \Delta_{\mathrm{det}}. \tag{28}$$

Thus, the explicit detection-delay contribution is bounded by the constant Δ_{det} per amendment, while a logarithmic relearning term remains for every stationary phase.

The corollary separates two costs of legal change. Detection delay contributes at most a constant term per amendment, while each stationary phase contributes its own logarithmic relearning term. Thus, the learner need not be discarded whenever the legislator moves. Its statistical state is restarted when the environment changes, while the updated hard gate applies as soon as the amendment takes effect.

7.3 Offline sequential policy selection

The bandit model treats each review route as one decision with one resulting loss. Offline RL can be used when retrieval, agent selection, drafting, critique, and escalation jointly form a trajectory whose early decisions alter later states and costs. Let d^π and d^b be time-indexed occupancy measures of a candidate and behavior policy on the information states used for control. Define the coverage coefficient by

$$C(\pi, b) = \begin{cases} \operatorname{ess\,sup}_{d^b} \dfrac{\mathrm{d}d^\pi}{\mathrm{d}d^b}, & d^\pi \ll d^b, \\ +\infty, & \text{otherwise.} \end{cases}$$

In a finite space this is the largest ratio $d^\pi(s,a)/d^b(s,a)$ on the behavior support, with value $+\infty$ if the target assigns positive mass outside that support. A large value indicates weak

overlap, while missing support prevents unrestricted off-policy identification (see Assumption A6, Appendix A).

Lemma 6 (trajectory importance sampling; Precup et al., 2000). *Consider a horizon-H trajectory ξ generated under a behavior policy b, and let π be a target policy operating under the same initial state distribution and environment dynamics. Suppose that*

$$\pi(a \mid h) > 0 \quad \Longrightarrow \quad b(a \mid h) > 0$$

at every π-reachable history. For every trajectory return G integrable under the target law, define

$$w(\xi) = \prod_{k=1}^{H} \frac{\pi(a_k \mid h_k)}{b(a_k \mid h_k)}.$$

Then

$$\mathbb{E}_b[w(\xi)G(\xi)] = \mathbb{E}_\pi[G(\xi)]\,.$$

If, in addition, $\pi(a \mid h)/b(a \mid h) \le C_\infty$ whenever $\pi(a \mid h) > 0$, then $0 \le w(\xi) \le C_\infty^H$.

The exponential dependence on H in the weight bound is the main reason the proposed framework prefers short decision horizons (per-repair episodes) and why pessimism governs offline selection (Rashidinejad et al., 2021; Jin et al., 2021). For more details on the interaction of offline RL with human feedback, see Zhan et al. (2023).

Theorem 5 (pessimistic offline policy selection; Thomas et al., 2015b). *Let Π be a finite candidate class fixed independently of the logged evaluation sample. For each $\pi \in \Pi$, suppose that an estimator $\widehat{V}(\pi)$ and a confidence width $w_n(\pi)$ satisfy*

$$\mathbb{P}\Big(\Big|\widehat{V}(\pi) - V(\pi)\Big| > w_n(\pi)\Big) \le \frac{\delta}{|\Pi|}.$$

Then every lower-confidence-bound maximizer

$$\widehat{\pi} \in \operatorname*{argmax}_{\pi\in\Pi} \Big\{\widehat{V}(\pi) - w_n(\pi)\Big\}$$

satisfies, with probability at least $1-\delta$,

$$V(\widehat{\pi}) \ge \max_{\pi\in\Pi} \{V(\pi) - 2w_n(\pi)\}\,. \tag{29}$$

Observe that an example for a suitable estimator $\hat{V}(\pi)$ can be obtained from the IPS estimator of Lemma 6 with Hoeffding width $w_n(\pi) = C_\infty^H\sqrt{\ln(2|\Pi|/\delta)/(2n)}$ when the return G lies in $[0,1]$ and the logged trajectories are i.i.d. A known bounded negative-loss return may first be shifted and rescaled. The selected policy in Theorem 5 competes with every candidate up to twice that candidate's own uncertainty. In particular, well-covered policies are compared at nearly full strength. Poorly covered ones are automatically discounted. The abstract selection result permits policy-specific coverage conditions, but every candidate still needs a valid confidence bound. Missing support generally prevents a nontrivial certificate without further assumptions. The evaluation reports coverage diagnostics, namely the effective sample size of the weights and estimates $C(\pi, b)$ per policy. Against this background, (29) states that honesty about coverage is built directly into the selection rule.

The preceding results govern how the system learns to spend effort. They do not yet authorize machine acceptance. That final decision requires a separate finite-sample statement about the probability of releasing a material defect, which is the subject of the next subsection.

7.4 Calibrated acceptance and end-to-end assurance

The preceding learning procedures determine how the system gathers evidence and allocates review. Selective acceptance addresses a different question. It determines when the evidence is sufficiently strong for the system to release a repair without further human review. This subsection first certifies the machine channel, then combines it with human review risk, and finally states what remains valid under bounded deployment shift.

Definition 11 (selective acceptance). *In the risk coverage framework of selective classification (El-Yaniv and Wiener, 2010), let $\hat{q}(x)$ be an estimate of the probability of a material error after all machine checks. For threshold τ, the machine acceptance indicator is*

$$A_\tau(x) = \mathbf{1}\{\hat{q}(x) \le \tau \text{ and } HardGate(x) = 1\}.$$

All other cases are escalated. The selective risk and the coverage are

$$R(\tau) = \mathbb{P}(\text{material miss} | A_\tau = 1) \text{ and } Cov(\tau) = \mathbb{P}(A_\tau = 1).$$

Two complementary guarantees are relevant for the choice of τ. The first is an exact binomial bound for the auditing threshold. The second is a distribution-free procedure that controls a marginal episode loss for a nested family of acceptance rules. The first requires i.i.d. episodes; the second uses exchangeable loss functions. Both are stated at the level of amendment episodes, because obligations within an episode share sources, drafts, and reviewers and are therefore not independent from each other. For calibration episodes $i = 1, ..., K$ that are i.i.d. under the calibration distribution, we define the level indicators

$$\begin{aligned} A_{\tau,i} =& \mathbf{1}\{\text{episode } i \text{ contains at least one machine-accepted repair at threshold } \tau\}, \\ M_{\tau,i} =& \mathbf{1}\{\text{episode } i \text{ contains at least one machine-accepted material miss at threshold } \tau\}, \end{aligned}$$

with counts $N_\tau = \sum_i A_{\tau,i}$ and $X_\tau = \sum_i A_{\tau,i} M_{\tau,i}$. The selective risk of Definition 11 is then the conditional probability $R(\tau) = \mathbb{P}(M_\tau = 1 | A_\tau = 1)$ per episode. With this, confidence limits are functions of observable counts rather than the unknown probability itself.

Lemma 7 (exact binomial upper bound). *Let $X \sim \mathrm{Bin}(N, p)$, where $N \ge 0$ is a fixed integer and $p \in [0, 1]$. For $0 < \alpha < 1$, define the one-sided Clopper–Pearson upper limit by*

$$U_\alpha(x; N) := \sup \{q \in [0, 1] : \mathbb{P}_q(\mathrm{Bin}(N, q) \le x) \ge \alpha\}, \tag{30}$$

with $U_\alpha(0; 0) = 1$. Then

$$\mathbb{P}_p(U_\alpha(X; N) \ge p) \ge 1 - \alpha.$$

Lemma 7 supplies the statistical building block for deployment auditing. Applied to the accepted calibration episodes, it converts the observed number of material misses into a one-sided upper confidence limit. Since the threshold may itself be chosen after inspecting the same calibration sample, the next proposition makes the certificate simultaneous over a finite threshold grid.

Proposition 12 (selection-valid audit of the selective risk). *Let Θ be a non-empty finite threshold grid and $0 < \delta < 1$. For each fixed $\tau \in \Theta$, let $(A_{\tau,i}, M_{\tau,i})_{i=1}^K$ be i.i.d., where $A_{\tau,i} \in \{0, 1\}$ records machine acceptance and $M_{\tau,i} \in \{0, 1\}$ records a material miss. Assume that the complete workflow for each threshold, including the fitted risk model, hard gate, and candidate grid Θ, is fixed independently of these calibration episodes. Miss labels must be obtained by valid adjudication for the workflow evaluated at each threshold. Labels observed only on a selectively reviewed subset do not automatically satisfy these assumptions. Define*

$$R(\tau) := \begin{cases} \mathbb{P}(M_\tau = 1 \mid A_\tau = 1), & \mathbb{P}(A_\tau = 1) > 0, \\ 0, & \mathbb{P}(A_\tau = 1) = 0, \end{cases} \qquad N_\tau := \sum_{i=1}^K A_{\tau,i}, \qquad X_\tau := \sum_{i=1}^K A_{\tau,i} M_{\tau,i}.$$

Conditionally on the acceptance pattern $(A_{\tau,i})_{i=1}^{K}$*, and hence also conditionally on* N_τ*,*

$$X_\tau \mid N_\tau \sim \mathrm{Bin}(N_\tau, R(\tau))\,.$$

Let $\widehat{\tau} \in \Theta$ *be selected by any rule based on the same calibration data, and put*

$$\overline{R}(\tau) := U_{\delta/|\Theta|}(X_\tau; N_\tau).$$

Then

$$\mathbb{P}\big(R(\widehat{\tau}) \le \overline{R}(\widehat{\tau})\big) \ge 1 - \delta. \tag{31}$$

Consequently, if deployment is permitted only when $\overline{R}(\widehat{\tau}) \le \varepsilon$*, then, on the coverage event in* (31)*,* $R(\widehat{\tau}) \le \varepsilon$*. For a fresh deployment episode from the same episode law, independent of the calibration sample, its joint machine miss risk*

$$J(\tau) := \mathbb{P}(M_\tau = 1, A_\tau = 1) = \mathbb{P}(A_\tau = 1)R(\tau)$$

therefore satisfies $J(\widehat{\tau}) \le R(\widehat{\tau}) \le \varepsilon$ *on the same event.*

Proposition 12 supports periodic approval or audit of a threshold selected from a prespecified grid. The second route calibrates a nested acceptance rule directly. It does not require $\hat{q}$ to be a correctly calibrated probability, but it still requires a score or another ordered rule that induces the monotone loss family. It requires the boundedness, regularity, and exchangeability of loss functions stated in Theorem 6 (see also Assumption A7, Appendix A).

Theorem 6 (conformal risk control for selective acceptance; Angelopoulos et al., 2024). *Let* $K \ge 1$ *be an integer, let* $\Lambda = [\lambda_{\min}, \lambda_{\max}] \subset \mathbb{R}$ *be a non-empty compact interval, and let* $\ell_1, \ldots, \ell_{K+1} : \Lambda \to [0,1]$ *be exchangeable random loss functions. Suppose that, almost surely, every* ℓ_i *is non-increasing and right-continuous and satisfies* $\ell_i(\lambda_{\max}) = 0$*. For* $\varepsilon \in [1/(K+1), 1]$*, define*

$$\widehat{\lambda} := \min\left\{\lambda \in \Lambda : \frac{\sum_{i=1}^{K} \ell_i(\lambda) + 1}{K+1} \le \varepsilon\right\}. \tag{32}$$

Then

$$\mathbb{E}\Big[\ell_{K+1}(\widehat{\lambda})\Big] \le \varepsilon, \tag{33}$$

where the expectation is over both the calibration episodes and the new episode.

Suppose each episode has a finite candidate set whose drafts, defect labels, scores, and hard-gate outcomes are fixed as the threshold varies. For the strict acceptance rule $\widehat{q} < \tau$, take $\lambda = -\tau$, so larger λ means stricter acceptance. Choose $\lambda_{\max}$ to reject every candidate. The episode-level loss

$$\ell_i(\lambda) = \mathbf{1}\{\text{episode } i \text{ contains a machine-accepted repair with a material defect}\}$$

is then non-increasing and right-continuous. These properties do not follow from thresholding alone if changing the threshold changes drafting, routing, or interactions between edits. In that case the full loss family needs a separate monotonicity argument, or the workflow may use the finite-grid audit in Proposition 12 instead. With the non-strict rule $\widehat{q} \le \tau$, the corresponding loss is generally only left-continuous. One should therefore use the strict convention or a finite threshold grid. In the strict case the left-hand side of (33) is the marginal probability that a new episode contains at least one machine-accepted material miss. A materiality-weighted fraction loss instead controls the expected defect fraction and, by itself, does not control the probability of any miss. If Λ is a finite ordered grid, right-continuity is unnecessary. For the displayed calibration rule, $\varepsilon = 10^{-2}$ requires $K \ge 99$ for its feasible set to be non-empty. This is not a guarantee of useful machine coverage or a high-probability audit. A rule that always defers has zero machine loss regardless of this sample-size condition.

Observe that the two routes in Proposition 12 and Theorem 6 are operationally complementary. Proposition 12 gives a high-probability and selection-valid certificate for the conditional selective risk of a threshold (chosen via grid) and suits the periodic audit. Theorem 6 gives a marginal expected-loss guarantee for the stated calibration procedure. Repeated adaptive recalibration or workflow changes require a fresh validity argument and are not covered automatically. Moreover, with the indicator loss, it yields the episode-level machine term $\varepsilon_{\text{machine}}$ directly. Either way, calibration is assessed by amendment episode and the legal instrument family. For a complete policy update whose machine channel carries an episode-level budget and whose m' escalated repairs are adjudicated by humans with conditional error risks $p_h(x)$, the composed guarantee is as follows.

Proposition 13 (composed assurance with imperfect human review). *Under the deployment law $\mathbb{P}$, let E_{M} be the event that an episode contains a material defect among its machine-accepted repairs, and suppose that*

$$\mathbb{P}(E_{\mathrm{M}}) \leq \varepsilon_{\mathrm{M}}.$$

Let N_{H} be the almost surely finite number of escalated repairs in the episode. For $j \geq 1$, let H_j be the event that the expert assigned to the jth escalation fails to detect a material defect, and let $\mathcal{F}_{j-1}$ contain the routing history available before that review. Assume that $\{j \leq N_{\mathrm{H}}\} \in \mathcal{F}_{j-1}$ and that there is an $\mathcal{F}_{j-1}$-measurable bound $p_j \in [0,1]$ such that, on $\{j \leq N_{\mathrm{H}}\}$,

$$\mathbb{P}(H_j \mid \mathcal{F}_{j-1}) \leq p_j.$$

If every material miss is caused either by machine acceptance or by a failed human review, then

$$\mathbb{P}(\textit{any material miss}) \leq \varepsilon_{\mathrm{M}} + \mathbb{E}\left[\sum_{j=1}^{N_{\mathrm{H}}} p_j\right]. \tag{34}$$

No independence among machine and human errors, or among the human errors, is required.

If the number of escalations is the deterministic value m' and repair j is assigned to expert h_j with the deterministic bound p_{h_j}, then (34) reduces to

$$\mathbb{P}(\text{any material miss}) \leq \varepsilon_{\mathrm{M}} + \sum_{j=1}^{m'} p_{h_j}. \tag{35}$$

For Theorem 6 with the episode-level indicator loss, the machine premise is a marginal bound over the calibration and deployment episodes. For Proposition 12, (34) applies conditionally on the calibration data on its simultaneous-coverage event. Hence, the composed deployment bound is itself a confidence statement with probability at least $1-\delta$ over calibration samples, provided the human-review bounds remain valid under the same conditioning. If a single unconditional bound is needed instead, the calibration failure probability may be added to the right-hand side, giving the conservative bound

$$\mathbb{P}(\text{any material miss}) \leq \min\left\{1, \varepsilon_{\mathrm{M}} + \delta + \mathbb{E}\left[\sum_{j=1}^{N_{\mathrm{H}}} p_j\right]\right\}.$$

If, instead, the machine channel supplies valid bounds for individual machine-accepted repairs, the same argument may replace ε_{M} by the corresponding union-bound term. For a fixed collection of m repairs with $\mathbb{P}(D_r) \leq \varepsilon_r$, this term is $\sum_{r=1}^{m} \varepsilon_r$. When the accepted repairs are selected adaptively, the per-repair bounds must themselves remain valid after that selection, for example through predictable conditional bounds or a simultaneous selection-valid certificate. Such bounds do not follow from Theorem 6 or Proposition 12, which certify an episode-level machine term.

Proposition 14 (assurance under bounded deployment shift). *Let $\mathbb{P}$ and $\mathbb{Q}$ be the calibration and deployment laws on the episode space, respectively, with $\mathbb{Q} \ll \mathbb{P}$. Let A be the machine-acceptance event and let $B \subseteq A$ be the joint episode-level machine-miss event. Suppose that*

$$\mathbb{P}(B) \leq \varepsilon_{\mathrm{M}} \qquad \textit{and} \qquad \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} \leq \Lambda \quad \mathbb{P}\textit{-almost surely on } A.$$

Then

$$\mathbb{Q}(B) \leq \Lambda \varepsilon_{\mathrm{M}}. \tag{36}$$

Consequently, if the human-review bounds in Proposition 13 are valid under $\mathbb{Q}$, then

$$\mathbb{Q}(\textit{any material miss}) \leq \Lambda \varepsilon_{\mathrm{M}} + \mathbb{E}_{\mathbb{Q}}\left[\sum_{j=1}^{N_{\mathrm{H}}} p_j\right]. \tag{37}$$

The joint form in Proposition 14 is essential. If only the conditional selective risk is bounded under $\mathbb{P}$, say

$$\mathbb{P}(M = 1 \mid A) \leq \rho,$$

then the joint density-ratio assumption yields, whenever $\mathbb{Q}(A) > 0$,

$$\mathbb{Q}(M = 1 \mid A) \leq \Lambda \rho \, \frac{\mathbb{P}(A)}{\mathbb{Q}(A)}. \tag{38}$$

A clean factor-Λ transfer of conditional risk instead requires

$$\frac{\mathrm{d}\mathbb{Q}(\,\cdot \mid A)}{\mathrm{d}\mathbb{P}(\,\cdot \mid A)} \leq \Lambda \quad \mathbb{P}(\,\cdot \mid A)\text{-almost surely.}$$

Likewise, a drift monitor supports (36) only if it provides a valid upper certificate $\overline{\Lambda}$ for the density ratio. If machine acceptance is allowed only when $\overline{\Lambda} \leq \Lambda_{\max}$ and is suspended otherwise, then the machine term is bounded by $\Lambda_{\max}\varepsilon_{\mathrm{M}}$ whenever the certificate is valid.

Propositions 13 and 14 complete the assurance chain. The machine certificate controls the released machine channel. The expert bounds account for escalated cases. The density ratio bound states how much of the machine guarantee can be transferred when the deployment population differs from calibration. None of these terms repairs an omitted source or dependency edge. Those failures remain governed by the structural assumptions behind the impact cone and must be evaluated separately.

7.5 Expert routing, drift response, and conservative policy improvement

This subsection closes the theoretical development by addressing three decisions that remain before the algorithm can be stated. The first concerns the value of an additional expert when reviewer errors are correlated. The second concerns the action taken when the legal or statistical environment changes. The third concerns the evidence required before a learned orchestration policy may replace the approved baseline. These decisions connect the routing, drift, and assurance results above to the implementation in Section 8.

7.5.1 Cost sensitive selection of correlated experts

The one-step router in Proposition 10 selects the action with the smallest conditional risk and review burden. For an expert h, a materiality weighted version of this criterion is

$$w(x)p_h(x) + c_h + \lambda_d d_h, \tag{39}$$

where $w(x)$ is the materiality of case x, $p_h(x)$ is the expert's conditional miss probability, c_h is the review cost, and d_h is the delay. The error probability can be estimated from adjudicated outcomes using reviewer error models such as Dawid and Skene (1979). When a second opinion is considered, its value depends on the joint error probability rather than the number of reviewers alone. This is related to the role of error dependence in classifier ensembles (Kuncheva and Whitaker, 2003).

Proposition 15 (value of a correlated second opinion). *Fix a case x. Let M_1 and M_2 be the miss indicators of two experts, with $0 < p_i := \mathbb{P}(M_i = 1) < 1$ and $\mathrm{Corr}(M_1, M_2) = \varrho$. Suppose that the dual review scheme misses a material defect exactly when both experts miss it. Put $q_i = 1 - p_i$ and $s_{12} = \sqrt{p_1 q_1 p_2 q_2}$. Then*

$$\mathbb{P}(M_1 = 1, M_2 = 1) = p_1 p_2 + \varrho s_{12}. \tag{40}$$

If expert 1 is already assigned and the incremental burden of adding expert 2 is $\widetilde{c}_2 = c_2 + \lambda_d d_2$, then adding expert 2 strictly reduces the objective if and only if

$$w(x)\,(p_1 - p_1 p_2 - \varrho s_{12}) > \widetilde{c}_2. \tag{41}$$

The admissible correlation is restricted by

$$\frac{\max\{0, p_1 + p_2 - 1\} - p_1 p_2}{s_{12}} \le \varrho \le \frac{\min\{p_1, p_2\} - p_1 p_2}{s_{12}}. \tag{42}$$

In particular, if $p_1 = p_2 = p$ and $\varrho = 1$, then the two miss indicators coincide almost surely, the dual review miss probability equals p, and the risk reduction from the second opinion is zero.

Proposition 15 shows why two reviews need not provide the protection suggested by the reviewer count. A second opinion is useful when it reduces the joint miss probability enough to justify its cost and delay. In practice, source exposure, specialization, and briefing are relevant insofar as they affect the dependence between reviewer errors. The routing system should therefore estimate or conservatively bound joint miss risk whenever paired adjudicated outcomes are available.

7.5.2 Drift response

Let $P_t(X, Y)$ be the joint law of task features and adjudicated outcomes at legal time t. Amendments, entity changes, new guidance, model updates, and reviewer turnover may change this law. A drift monitor in the tradition surveyed by Gama et al. (2014) may compare source novelty, graph topology, embedding distributions, calibration error, escalation frequency, and reviewer disagreement with prespecified alarm thresholds.

When an alarm fires, machine acceptance is suspended or narrowed while the normative state is updated and the acceptance rule is recalibrated. The alarm alone does not establish the density ratio bound required by Proposition 14. The machine guarantee can be re-established by valid calibration for the new deployment law, or transferred from the old law using a valid density-ratio certificate $\overline{\Lambda}$. Recalibration need not estimate a density ratio. If the monitor also satisfies the detection assumptions of Corollary 1, the alarm triggers a restart of the route learner. The same operational event can therefore activate two distinct safeguards. It suspends unsupported machine acceptance and restarts statistical estimates that no longer describe the current phase.

7.5.3 Conservative policy improvement

Let π_0 be the approved baseline orchestration policy and let π_1 be a candidate. In the strictest instantiation, π_0 sends every material repair to mandatory human review. Define the risk difference and cost saving by

$$D_R = \mathrm{Risk}(\pi_1) - \mathrm{Risk}(\pi_0), \qquad D_C = \mathrm{Cost}(\pi_0) - \mathrm{Cost}(\pi_1).$$

Let U_R be an upper confidence bound for D_R and let L_C be a lower confidence bound for D_C. For a tolerated risk increase $\Delta_{\text{safe}} \geq 0$, the candidate is adopted only when

$$U_R \leq \Delta_{\text{safe}} \qquad \text{and} \qquad L_C > 0. \tag{43}$$

This rule follows the high-confidence policy improvement pattern of Thomas et al. (2015a).

Theorem 7 (conservative adoption). *Assume that π_0 is fixed and that π_1 is fixed independently of the evaluation sample. Alternatively, assume that the confidence bounds below remain valid after the selection of π_1. Suppose that*

$$\mathbb{P}(D_R > U_R) \leq \delta_R, \qquad \mathbb{P}(D_C < L_C) \leq \delta_C. \tag{44}$$

If π_1 is adopted exactly when (43) holds, then

$$\mathbb{P}\left(\pi_1 \text{ is adopted and } D_R > \Delta_{\text{safe}}\right) \leq \delta_R, \tag{45}$$

$$\mathbb{P}\left(\pi_1 \text{ is adopted and } D_C \leq 0\right) \leq \delta_C. \tag{46}$$

Consequently,

$$\mathbb{P}\left(\pi_1 \text{ is adopted and } \{D_R > \Delta_{\text{safe}} \text{ or } D_C \leq 0\}\right) \leq \delta_R + \delta_C. \tag{47}$$

In particular, choosing $\delta_R = \delta_C = \delta/2$ controls the probability of any unsafe or cost ineffective adoption by δ.

The bounds in (44) may be constructed from held-out amendment episodes using a concentration result suited to the relevant loss. Lemma 7 applies when the target is a binomial risk. Theorem 5 provides a policy selection guarantee for its specified value estimator and candidate class. Neither result automatically supplies both bounds in (44). Their validity must be established for the risk and cost estimands used by the adoption rule.

The practical consequence is that a candidate does not replace the baseline because its point estimates look preferable. It is adopted only when the upper bound on additional risk remains within the approved tolerance and the lower bound on cost saving is positive. The adoption decision is therefore a statistical test whose error budget is explicit. A bound on risk increase does not by itself establish an absolute risk ceiling. Deployment also requires the relevant absolute risk certificate, such as (51).

8 End-to-end algorithm and implementation

This section combines the preceding formal components into a specified amendment-to-policy pipeline. It first combines source validation, impact analysis, constrained repair, orchestration, and adjudication in one updating algorithm. It then states the end-to-end guarantee for the resulting policy and specifies the audit record required to reconstruct that guarantee. In this way, the section connects the mathematical results to the records and decisions that the system must produce in practice.

Figure 1 summarizes the technical architecture. It also makes the separation principle visible. The legal graph, applicability semantics, impact cone, and hard feasibility checks belong to the normative layer. Calibration and learning affect retrieval, verification, revision, routing, and acceptance, but cannot alter the legal propositions against which the repair is assessed.

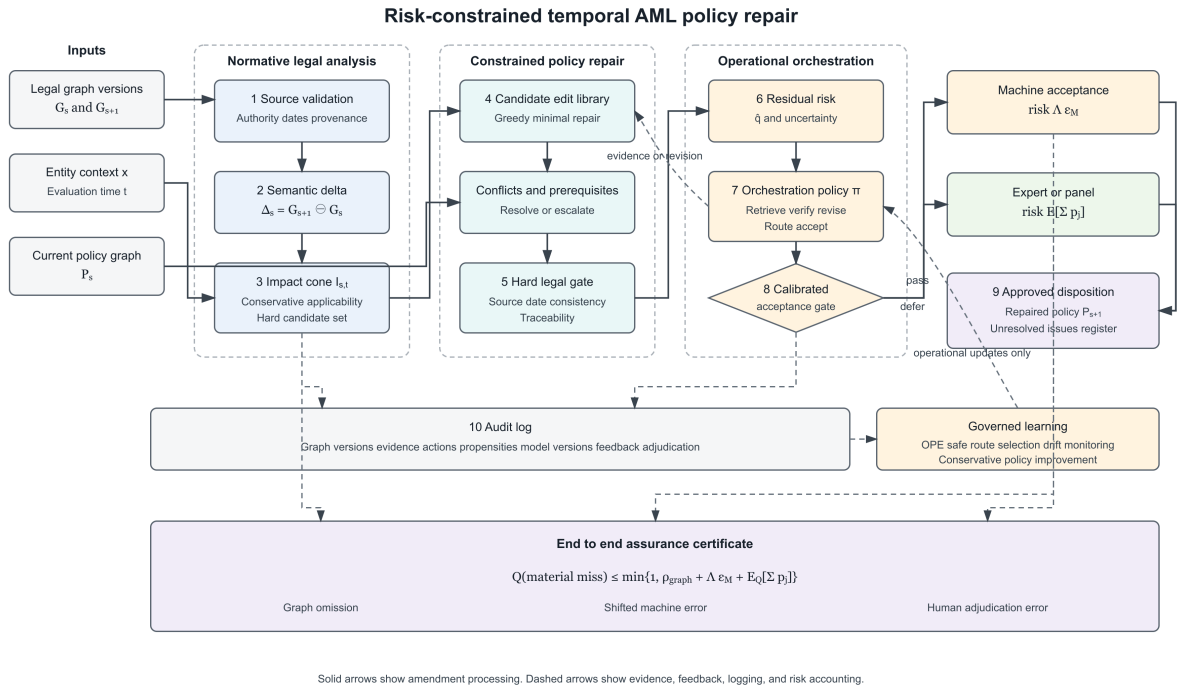

Figure 1: System architecture

8.1 Updating algorithm

At each ingestion event, the system authenticates the source and updates the normative layer through the ingestion operator of Definition 1. It then computes the semantic delta and the impact cone, resolves entity facts where possible, and treats every retained obligation as a coverage requirement for the repair stage. The repair engine returns candidate edits and explicit escalations. The orchestration policy determines whether additional evidence, verification, revision, or expert review is required. Machine acceptance is available only after the hard gate and the calibrated risk gate both pass.

Algorithm 1 separates deterministic legal-data operations from stochastic agent operations and human adjudication. The knowledge time s indexes the graph and policy versions, while t remains the legal evaluation time. This distinction follows the temporal convention of Section 3 and avoids using one index for two different notions of time.

Algorithm 1 Risk-constrained temporal AML policy repair

Require: Legal graph versions G_s and G_{s+1}, entity context x, current policy P_s, legal evaluation time t, risk budget ε, and shift bound Λ

Ensure: Proposed policy version P_{s+1}, unresolved-issues register, and an assurance certificate only for the scope whose conditions have been established

1: Validate the source versions, authority metadata, effective dates, and provenance under Assumption A1
2: Compute the semantic delta $\Delta_s = G_{s+1} \ominus G_s$ and attach confidence and provenance
3: Construct the impact cone $I_{s,t}$ on the union graph according to Definition 7, retain every member as a hard candidate, and rank review priority using the path score of Lemma 2
4: Generate edits from $\mathcal{E}(P_s)$; use Theorem 3 only when its coverage assumptions hold, and otherwise perform constrained search or escalate
5: Check conflicts and prerequisites; escalate certified infeasibility or unresolved search, distinguishing an infeasibility certificate from a timeout
6: Reject every candidate that fails a source, date, authority, consistency, or traceability condition of the hard gate in Proposition 8
7: Estimate the residual material error risk $\widehat{q}$ and its uncertainty for every surviving candidate
8: Let the orchestration policy π select among retrieval, verification, revision, expert routing, and acceptance as specified in Sections 6 and 7
9: Apply an acceptance rule certified at the amendment-episode-level under Proposition 12 or Theorem 6; do not treat that certificate as a per-edit bound
10: Escalate if the certified risk exceeds the allocated budget, a novelty or drift alarm is active, or a material $\perp$ value remains
11: Recheck the combined approved edits and their control impacts; apply only approved changes and record their rationale and citations
12: Log the states, actions, propensities, model versions, evidence, feedback, and adjudication required by Lemmas 3–6 and Theorem 5

The order of these operations is an implementation invariant. In particular, the system may rank the members of the impact cone, but it may not remove one from consideration without an approved disposition. Every retained obligation must therefore be resolved by a machine-accepted repair or passed to human adjudication. Escalation records an outstanding decision and does not close it. A claim about a completed policy update requires resolution of every material issue in its certified scope. The assurance result below also requires explicit accounting for pre-existing defects and interactions among edits.

The point of the architecture is that its guarantees compose. The following result combines the structural, machine, and human failure channels permitted by Algorithm 1.

Theorem 8 (end-to-end assurance for Algorithm 1). *Consider one amendment episode. Let $\mathbb{P}$ be the reference law under which the machine channel was calibrated and let $\mathbb{Q}$ be the deployment law. Let M be the event that the completed policy update contains a material documentary or control-design miss within the declared legal scope. Let B be the event that an affected applicable obligation falls outside the computed impact cone, and let E_{M} be the event that the episode contains a material defect among its machine-accepted repairs. Suppose that*

$$\mathbb{Q}(B) \le \rho_{\mathrm{graph}}, \qquad \mathbb{P}(E_{\mathrm{M}}) \le \varepsilon_{\mathrm{M}}.$$

Let A be the event that the episode contains a machine-accepted repair and assume that $E_{\mathrm{M}} \subseteq A$, $\mathbb{Q} \ll \mathbb{P}$, and

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} \le \Lambda \quad \mathbb{P}\text{-almost surely on } A.$$

Let N_{H} be the almost surely finite number of escalations. For $j \geq 1$, let H_j be the event that the jth escalated repair contains a material defect that is missed in adjudication, and let $\mathcal{F}_{j-1}$ contain the routing history available before that adjudication. Assume that $\{j \leq N_{\mathrm{H}}\} \in \mathcal{F}_{j-1}$ and that there is an $\mathcal{F}_{j-1}$-measurable $p_j \in [0,1]$ satisfying

$$\mathbb{Q}(H_j \mid \mathcal{F}_{j-1}) \leq p_j \quad \text{on } \{j \leq N_{\mathrm{H}}\}.$$

Finally, assume the exhaustive-disposition invariant of Algorithm 1, and that every escalated material issue has been adjudicated before a completed update is certified. Assume also that the starting policy has no pre-existing material defect outside the reviewed scope and that defects introduced by edits, including interactions and effects on otherwise unchanged obligations, are included in the machine or human error events. In particular, require the failure-accounting condition

$$M \subseteq B \cup E_{\mathrm{M}} \cup \bigcup_{j \geq 1} \big(\{j \leq N_{\mathrm{H}}\} \cap H_j\big).$$

Merely routing each retained obligation does not imply this inclusion. Under these assumptions,

$$\mathbb{Q}(M) \leq \min\left\{1, \rho_{\mathrm{graph}} + \Lambda\varepsilon_{\mathrm{M}} + \mathbb{E}_{\mathbb{Q}}\left[\sum_{j=1}^{N_{\mathrm{H}}} p_j\right]\right\}. \tag{48}$$

No independence among the three channels or among the human errors is required. If $N_{\mathrm{H}} = m'$ and $p_j = \bar{p}_{h_j}$ are deterministic, then

$$\mathbb{Q}(M) \leq \min\left\{1, \rho_{\mathrm{graph}} + \Lambda\varepsilon_{\mathrm{M}} + \sum_{j=1}^{m'} \bar{p}_{h_j}\right\}. \tag{49}$$

For the encoded evaluations, Theorem 1 excludes omissions under complete semantic change encoding and D-locality. Taking $\rho_{\mathrm{graph}} = 0$ for actual legal obligations additionally requires correct legal interpretation and applicability encoding, including Assumption A4. Source availability alone does not establish semantic correctness. In the relaxed regime, Proposition 3 can be used to construct ρ_{graph} by bounding the omission probability for each affected obligation and applying a union bound. When the machine premise is obtained from Proposition 12, equation (48) holds on the simultaneous calibration-success event. An unconditional statement then acquires the additive calibration term δ_{cal}. When the premise is obtained from Theorem 6 with the episode-level indicator loss, the machine term is marginal over calibration and deployment jointly. The density-ratio condition must in that case hold on the same joint probability space, or conditionally on each realized calibration sample.

If a residual event E_0 covers pre-existing or otherwise unaccounted-for defects and $\mathbb{Q}(E_0) \leq \rho_0$, the same union-bound proof adds ρ_0 to the right-hand side. Implementation and operating-effectiveness failures remain outside the documentary and control-design claim unless separately incorporated.

Under the failure-accounting assumptions above, exact graph assumptions, no deployment shift, a machine budget $\varepsilon_{\mathrm{M}} = \varepsilon$, and oracle adjudication, the bound reduces to $\mathbb{Q}(M) \leq \varepsilon$. Equation (48) therefore makes explicit how graph construction, calibrated machine acceptance, deployment shift, and human adjudication contribute to the episode-level material miss risk.

8.2 Assurance record and auditability

For every repaired policy version, the system preserves the identities and versions of the legal sources used, the semantic delta and impact cone, and the mapping from affected norms to policy clauses and controls. It also preserves the candidate edits, agent actions, verification results,

rejected alternatives, risk estimates, calibration version, thresholds, and allocated assurance budgets. Human reviewer identities or roles, feedback, adjudications, and unresolved disagreements remain linked to the corresponding repair. The final policy difference carries a rollback-capable version identifier.

The assurance certificate is consequently not only a reported probability. It is the constructive record of the assumptions entering equation (48). An auditor can trace the graph term to the source and dependency record, the machine term to the calibration and shift certificates, and the human term to the realized routing and adjudication history. The same record supplies the logged state, action, propensity, and outcome information required for the estimators in Section 7.

9 Practical evaluation framework

The present paper develops the formal and methodological foundations of the proposed system. It does not report the implementation and empirical evaluation of a complete prototype. Such an evaluation requires an adjudicated collection of regulatory amendments, entity-specific policy states, repeated expert assessments, and a temporally separated calibration and test design. Constructing this environment constitutes a substantial empirical undertaking of its own and is therefore beyond the scope of the present paper.

Nevertheless, the theoretical guarantees are intended to support an operational system and must ultimately be assessed against observable outcomes. This section therefore specifies how a prototype can be evaluated in practice. It identifies the principal evaluation claim, the experimental unit, the required comparison systems, and the outcome families corresponding to the components of the assurance argument. The resulting protocol prevents an operational improvement, such as a reduction in specialist workload, from being reported without the legal-safety certificate under which that improvement was obtained.

9.1 Primary evaluation claim

Let $H_e(\pi)$ denote the specialist minutes required when orchestration policy π processes amendment episode e. Let $M_e(\pi)$ be the corresponding material miss indicator and define the episode-level material omission risk by

$$R(\pi) := \mathbb{P}\big(M_e(\pi) = 1\big).$$

For the approved baseline π_0 and a candidate policy π_1, the primary workload estimand is

$$\Delta_H(\pi_0, \pi_1) := \mathbb{E}[H_e(\pi_0) - H_e(\pi_1)]\,. \tag{50}$$

The primary empirical claim would be that $\Delta_H(\pi_0, \pi_1) > 0$ while $R(\pi_1) \le \varepsilon$. The unknown risk cannot enter the deployment rule directly. Instead, the empirical study must construct a data-computed end-to-end upper certificate satisfying

$$\widehat{U}^{\mathrm{e2e}}_{\delta_R}(\pi_1) \le \varepsilon. \tag{51}$$

The certificate combines valid upper bounds for the graph, machine, deployment-shift, and human terms of Theorem 8. Proposition 12 may supply the machine component, but does not by itself certify the policy-level risk within the declared and fully accounted-for scope. Its audit route is the natural choice when the empirical study requires a high-probability certificate. The marginal guarantee of Theorem 6 may be reported separately, but cannot be substituted without changing the empirical claim.

The workload comparison should use a paired design in which π_0 and π_1 process the same amendment episodes. Proposition 16 in Appendix B gives a finite-sample lower confidence bound for the expected reduction in specialist minutes. Together with (51), it supports a joint claim of

workload reduction and an absolute risk ceiling. This is distinct from the relative risk-difference criterion in Theorem 7. A secondary analysis may consider total decision loss, including missed obligations, unnecessary edits, delay, and machine cost. Since raw accuracy hides how often the system acts and how often it escalates, the coverage and selective risk pair

$$\big(\mathrm{Cov}(\tau), R_{\mathrm{sel}}(\tau)\big)$$

should be reported as a curve over the acceptance threshold τ.

9.2 Experimental unit and temporal split

The top-level experimental unit is an amendment episode combined with an entity policy pair, rather than an isolated clause. Training should use earlier legal states. Calibration should use later states that precede the test period, while the newest amendment episodes remain untouched for final evaluation. This temporal split reflects the intended deployment setting and prevents information about a later legal state from entering an earlier model or threshold.

Paired comparisons should be used when two systems process the same tasks. Review order should be randomized and experts should be blinded to system identity where feasible, which also provides the intervention required by the identifiability result in Proposition 9. Reviewers who provide training feedback should remain separate from final evaluation reviewers whenever the available sample permits this separation.

9.3 Baselines and outcome families

The empirical comparison should include a manual or mandatory-expert review baseline. Retrieval and drafting comparisons should include keyword and embedding retrieval, conventional retrieval-augmented generation, GraphRAG without orchestration, single-agent drafting, and multi-agent drafting without independent verification. Routing comparisons should include fixed-threshold deferral, supervised learning to defer, a contextual bandit, and offline reinforcement learning when the support diagnostics of Section 7 justify its use. Oracle retrieval and oracle impact conditions should be used to separate retrieval, graph, reasoning, and drafting errors. In particular, the oracle impact condition permits the graph term ρ_{graph} in Theorem 8 to be assessed independently.

The outcome families are summarized in Table 1. Measures should first be aggregated at the amendment-episode-level before uncertainty is computed. Obligation-level results remain useful for error analysis, but should not be treated as independent episodes.

9.4 Statistical analysis

The primary workload comparison should be paired at the amendment-episode level. The workload and risk certificates may be combined by allocating separate error probabilities δ_H and δ_R. Boole's inequality then gives joint confidence at least $1 - \delta_H - \delta_R$ without requiring the two certificates to be independent. Proposition 16 states this argument formally.

The sampling unit used for uncertainty must reflect the dependence in the evaluation data. Proposition 16 applies when the paired episode-level workload differences are independent. If several episodes share a legal instrument or another source of dependence, inference should instead operate on independent clusters or use a bootstrap that resamples complete clusters. A multilevel model may additionally separate average system effects from heterogeneity across episodes, entity policy pairs, obligations, and reviewers when the replication structure supports the required variance components. A formal specification is provided in Appendix B.

The primary comparison, legal risk tolerance, and confidence allocation should be specified before the final test episodes are opened. The empirical application of this protocol, including

Table 1: Outcome families and representative evaluation measures.

Outcome family	Representative measures
Legal safety	Material obligation recall, severe false-negative rate, and unresolved ambiguity rate
Evidence	Citation entailment, source authority, temporal validity, and provenance completeness
Repair quality	Expert sufficiency, contradiction rate, unnecessary-edit rate, and semantic edit distance
Human system	Review minutes, escalation volume, override rate, and inter-reviewer agreement
Learning	Calibration, selective risk curve, regret, and off-policy effective sample size
Operations	Delay, token and tool cost, failure recovery, and reproducibility across model runs

the construction of an adjudicated benchmark, prototype implementation, preregistered comparisons, and analysis of reviewer behaviour, is left to separate work. Consequently, the present paper establishes how the system should be evaluated, but does not claim that the proposed performance or workload improvements have already been demonstrated empirically.

10 Discussion and limitations

The principal contribution of the proposed framework is not an autonomous legal decision maker, but an assurance architecture for the controlled use of AI in policy maintenance. The architecture separates the normative state from the operational decisions that may be learned from data. It also connects the identification of affected obligations, the construction of a feasible repair, the allocation of machine and human effort, and the final acceptance decision. This connection is important because good performance in one component does not compensate for failure in another. A well calibrated drafting model cannot recover an obligation that is absent from the legal graph, and an accurate impact analysis does not establish that a proposed edit is legally sufficient. Theorem 8 makes this distinction explicit by assigning separate terms to graph omission, machine acceptance, deployment shift, and human adjudication.

The resulting framework should therefore be understood as a way of delimiting and documenting automation. It supports machine acceptance only when the represented legal state, the hard feasibility conditions, and the statistical certificate all permit it. Otherwise, the required system output is an escalation with the unresolved proposition and the available evidence. This is a substantive outcome rather than a system failure. In a legally sensitive setting, the ability to identify what cannot be resolved safely is part of the intended functionality.

10.1 Limitations of the system

The first limitation concerns the construction of the normative layer. The completeness result of Theorem 1 is conditional on stable identities across legal versions, a complete semantic delta, and the locality of legal evaluation with respect to the encoded dependency relation. These conditions are meaningful design requirements, but they are demanding in practice. Legal effects may depend on implicit doctrine, cross references, source hierarchy, transitional provisions, or interpretations that are not represented as local graph dependencies. The three-valued semantics prevents an unknown entity fact from being treated as false, but it does not resolve the underlying legal ambiguity. An incorrect interpretation can also be represented consistently. Consequently, source authentication, legal modelling, identity alignment, and dependency validation remain

expert tasks. In the relaxed setting, the graph term ρ_{graph} must be supported by a separate audit rather than treated as negligible by assumption.

The second limitation concerns the repair representation. The repair problem operates on a finite library of admissible edits and a specified feasibility predicate. A legally sufficient and operationally suitable edit may be absent from that library, while interactions among clauses and controls may be richer than the coverage model captures. The existence result therefore applies only when a feasible repair is represented. Escalation preserves a safe disposition when this condition fails, but it does not produce the missing legal solution. The complexity results also show that exact minimal repair cannot generally be expected at scale. The greedy guarantee applies under the stated submodularity and cost conditions, while conflicts and prerequisite relations require the additional structures imposed in Propositions 5 and 6.

The orchestration layer introduces further modelling choices. Its losses, materiality weights, review costs, delay costs, and risk tolerances are not determined by the mathematical framework. They require organizational and legal governance. The finite horizon POMDP establishes a coherent decision problem and Proposition 7 establishes attainment under finite action and observation sets, but neither result provides an efficient solution for every realistic implementation. A conservative risk tolerance may also lead to low machine coverage and substantial expert workload. This is an intended safety tradeoff, although it may limit the operational benefit of the system.

Human review is not an oracle. Reviewer errors may be correlated because experts rely on the same sources, receive the same briefing, or share the same interpretive practice. Fatigue, time pressure, automation bias, and changes in review personnel may further affect the bounds p_j. Proposition 15 captures one aspect of dependence between two reviewers, but it does not model the full organizational process through which an adjudication is produced. The separation principle protects the normative layer from direct modification by operational learning. It cannot prevent an erroneous authenticated source, adjudicated interpretation, or manual graph update from entering that layer.

Finally, the formal model does not exhaust the engineering risks of a deployed system. Model updates and nondeterministic outputs require versioned testing. Internal policies, controls, and reviewer records require confidentiality and access controls. Source ingestion and tool use also create cybersecurity risks, including manipulated documents and instructions contained in retrieved material. Provenance records support investigation of these events, but the present analysis does not provide a security or privacy guarantee.

10.2 Limitations of the analysis

The mathematical guarantees are conditional and address different probability statements. They should not be read as interchangeable measures of legal correctness. Theorem 6 controls a marginal expected episode loss over calibration and deployment. Proposition 12 instead gives a high-probability certificate for a threshold selected from a fixed grid under its calibration conditions. Proposition 14 transfers a joint machine risk bound only under the stated absolute continuity and density ratio condition. In a high dimensional and changing legal environment, a useful upper bound for this density ratio may be difficult to establish. If the bound is large, the transferred guarantee may be too conservative to permit machine acceptance.

The end-to-end result composes valid component bounds. It does not make an invalid component valid, and the addition of the component terms can be loose when failures overlap. The benefit of this form is that it does not require independence. Its cost is that the certificate may be conservative, especially when many matters are escalated or when human error bounds are imprecise. The result also controls the defined event of a material miss. It does not by itself

measure the quality of legal reasoning, the clarity of the revised policy, the cost of unnecessary edits, or the effectiveness of the resulting controls in practice.

The learning results likewise depend on their stated data conditions. Off policy estimators require reliable logging and sufficient support for the actions selected by the candidate policy. The contextual bandit result relies on stationarity within the relevant context stratum. Offline sequential policy selection becomes uninformative when trajectory coverage is weak. Drift alarms can indicate that these conditions are doubtful, but an alarm alone does not certify the density ratio or restore exchangeability. These limitations are particularly relevant when rare legal amendments, changes in model versions, and selective human review produce limited and nonrepresentative feedback.

The analysis performed in this paper is primarily formal. Section 9 specifies an empirical evaluation protocol, but does not report results from a complete prototype. In particular, the paper does not yet establish graph recall on an independently adjudicated corpus, the legal sufficiency of generated repairs, the accuracy of reviewer error estimates, or the reduction in specialist workload. Proposition 16 provides a finite-sample certificate for a paired workload comparison under independent and bounded episode-level differences. Shared legal instruments, repeated entities, and repeated reviewers may violate this independence condition. Cluster based inference or a supported multilevel model is then required. Rare material misses and disagreement about the reference legal judgment may also produce wide uncertainty intervals even in a comparatively large study.

10.3 Implications for practical validation

These limitations determine the appropriate next step. A prototype should be evaluated prospectively on temporally ordered amendments, with the final test episodes separated from system development and threshold selection. The study should include several legal instruments, entity types, and reviewers. It should use independent legal adjudication, report disagreement, and preserve the amendment episode as the primary sampling unit. Component studies should test graph recall, impact closure, repair feasibility, evidence quality, reviewer behaviour, and calibration separately before the end-to-end claim is assessed. Stress tests should include missing dependency edges, unresolved entity facts, novel sources, model changes, and simulated deployment shift.

Accordingly, the framework provides testable conditions for safe deployment, but the satisfaction of those conditions remains an empirical matter. Until the evaluation described in Section 9 has been completed, the system should be regarded as a formally specified decision support system whose unresolved or insufficiently certified outputs require accountable human adjudication. If an assurance condition fails, the appropriate response is to narrow machine acceptance or escalate the case, rather than to infer safety from average predictive performance alone.

11 Conclusions

This paper develops a formal framework for the AI supported updating of AML/CFT policies. Its central premise is that policy maintenance cannot be reduced to the generation of revised text. It requires the representation of the applicable legal state, the identification of obligations affected by an amendment, the construction of a feasible policy repair, the allocation of machine and human effort, and a justified decision on whether the resulting repair may be accepted. The separation principle ensures that learning can improve these operational decisions without altering the normative content against which they are assessed.

The formal results establish conditions under which the impact cone contains every conservatively

applicable obligation whose encoded legal evaluation may have changed. They also characterize the computational difficulty of minimal policy repair and identify conditions under which a greedy procedure provides a controlled approximation. The orchestration results connect machine processing, expert routing, and calibrated acceptance. The final assurance result shows how graph omissions, machine errors, deployment shift, and imperfect human adjudication contribute to the probability of a material miss. The resulting guarantee is conditional on the validity of its component bounds. If one of these conditions cannot be established, the appropriate system response is escalation rather than unsupported machine acceptance.

The paper provides a theoretical and methodological foundation rather than an empirical validation of a complete prototype. The practical evaluation framework specifies how legal safety and specialist workload can be assessed jointly in future work. Such an evaluation will require temporally separated amendment episodes, independent legal adjudication, realistic entity contexts, and explicit tests of graph quality, repair sufficiency, reviewer behaviour, and distribution shift. Subject to this validation, the proposed framework offers a transparent basis for combining legal knowledge representation, constrained optimization, statistical learning, and accountable human review in regulatory policy maintenance.

References

Altman, E. (1999). *Constrained Markov Decision Processes*. Chapman & Hall/CRC.

Angelopoulos, A. N., S. Bates, E. J. Candès, M. I. Jordan, and L. Lei (2025). Learn then test: Calibrating predictive algorithms to achieve risk control. *Annals of Applied Statistics 19*(2), 1641–1662.

Angelopoulos, A. N., S. Bates, A. Fisch, L. Lei, and T. Schuster (2024). Conformal risk control. In *The Twelfth International Conference on Learning Representations*.

Auer, P., N. Cesa-Bianchi, and P. Fischer (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning 47*(2-3), 235–256.

Barrientos, M., K. Winter, and S. Rinderle-Ma (2026). Impact analysis of regulatory requirement changes on business process compliance. *Information and Software Technology 194*, 108079.

Bates, S., A. N. Angelopoulos, L. Lei, J. Malik, and M. I. Jordan (2021). Distribution-free, risk-controlling prediction sets. *Journal of the ACM 68*(6), 43:1–43:34.

Bradley, R. A. and M. E. Terry (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika 39*, 324–345.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics 90*(3), 414–427.

Charnes, A. and W. W. Cooper (1959). Chance-constrained programming. *Management Science 6*(1), 73–79.

Charusaie, M.-A., H. Mozannar, D. Sontag, and S. Samadi (2022). Sample efficient learning of predictors that complement humans. ICML 2022.

Chen, Z., L. D. Van Khoa, E. N. Teoh, A. Nazir, E. K. Karuppiah, and K. S. Lam (2018). Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: A review. *Knowledge and Information Systems 57*(2), 245–285.

Chen, Z., Q. Zhang, Z. Xiang, Z. Wei, L. Gao, X. Huang, Z. Zhang, and J. Su (2026). LegalGraphRAG: Multi-agent graph retrieval-augmented generation for reliable legal reasoning. *ACL 2026*.

Chow, Y., M. Ghavamzadeh, L. Janson, and M. Pavone (2018). Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research 18*(167), 1–51.

Christiano, P. F., J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei (2017). Deep reinforcement learning from human preferences. *NeurIPS 2017*.

Chvátal, V. (1979). A greedy heuristic for the set-covering problem. *Mathematics of Operations Research 4*(3), 233–235.

Constant, A. (2024). A bayesian model of legal syllogistic reasoning. *Artificial Intelligence and Law 32*, 441–462.

Davidson, R. R. and R. J. Beaver (1977). On extending the bradley–terry model to incorporate within-pair order effects. *Biometrics 33*(4), 693–702.

Dawid, A. P. and A. M. Skene (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 28*(1), 20–28.

Devroye, L., L. Györfi, and G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition*. Springer.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik 1*, 269–271.

Dinur, I. and D. Steurer (2014). Analytical approach to parallel repetition. STOC 2014.

Dudík, M., J. Langford, and L. Li (2011). Doubly robust policy evaluation and learning. ICML 2011.

Edge, D., H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson (2024). From local to global: A graph RAG approach to query-focused summarization.

El-Yaniv, R. and Y. Wiener (2010). On the foundations of noise-free selective classification. *Journal of Machine Learning Research 11*(53), 1605–1641.

Ershov, V. D. (2023). A case study for compliance as code with graphs and language models: Public release of the regulatory knowledge graph.

European Parliament and Council of the European Union (2024a). Directive (EU) 2024/1640 of 31 may 2024 on the mechanisms to be put in place by the member states for the prevention of the use of the financial system for the purposes of money laundering or terrorist financing. Official Journal of the European Union.

European Parliament and Council of the European Union (2024b). Regulation (EU) 2024/1624 of 31 may 2024 on the prevention of the use of the financial system for the purposes of money laundering or terrorist financing. Official Journal of the European Union.

Falcarin, P., P. Chowdhury, E. Carbone, T. Scantamburlo, R. Tripodi, and S. Vascon (2025). Legal requirements compliance using NLP and knowledge graphs. In *2025 IEEE 33rd International Requirements Engineering Conference Workshops*, pp. 412–418.

Feige, U. (1998). A threshold of ln n for approximating set cover. *Journal of the ACM 45*(4), 634–652.

Financial Action Task Force (2021). Opportunities and challenges of new technologies for AML/CFT. Technical report, Financial Action Task Force, Paris.

Fitting, M. (1985). A Kripke–Kleene semantics for logic programs. *Journal of Logic Programming 2*(4), 295–312.

Ford, L. R. (1957). Solution of a ranking problem from binary comparisons. *American Mathematical Monthly 64*(8, Part 2), 28–33.

Gama, J., I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia (2014). A survey on concept drift adaptation. *ACM Computing Surveys 46*(4), 44:1–44:37.

Garivier, A. and E. Moulines (2011). On upper-confidence bound policies for switching bandit problems. In *ALT'11: Proceedings of the 22nd international conference on Algorithmic learning theory*, pp. 174–188.

Gelman, A. and J. Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Governatori, G. (2013). Business process compliance: An abstract normative framework. *it - Information Technology 55*(6), 231–238.

Governatori, G. and A. Rotolo (2010). Changing legal systems: Legal abrogations and annulments in defeasible logic. *Logic Journal of the IGPL 18*(1), 157–194.

Guha, N., J. Nyarko, D. E. Ho, C. Ré, A. Chilton, et al. (2023). LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, Volume 36.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*, 663–685.

Ionides, E. L. (2008). Truncated importance sampling. *Journal of Computational and Graphical Statistics 17*(2), 295–311.

Jin, Y., Z. Yang, and Z. Wang (2021). Is pessimism provably efficient for offline rl? ICML 2021.

Kaelbling, L. P., M. L. Littman, and A. R. Cassandra (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence 101*, 99–134.

Karp, R. M. (1972). Reducibility among combinatorial problems. In R. E. Miller, J. W. Thatcher, and J. D. Bohlinger (Eds.), *Complexity of Computer Computations*, pp. 85–103. Plenum Press.

Kleene, S. C. (1952). *Introduction to Metamathematics*. D. Van Nostrand Company Inc.

Kuncheva, L. I. and C. J. Whitaker (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning 51*(2), 181–207.

Lai, T. L. and H. Robbins (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics 6*(1), 4–22.

Li, L., W. Chu, J. Langford, and T. Moon (2012). An unbiased offline evaluation of contextual bandit algorithms with generalized linear models. *Proceedings of Machine Learning Research 26*, 19–36.

Madras, D., T. Pitassi, and R. Zemel (2018). Predict responsibly: Improving fairness and accuracy by learning to defer. NeurIPS 2018.

Mao, A., M. Mohri, and Y. Zhong (2025). Mastering multiple-expert routing: Realizable h-consistency and strong guarantees for learning to defer. ICML 2025.

Mozannar, H. and D. Sontag (2020). Consistent estimators for learning to defer to an expert. ICML 2020.

Nemhauser, G. L., L. A. Wolsey, and M. L. Fisher (1978). An analysis of approximations for maximizing submodular set functions - i. *Mathematical Programming 14*, 264–294.

Prakken, H. and G. Sartor (2015). Law and logic: A review from an argumentation perspective. *Artificial Intelligence 227*, 214–245.

Precup, D., R. S. Sutton, and S. Singh (2000). Eligibility traces for off-policy policy evaluation. ICML 2000.

Rashidinejad, P., B. Zhu, C. Ma, J. Jiao, and S. Russell (2021). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. NeurIPS 2021.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association 89*, 846–866.

Rockafellar, R. T. and S. Uryasev (2000). Optimization of conditional value-at-risk. *Journal of Risk 2*, 21–41.

Sun, J., Z. Luo, and Y. Li (2025). A compliance checking framework based on retrieval augmented generation. *COLING 2025*.

Thomas, P., G. Theocharous, and M. Ghavamzadeh (2015a). High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning*, Volume 37 of *Proceedings of Machine Learning Research*, pp. 2380–2388. PMLR.

Thomas, P. S., G. Theocharous, and M. Ghavamzadeh (2015b). High-confidence off-policy evaluation. Proceedings of the AAAI Conference on Artificial Intelligence, 29(1).

von Wright, G. H. (1951). Deontic logic. *Mind 60*(237), 1–15.

Weiser, M. (1984). Program slicing. *IEEE Transactions on Software Engineering 10*(4), 352–357.

Wolsey, L. A. (1982). An analysis of the greedy algorithm for the submodular set covering problem. *Combinatorica 2*(4), 385–393.

Wu, Q., G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. Awadallah, R. W. White, D. Burger, and C. Wang (2024). AutoGen: Enabling next-gen LLM applications via multi-agent conversations. First Conference on Language Modeling.

Zhan, W., M. Uehara, N. Kallus, J. D. Lee, and W. Sun (2023). Provable offline reinforcement learning with human feedback. Open Review.

Appendix

A - Assumptions

A1 (legal source completeness). *All relevant legal source versions are available and correctly versioned.*

A2 (semantic delta recall). *Under the stable cross-version identity alignment, every node, edge, attribute, or evaluation-rule change that can affect the encoded evaluation is represented in* Δ_s. *Edge-type changes are included.*

A3 (dependency completeness / D-locality). *Evaluation of every norm factors through its rooted attributed D-ancestor subgraph, and every change to the corresponding evaluation rule is seeded by the semantic delta.*

A4 (applicability adequacy). *Applicability formulas correctly encode the declared legal scope, recorded entity facts are accurate, and missing facts produce* $\perp$ *rather than false. The full legal interpretation must also be validated before an encoded completeness result is read as a claim about actual law.*

A5 (edit feasibility). *The repair library contains a jointly feasible compliant repair or a jointly feasible escalation disposition for all identified requirements. Escalation denotes safe disposition, not a completed compliant policy repair.*

A6 (logged-policy adequacy). *Candidate learning policies remain within adequate support of logged data, and coverage is reported.*

A7 (episode exchangeability). *Theorem 6 requires exchangeable loss functions across calibration and deployment episodes for a workflow fixed before calibration. Proposition 12 and its binomial audit require i.i.d. episode pairs. A chronological data split alone establishes neither condition. If calibration is stratified, the relevant condition and certificate must hold within each declared stratum.*

A8 (feedback governance). *Reviewer feedback is attributable, versioned, and distinguishable from authoritative legal sources.*

A9 (bounded deployment shift). $\mathbb{Q} \ll \mathbb{P}$ *and a valid bound* $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P} \leq \Lambda$ *holds* $\mathbb{P}$*-almost surely on the acceptance event. For a data-selected rule, this statement must be valid under the same conditioning as its machine certificate, or on the joint calibration–deployment space. A drift alarm alone is not such a certificate.*

A10 (human error bounds). *The routed expert errors admit valid conditional bounds given the pre-review history, as in Propositions 13–14 and Theorem 8. Unconditional historical error rates do not automatically bound adaptively selected reviews. Any estimation-failure probability must also be included in the assurance budget.*

A11 (failure accounting and release). *A completed-update certificate requires resolved material escalations, a valid baseline within the declared scope, and accounting for interactions and collateral effects of edits. Any failure channel outside those in Theorem 8 requires a separate bound.*

B - Further theoretical results

Proposition 3 (robustness to missing edges). *Suppose that the seed set and the conservative applicability status of n are encoded correctly. For every true dependency edge e, let Z_e indicate whether e is captured, and suppose that the family $(Z_e)_e$ is mutually independent with $\mathbb{P}(Z_e = 1) \geq r$, $0 < r \leq 1$. Let n be an affected obligation possessing $k_n \geq 1$ pairwise edge-disjoint seed-to-n dependency paths in the true union graph, each of length at most the non-negative integer L_n. Then,*

$$\mathbb{P}(n \notin \hat{I}_{s,t}) \leq (1 - r^{L_n})^{k_n}, \tag{52}$$

where $\hat{I}_{s,t}$ is the cone computed on the encoded graph. Consequently, the expected number of missed affected obligations is at most $\sum_n (1 - r^{L_n})^{k_n}$ over affected n.

Lemma 2 (score computation) *Assume $0 < \gamma_e \leq 1$, $\lambda_{\text{path}} \geq 0$, and $w_n \geq 0$. Under the transform $e \to -\ln \gamma_e + \lambda_{\text{path}}$, every edge weight is non-negative and the maximization in*

$$score(n|\Delta_s) = w_n \cdot \max_{p:Seed(\Delta_s)\to n} \prod_{e\in p} \gamma_e \cdot \exp(-\lambda_{\text{path}}|p|) \tag{53}$$

is a multi-source shortest-path problem, or equivalently a single-source problem after adding a super-source with zero-weight edges to all seeds. Dijkstra's algorithm (Dijkstra, 1959) computes the transformed distances in $O((|V| + |E_D|) \log |V|)$ arithmetic operations with a binary heap. Scores are recovered as $w_n \exp(-d(n))$. Exactness here assumes exact real arithmetic; this is not a bit-complexity or floating-point accuracy guarantee. An unreachable node has distance $+\infty$ and score zero.

Proposition 16 (finite-sample certificate for the workload claim). *Assume that π_0 and π_1 are fixed independently of the evaluation sample, or that the bound below remains valid after the selection of π_1. For a fixed $B > 0$ and integer $E \geq 1$, let*

$$D_e := H_e(\pi_0) - H_e(\pi_1) \in [-B, B]$$

be the paired workload difference on amendment episode e. Suppose that $D_1, \ldots, D_E$ are independent and identically distributed, and put $\overline{D}_E := E^{-1} \sum_{e=1}^{E} D_e$. Then, for every $\delta_H \in (0, 1)$, with probability at least $1 - \delta_H$,

$$\overline{D}_E - B\sqrt{\frac{2\ln(1/\delta_H)}{E}} \leq \Delta_H(\pi_0, \pi_1). \tag{54}$$

If, in addition,

$$\mathbb{P}\Big(R(\pi_1) \leq \widehat{U}^{\text{e2e}}_{\delta_R}(\pi_1)\Big) \geq 1 - \delta_R,$$

then both certificates hold jointly with probability at least $1-\delta_H-\delta_R$. No independence between the workload and risk certificates is required. Consequently, a positive lower bound in (54) *together with* (51) *certifies a reduction in expected specialist minutes at material omission risk at most ε.*

Taking $\delta_H = \delta_R = \delta$ gives joint confidence at least $1 - 2\delta$. The assumption that the D_e are independent is essential for the stated Hoeffding bound. When dependence persists across amendment episodes, the unit entering Proposition 16 must be replaced by an independent cluster-level aggregate, or another confidence construction must be used.

Hierarchical evaluation model

Let Y_{eijr} be a binary error indicator for amendment episode e, entity policy pair i, obligation j, and model run or reviewer r. Let S_{eijr} indicate the system under evaluation. A possible multilevel logistic specification is

$$\operatorname{logit} \mathbb{P}(Y_{eijr} = 1 \mid u_e, v_i, w_j, b_r) = \alpha + \beta S_{eijr} + u_e + v_i + w_j + b_r. \tag{55}$$

The terms u_e, v_i, w_j, and b_r are zero-mean random intercepts for episodes, entity policy pairs, obligations, and runs or reviewers. They should be included only when the replication structure identifies the corresponding variance components and the model diagnostics support the specification (Gelman and Hill, 2007). Otherwise, a bootstrap that resamples complete independent clusters provides a more robust alternative while preserving the dependence within an amendment episode or legal instrument (Cameron et al., 2008).

C - Proofs

Proof of Proposition 1

Proof. We perform an induction over the length k of the update sequence. For $k = 0$, there is nothing to prove as the system is not updated. Assume the claim for sequences of length k and append one-step. If the step is an ingestion step $\mathcal{I}_r$, then its normative component is $\iota_r(\mathcal{N})$, independently of ω, so both the full sequence and the ingestion-only sequence apply the same next normative map. If the step is a learning step $\mathcal{L}$, it maps $(\mathcal{N}, \omega)$ to $(\mathcal{N}, \omega')$ by Definition 1, so the normative layer is unchanged on the left side and the ingestion-only sequence is unchanged on the right side. In either case, the normative layers agree after $k+1$ steps. Since by Definition 1 every transition factors into such steps, the claim holds for all finite update sequences. The listed objects (legal propositions, source ranking, validity interval, predicate) are components of $\mathcal{N}$, which proves the second part. □

Proof of Lemma 1

Proof. The proof relies on Kleene (1952) and Fitting (1985). We perform a structural induction on q. For an atom p, if x is undefined at p, then $[\![p]\!]_x = \bot \sqsubseteq [\![p]\!]_{x'}$ follows trivially. If x is defined at p, then x' extending x assigns the same value, which means that valuations are equal. For $\neg q$, the map $1 \to 0$, $0 \to 1$, $\bot \to \bot$ is monotone for $\sqsubseteq$ as the only strict order relations are $\bot \sqsubset 0$ and $\bot \sqsubset 1$. These are mapped to $\bot \sqsubset 1$ and $\bot \sqsubset 0$, respectively. Composing a monotone map with the induction hypothesis preserves the claimed inequality. For $q_1 \wedge q_2$, the strong Kleene conjunction $\wedge : \{0, 1, \bot\}^2 \to \{1, 0, \bot\}$ is monotone in each argument for $\sqsubseteq$. This can be verified by the strict moves $\bot \to 0$ and $\bot \to 1$ in one argument with the other fixed. If the fixed argument is 0, the conjunction is 0 before and after. If the fixed argument is 1, the conjunction moves $\bot \to 0$ or $\bot \to 1$, which are both allowed. If the fixed argument is $\bot$, the conjunction moves $\bot \to 0$ when the moving argument becomes 0 or stays $\bot$ when the moving argument becomes 1. Again, both are allowed. Disjunction is dual, and bounded quantifiers expand into finite conjunctions and disjunctions over the domain, which is fixed and finite. Therefore, the induction covers them, and monotone composition of monotone maps completes the induction. Since 0 and 1 are maximal in $\sqsubseteq$, the conditions $[\![q]\!]_x \in \{0, 1\}$ and $[\![q]\!]_x \sqsubseteq [\![q]\!]_{x'}$ force equality and complete the proof. □

Proof of Proposition 2

Proof. The proof consists of two parts, as in the proposition. (i) Let x' be a completion of x and $n \in \mathrm{App}_{s,t}(x')$, such that $t \in \nu(n)$ and $[\![q_n]\!]_{x'} = 1$. If $n \notin \mathrm{App}^{\dagger}_{s,t}(x)$, then, because validity is

already established, $[\![q_n]\!]_x = 0$. By Lemma 1, every extension of x evaluates q_n to 0, which is a contradiction. Thus, $n \in \mathrm{App}^{\dagger}_{s,t}(x)$ holds. (ii) Let $x \sqsubseteq x'$. If $n \in \mathrm{App}^{\dagger}_{s,t}(x')$, then $[\![q_n]\!]_{x'} \in \{1, \bot\}$. If the later value is $\bot$, Lemma 1 forces the earlier value to be $\bot$. If the later value is 1, the earlier value is either $\bot$ or 1. Thus, in either case, $n \in \mathrm{App}^{\dagger}_{s,t}(x)$. If $n \in \mathrm{App}_{s,t}(x)$, then $[\![q_n]\!]_x = 1$ and $[\![q_n]\!]_{x'} = 1$ by Lemma 1, hence, $n \in \mathrm{App}_{s,t}(x')$, which completes the proof. □

Proof of Theorem 1

Proof. We first consider norms that do not persist. A change of node type that makes an existing object enter or leave the norm set is an attribute change and seeds that object directly. If $n \in V^+ \cup V^-$, then $n \in \mathrm{Seed}(\Delta_s) \subseteq \mathrm{Reach}_D(\mathrm{Seed}(\Delta_s))$. By the absent convention, its evaluations differ. Hence, any such norm in the applicability union lies in $I_{s,t}$ already.

For what follows, assume that n is present in both versions. We prove the contrapositive. Write $S = \mathrm{Seed}(\Delta_s)$ and assume that $n \notin \mathrm{Reach}_D(S)$ in the union graph. Let $W_s = \mathrm{Anc}_D(n; G_s)$ and $W_{s+1} = \mathrm{Anc}_D(n; G_{s+1})$. Since every D-edge of either version is an edge of the union graph, no member of S belongs to either ancestor set.

First, $W_s = W_{s+1}$. Otherwise, suppose without loss of generality that $u \in W_{s+1} \setminus W_s$. A D-path from u to n in G_{s+1} must contain an edge absent from $E_D(G_s)$. Such an edge is newly added or has changed its dependency type. An endpoint of the last such edge lies in $\mathrm{ends}(E^+ \cup A^{\Delta}_E) \subseteq S$ and has a D-path to n in the union graph, contradicting $n \notin \mathrm{Reach}_D(S)$. The reverse inclusion follows symmetrically using $E^- \cup A^{\Delta}_E$. Write W for the common ancestor set.

Second, the rooted attributed subgraphs induced by W are identical. No node in W lies in A^{Δ}_V, and no changed, added, or removed edge within the induced subgraph has an endpoint in W. Otherwise, that node or endpoint would belong to S and reach n. A change to F_n would seed n itself. Thus, by D-locality,

$$\mathrm{Eval}_n(G_s, x, P) = \mathrm{Eval}_n(G_{s+1}, x, P).$$

This proves the contrapositive. Finally, an affected norm in $\mathrm{App}^{\dagger}_{s,t}(x) \cup \mathrm{App}^{\dagger}_{s+1,t}(x)$ therefore lies in both $\mathrm{Reach}_D(S)$ and the applicability union, and hence lies in $I_{s,t}$ by Definition 7. Assumption A2 guarantees that S contains every changed object, which completes the proof. □

Proof of Proposition 3

Proof. Fix n and let B_i be the event that the ith edge-disjoint path is broken. Path p_i is intact with probability

$$\prod_{e \in p_i} \mathbb{P}(Z_e = 1) \geq r^{|p_i|} \geq r^{L_n},$$

so $\mathbb{P}(B_i) \leq 1 - r^{L_n}$. The events B_i are independent because the paths are edge-disjoint and the edge indicators are mutually independent. If n is missed, every listed path is broken. Therefore,

$$\mathbb{P}(n \notin \hat{I}_{s,t}) \leq \mathbb{P}\left(\bigcap_{i=1}^{k_n} B_i\right) = \prod_{i=1}^{k_n} \mathbb{P}(B_i) \leq (1 - r^{L_n})^{k_n}.$$

Summing the miss indicators over affected obligations and applying linearity of expectation gives the second claim. □

Proof of Lemma 2

Proof. For every path p,

$$-\ln\left(\prod_{e\in p}\gamma_e\exp(-\lambda_{\text{path}}|p|)\right)=\sum_{e\in p}\left(-\ln\gamma_e+\lambda_{\text{path}}\right).$$

Because $0<\gamma_e\le 1$ and $\lambda_{\text{path}}\ge 0$, each transformed edge weight is non-negative. Moreover, $-\ln(\cdot)$ is strictly decreasing, so maximizing the path product is equivalent to minimizing the transformed path length. Adding the stated super-source converts the multi-source problem to an ordinary non-negative single-source shortest-path problem. If $d(n)$ is the resulting distance, then

$$\text{score}(n\mid\Delta_s)=w_n\exp(-d(n)).$$

The standard binary-heap implementation has the stated arithmetic-operation count. Numerical evaluation of logarithms and exponentials requires a separate precision choice. □

Proof of Proposition 4

Proof. The family

$$\mathcal{F}=\{R\subseteq\mathcal{E}(P_s):\Phi(P_s\oplus R,G_{s+1},x,t)=1\}$$

is a subset of the power set of a finite set, hence finite, and it is non-empty by hypothesis. A real-valued function on a non-empty finite set attains its minimum. □

Proof of Theorem 2

Proof. The proof consists of two parts, whereas the first part can be further separated into two elements: membership and hardness of the problem.

(i.a) *Membership.* A certificate is the repair R itself. By Definition 9, the certificate has polynomial length, all numerical inputs have polynomial-size rational encodings, and Φ and J are evaluable in polynomial time. The decision problem is therefore in NP.

(i.b) *Hardness.* Reduce from weighted set cover. Let an instance be given by the universe $U=\{o_1,\dots,o_N\}$, sets $S_1,\dots,S_m\subseteq U$ with non-negative rational weights $w_1,\dots,w_m$, and budget B. Construct a repair instance as follows: the impact cone $I_{s,t}$ consists of N obligations, one per universe element, each applicable and rendered unsatisfied by the update. The edit library contains edits $e_1,\dots,e_m$, where e_i has cost $\mathcal{C}^{\text{edit}}(e_i)=w_i$ and covers exactly the obligations in S_i. The remaining objective terms are switched off, that is, $\lambda=\mu=\xi=0$, which gives $J(R)=\sum_{e_i\in R}w_i$. Moreover,

$$\Phi(P_s\oplus R,G_{s+1},x,t)=1\quad\Longleftrightarrow\quad\bigcup_{e_i\in R}S_i=U.$$

The construction is polynomial. Feasible repairs are exactly covers with identical cost, so a feasible repair with $J(R)\le B$ exists if and only if a cover of weight at most B exists. Its decision version is in NP and is NP-hard already for unit weights by the classical set-cover reduction (Karp, 1972). Thus, the decision problem is NP-hard, and with membership from (i.a), it is NP-complete. The restricted setting in the claim, namely pure coverage with independent, non-conflicting edits, is precisely the constructed instance.

(ii) The reduction is cost-preserving: any polynomial-time algorithm returning a feasible repair within a factor $(1-\alpha)\ln N$ of optimal would return, on the constructed instances, a set cover within the same factor. The logarithmic inapproximability theorem for set cover rules this out,

for every fixed $0 < \alpha < 1$ and all sufficiently large N, unless P = NP (Dinur and Steurer, 2014; see also Feige, 1998). Since $N = |I_{s,t}|$ in the construction, the statement follows. □

Proof of Theorem 3

Proof. The argument is essentially Chvátal (1979)'s charging scheme for weighted set cover in the submodular-cover form of Wolsey (1982). The proof is split into three parts. Let A_k be the greedy selection after k steps, let $D_k = c_{\max} - c(A_k)$ be the remaining deficit, and let R^* be an optimal feasible repair of cost W^*. Fix a step at which $D_k > 0$.

(i) *Claim: some edit has cost-effectiveness at most W^*/D_k.* By monotonicity and feasibility of R^*, we have $c(A_k \cup R^*) = c_{\max}$. Order the elements of $R^* \setminus A_k$ arbitrarily as $f_1, \dots, f_r$. Telescoping and applying submodularity to each marginal give

$$\begin{aligned} D_k &= c(A_k \cup R^*) - c(A_k) \\ &= \sum_{j=1}^{r} \left[c(A_k \cup \{f_1, \dots, f_j\}) - c(A_k \cup \{f_1, \dots, f_{j-1}\}) \right] \\ &\le \sum_{j=1}^{r} \left[c(A_k \cup \{f_j\}) - c(A_k) \right]. \end{aligned}$$

Write $\Delta(f \mid A_k) = c(A_k \cup \{f\}) - c(A_k)$. Thus,

$$\sum_{f \in R^* \setminus A_k} \Delta(f \mid A_k) \ge D_k.$$

At least one of these marginals is positive. If every edit with positive marginal satisfied $w_f/\Delta(f \mid A_k) > W^*/D_k$, then

$$W^* \ge \sum_{f \in R^* \setminus A_k} w_f > \frac{W^*}{D_k} \sum_{f \in R^* \setminus A_k} \Delta(f \mid A_k) \ge W^*,$$

a contradiction. This proves the claim. Since greedy maximizes marginal coverage per unit cost, its chosen edit e_{k+1} satisfies

$$w_{k+1} \le \frac{W^*}{D_k} \Delta_{k+1}, \qquad \Delta_{k+1} = D_k - D_{k+1}.$$

(ii) *Termination and feasibility.* While $D_k > 0$, the claim exhibits an edit with positive marginal coverage. Greedy therefore always has a legal move, and the deficit, which is a non-negative integer, strictly decreases. Thus, greedy terminates feasibly with $c(A_k) = c_{\max}$.

(iii) *Cost.* Summing over the steps and using that the positive deficits are integers, we obtain

$$\begin{aligned} W^{\text{greedy}} &= \sum_k w_{k+1} \\ &\le W^* \sum_k \frac{D_k - D_{k+1}}{D_k} \\ &\le W^* \sum_k \sum_{j=D_{k+1}+1}^{D_k} \frac{1}{j} \\ &= W^* \sum_{j=1}^{c_{\max}} \frac{1}{j} = W^* H(c_{\max}) \le W^*(1 + \ln c_{\max}). \end{aligned}$$

The second inequality holds because each of the $D_k - D_{k+1}$ summands on its right is at least $1/D_k$. □

Proof of Proposition 5

Proof. Membership. A conflict-free feasible repair is a polynomial certificate, and both conflict-freeness and feasibility are polynomial-time checkable by Definition 9.

Hardness. Reduce from 3-SAT, which is NP-complete (Karp, 1972).

Given a formula φ with variables $v_1, ..., v_p$ and clauses $C_1, ..., C_q$, create one obligation per clause and create two unit-cost edits $e_i^\top$ and $e_i^\perp$ forming a conflict pair. $e_i^\top$ covers the obligations of clauses containing the literal v_i, and $e_i^\perp$ covers the obligations corresponding to clauses containing $\neg v_i$. If φ has a satisfying assignment, pick $e_i^\top$ or $e_i^\perp$ according to the assignment. The selection is conflict free, and every clause, being satisfied by some true literal, has its obligation covered, which constitutes a feasible repair.

Conversely, a conflict-free feasible repair selects at most one edit per pair. Extend it to an assignment by setting v_i true if $e_i^\top$ was selected, false if $e_i^\perp$ was selected, and arbitrarily otherwise. Every obligation is covered by some selected edit whose literal is true under this assignment, so every clause is satisfied. The conflict pairs are pairwise disjoint and costs are unit, as claimed. □

Proof of Proposition 6

Proof. Let the derived ground set be the family of closures and define

$$g(F) = c\left(\bigcup_{A \in F} A\right).$$

The function g is normalized and monotone. It is submodular because, for $F \subseteq F'$ and a closure $A \notin F'$, the set-marginal form of submodularity gives

$$c(X \cup A) - c(X) \geq c(Y \cup A) - c(Y),$$

where $X = \bigcup_{B \in F} B \subseteq Y = \bigcup_{B \in F'} B$. Theorem 3 therefore bounds the greedy accounted cost by $(1 + \ln c_{\max}) W^*_{\mathrm{cl}}$.

It remains to compare optima. Every finite down-closed repair is the union of the closures of its maximal elements. Distinct maximal elements are incomparable. Laminarity therefore makes their closures disjoint: a nesting would imply comparability. Hence an optimal precedence-constrained repair has a derived representation whose accounted closure cost equals its actual edit cost, and

$$W^*_{\mathrm{cl}} \leq W^*_{\prec}.$$

Conversely, a union of closures is down-closed and its actual cost is no greater than the sum of the selected full closure costs. Thus,

$$\mathrm{cost}(R_{\mathrm{greedy}}) \leq W^{\mathrm{accounted}}_{\mathrm{greedy}} \leq (1 + \ln c_{\max}) W^*_{\mathrm{cl}} \leq (1 + \ln c_{\max}) W^*_{\prec}.$$

□

Proof of Proposition 7

Proof. With finite action set $\mathcal{A}$, finite observation set, and finite horizon K, the set $\mathfrak{H}_K$ of observable histories of length at most K is finite. A history-dependent randomized policy is a point

$$\pi = (\pi(a \mid h))_{h \in \mathfrak{H}_K,\, a \in \mathcal{A}}$$

in a finite product of simplices, which is a compact polytope in finite dimension.

For fixed transition and observation kernels, integrate over any latent state trajectories. The probability of each finite action–observation history, and the expectation of any bounded loss attached to such a history, is a finite multilinear expression in the policy coordinates. Consequently, the expected discounted loss and $\mathbb{P}_\pi(M = 1)$ are continuous in π, so the chance-constraint set is closed.

For the CVaR constraint, the Rockafellar and Uryasev (2000) representation, using the tail-mass convention of Definition 10, is

$$\mathrm{CVaR}_\eta(L^\pi) = \min_{c \in [0, \bar{L}]} \left\{ c + \eta^{-1} \mathbb{E}_\pi \max(L^\pi - c, 0) \right\}.$$

The expression inside braces is jointly continuous in (π, c), and the minimum over the compact interval $[0, \bar{L}]$ is therefore continuous in π. Thus, the CVaR sublevel set is closed. No convexity of these functionals in π is claimed, and none is needed for attainment. The feasible set is a non-empty closed subset of a compact set and is therefore compact. A continuous function on a non-empty compact set attains its minimum. Randomization is permitted by construction of the policy space. □

Proof of Proposition 8

Proof. Let Π denote the ambient policy class and let

$$\Pi_0 = \{\pi \in \Pi : \mathbb{P}_\pi(\text{gate violation}) = 0\} \neq \emptyset.$$

For $\pi \in \Pi_0$, we have $V(\pi) = \mathbb{E}_\pi[u - c] > -\infty$. For $\pi \notin \Pi_0$, we have $V(\pi) = -\infty$. Hence,

$$\sup_{\pi \in \Pi} V(\pi) = \sup_{\pi \in \Pi_0} \mathbb{E}_\pi[u - c],$$

and, under the stated attainment hypothesis, the maximizers of V over Π are exactly the maximizers of $\mathbb{E}_\pi[u - c]$ over Π_0, which is the first claim.

For the second claim, consider the finite penalty objective

$$J_\kappa(\pi) = \mathbb{E}_\pi[u - c] - \kappa \, \mathbb{P}_\pi(\text{gate violation}),$$

where $0 \leq \kappa < \infty$. Construct an instance with two available policies: π_g, which respects the gate and has $\mathbb{E}_{\pi_g}[u - c] = 0$, and π_v, which has $\mathbb{E}_{\pi_v}[u - c] = 1$ and violates the gate with probability

$$p = \begin{cases} 1/2, & \kappa = 0, \\ \min\{1/2, 1/(2\kappa)\}, & \kappa > 0. \end{cases}$$

Then $J_\kappa(\pi_g) = 0$, whereas $J_\kappa(\pi_v) = 1 - \kappa p \geq 1/2 > 0$. The penalized optimum therefore violates the gate with positive probability. Since κ was arbitrary, no single finite penalty reproduces the constrained solution uniformly over all instances. □

Proof of Proposition 9

Proof. For a fixed reviewer h and pair (i, j), let

$$p^+_{h,ij} = \mathbb{P}(y_i \succ y_j \mid h, s_{ij} = +1),$$
$$p^-_{h,ij} = \mathbb{P}(y_i \succ y_j \mid h, s_{ij} = -1).$$

By the model in Proposition 9,

$$\begin{aligned}\operatorname{logit}(p^{+}_{h,ij}) &= \beta_h(u(y_i) - u(y_j)) + b_h,\\ \operatorname{logit}(p^{-}_{h,ij}) &= \beta_h(u(y_i) - u(y_j)) - b_h.\end{aligned}$$

Therefore,

$$\begin{aligned}\beta_h(u(y_i) - u(y_j)) &= \frac{1}{2}\left(\operatorname{logit}(p^{+}_{h,ij}) + \operatorname{logit}(p^{-}_{h,ij})\right),\\ b_h &= \frac{1}{2}\left(\operatorname{logit}(p^{+}_{h,ij}) - \operatorname{logit}(p^{-}_{h,ij})\right).\end{aligned}$$

Thus, every b_h and every observed scaled utility difference are identified. For the reference reviewer, $\beta_{h_0} = 1$ identifies the utility difference on every pair reviewed by h_0. If reviewer h shares a non-tied pair with h_0, division of the two identified scaled differences identifies β_h. Consequently, every observed edge of the union comparison graph identifies its unscaled utility difference. Connectivity identifies all observed-repair utilities up to one additive constant, and $u(y_{i_0}) = 0$ fixes that constant.

To see why order variation matters, consider one reviewer and one pair shown in only one deterministic order. The observable log-odds is $\beta_h(u(y_i) - u(y_j)) + b_h$. Even with β_h fixed, replacing $u(y_i) - u(y_j)$ by $u(y_i) - u(y_j) + d$ and b_h by $b_h - \beta_h d$ leaves the observable probability unchanged. Joint identification can therefore fail without order variation. The claim concerns the finite vector of observed-repair utilities. Identification of a parameter θ requires the additional injectivity condition stated in the proposition. □

Proof of Proposition 10

Proof. The argument is the standard pointwise Bayes-optimality computation of the learning to defer literature (Mozannar and Sontag, 2020; Mao et al., 2025). Condition on x. A randomized rule selects action a with probabilities $\lambda_a(x)$ summing to one, incurring $\sum_a \lambda_a(x)\rho_a(x) \geq \min_a \rho_a(x)$, with equality iff the rule is supported on minimizers. Since the one-step action affects no future state by hypothesis, the total expected loss is $\mathbb{E}_x[\sum_a \lambda_a \rho_a(x)]$, minimized pointwise by $a^*(x)$.

The characterization of optimal deferral is the definition of the minimizer restricted to $\mathcal{H}$. When review generates information that affects later actions, total loss need not decompose pointwise in the present action. Consequently, the one-step rule need not remain optimal. □

Proof of Proposition 11

Proof. Fix x and abbreviate $\hat{a} = \hat{a}(x)$ and $a^* = a^*(x)$. Then,

$$\begin{aligned}\rho_{\hat{a}}(x) - \rho_{a^*}(x) &= [\rho_{\hat{a}}(x) - \hat{\rho}_{\hat{a}}(x)] + [\hat{\rho}_{\hat{a}}(x) - \hat{\rho}_{a^*}(x)] + [\hat{\rho}_{a^*}(x) - \rho_{a^*}(x)]\\ &\leq \varepsilon_{\hat{a}}(x) + \varepsilon_{a^*}(x),\end{aligned}$$

because $\hat{\rho}_{\hat{a}}(x) \leq \hat{\rho}_{a^*}(x)$. The second bound follows by taking the maximum over actions. Since the inequality holds pointwise, taking expectation over x proves the final claim. □

Proof of Lemma 3

Proof. For one logged observation, condition on $X = x$ and use the tower property:

$$\mathbb{E}_b\left[\frac{\pi(A \mid X)}{b(A \mid X)} R \,\middle|\, X = x\right] = \sum_{a \in \mathcal{A}_{\text{safe}}(x)} b(a \mid x) \frac{\pi(a \mid x)}{b(a \mid x)} \mathbb{E}[R \mid X = x, A = a]$$
$$= \sum_{a \in \mathcal{A}_{\text{safe}}(x)} \pi(a \mid x) \mu(x, a).$$

Positivity makes every ratio on the support of π well defined. Taking expectation over X gives $V(\pi)$. Averaging n independent copies preserves the expectation, which proves

$$\mathbb{E}_b[\widehat{V}_{\text{IPS}}(\pi)] = V(\pi).$$

□

Proof of Lemma 4

Proof. Condition on the fitted nuisance functions and on $X = x$. The conditional expectation of one summand in (23) is

$$\sum_a \pi(a \mid x) \widehat{\mu}(x, a) + \sum_a b(a \mid x) \frac{\pi(a \mid x)}{\widehat{b}(a \mid x)} \big(\mu(x, a) - \widehat{\mu}(x, a)\big).$$

If $\widehat{b} = b$ on the support of π, this expression becomes

$$\sum_a \pi(a \mid x) \widehat{\mu}(x, a) + \sum_a \pi(a \mid x) \big(\mu(x, a) - \widehat{\mu}(x, a)\big) = \sum_a \pi(a \mid x) \mu(x, a).$$

If instead $\widehat{\mu} = \mu$, the residual term has conditional expectation zero and the direct term again equals $\sum_a \pi(a \mid x) \mu(x, a)$, regardless of $\widehat{b}$. Integrating over x and averaging over observations gives

$$\mathbb{E}[\widehat{V}_{\text{DR}}(\pi)] = V(\pi).$$

Under cross-fitting, this calculation is applied conditionally within each held-out fold and then averaged over folds. □

Proof of Lemma 5

Proof. By Lemma 3, $V(\pi) = \mathbb{E}_b[WR]$. Since $\min\{W, M\} - W = -(W - M)_+$,

$$\mathbb{E}_b[\widehat{V}_M(\pi)] - V(\pi) = -\mathbb{E}_b[(W - M)_+ R]\,.$$

Therefore,

$$\left|\mathbb{E}_b[\widehat{V}_M(\pi)] - V(\pi)\right| \le R_{\max} \mathbb{E}_b[(W - M)_+],$$

which proves the bias bound. Moreover, $(W - M)_+ \to 0$ pointwise as $M \to \infty$ and

$$0 \le (W - M)_+ \le W, \qquad \mathbb{E}_b[W] = \mathbb{E}_X\left[\sum_a \pi(a \mid X)\right] = 1.$$

Dominated convergence therefore gives $\mathbb{E}_b[(W - M)_+] \to 0$. □

Proof of Theorem 4

Proof. If $T \le K$, each route is selected at most once and the asserted bound follows from the initialization rule and the positive constant term in (27). Suppose henceforth that $T > K$.

Transform losses into rewards by setting

$$Y_{a,s} = 1 - L_{a,s}, \qquad \theta_a = 1 - \mu_a.$$

Then an optimal route has mean reward $\theta^* = \max_a \theta_a = 1 - \mu^*$, and $\theta^* - \theta_a = \Delta_a$. The loss rule in (26) is equivalent to selecting the largest reward index

$$\widehat{\theta}_{a,s} + c_{t,s}, \qquad c_{t,s} = \sqrt{\frac{2 \log t}{s}},$$

where $\widehat{\theta}_{a,s}$ is the empirical mean of the first s rewards from route a.

Fix a suboptimal route a, choose one optimal route a^*, abbreviate $\Delta = \Delta_a > 0$, and put

$$\ell_a = \left\lceil \frac{8 \log T}{\Delta^2} \right\rceil.$$

The number of pulls of a satisfies

$$N_a(T) \le \ell_a + \sum_{t=K+1}^{T} \mathbf{1}\{A_t = a,\ N_a(t-1) \ge \ell_a\}.$$

On an event in the sum, let $s = N_a(t-1)$ and $s^* = N_{a^*}(t-1)$. Selection of a implies

$$\widehat{\theta}_{a,s} + c_{t,s} \ge \widehat{\theta}_{a^*,s^*} + c_{t,s^*}.$$

Consequently, at least one of the following events occurs:

$$E_{1,t,s^*} = \left\{ \widehat{\theta}_{a^*,s^*} \le \theta^* - c_{t,s^*} \right\},$$
$$E_{2,t,s} = \left\{ \widehat{\theta}_{a,s} \ge \theta_a + c_{t,s} \right\},$$
$$E_{3,t,s} = \left\{ \theta^* < \theta_a + 2c_{t,s} \right\}.$$

Indeed, if all three events failed, the index of a^* would be larger than θ^*, whereas the index of a would be smaller than $\theta_a + 2c_{t,s} \le \theta^*$. This contradicts the selection of a.

For $s \ge \ell_a$ and $t \le T$,

$$2c_{t,s} \le 2\sqrt{\frac{2 \log T}{\ell_a}} \le \Delta,$$

so $E_{3,t,s}$ is impossible. Hoeffding's inequality gives, for every fixed positive s and s^*,

$$\mathbb{P}(E_{1,t,s^*}) \le t^{-4}, \qquad \mathbb{P}(E_{2,t,s}) \le t^{-4}.$$

A union bound over the at most t^2 feasible pairs (s, s^*) therefore gives

$$\mathbb{P}(A_t = a,\ N_a(t-1) \ge \ell_a) \le 2t^{-2}.$$

It follows that

$$\begin{aligned} \mathbb{E}[N_a(T)] &\le \ell_a + \sum_{t=1}^{\infty} 2t^{-2} \\ &\le \frac{8 \log T}{\Delta_a^2} + 1 + \frac{\pi^2}{3}. \end{aligned}$$

Finally,

$$\mathbb{E}[\mathcal{R}_T] = \sum_{a:\,\Delta_a>0} \Delta_a\, \mathbb{E}[N_a(T)].$$

Substitution of the pull count bound proves (27). Since the rule optimizes only over $\mathcal{A}_{\text{safe}}$, it never selects a route outside the hard gate. □

Proof of Corollary 1

Proof. Partition the dynamic pseudo regret over the Φ stationary phases. In the first phase, and in each later phase after the alarm and restart, the remaining observations before the next change form a stationary safe route bandit instance. Theorem 4 applies to this restarted portion, with the trivial one round bound covering a portion of length less than two. Its length is at most T_φ, so its expected pseudo regret is at most $B_\varphi(T_\varphi)$.

There are $\Phi - 1$ phase changes. Between a phase change and its alarm, at most Δ_{det} episodes elapse. Because losses lie in $[0, 1]$, the instantaneous pseudo regret is at most one, and the total detection delay contribution is therefore at most $(\Phi-1)\Delta_{\text{det}}$. Summing the stationary phase and delay contributions proves the first inequality in (28). The second follows from $\Phi - 1 \le \Phi$. □

Proof of Lemma 6

Proof. Let P_b and P_π denote the trajectory laws induced by b and π. Because the two policies operate under the same initial state distribution and environment dynamics, their trajectory factorizations differ only in the action probabilities. Positivity therefore gives the Radon Nikodym derivative

$$\frac{\mathrm{d}P_\pi}{\mathrm{d}P_b}(\xi) = \prod_{k=1}^{H} \frac{\pi(a_k \mid h_k)}{b(a_k \mid h_k)} = w(\xi)$$

on every trajectory to which P_π assigns positive probability. Hence, by change of measure,

$$\mathbb{E}_b[w(\xi)G(\xi)] = \int G(\xi)w(\xi)\,\mathrm{d}P_b(\xi) = \int G(\xi)\,\mathrm{d}P_\pi(\xi) = \mathbb{E}_\pi[G(\xi)]\,.$$

If each likelihood ratio is at most C_∞, their product is at most C_∞^H. □

Proof of Theorem 5

Proof. Define the simultaneous good event

$$\mathcal{G} := \bigcap_{\pi\in\Pi} \left\{ \left| \widehat{V}(\pi) - V(\pi) \right| \le w_n(\pi) \right\}.$$

The union bound gives $\mathbb{P}(\mathcal{G}) \ge 1 - \delta$. On $\mathcal{G}$, for every $\pi' \in \Pi$,

$$\begin{aligned} V(\widehat{\pi}) &\ge \widehat{V}(\widehat{\pi}) - w_n(\widehat{\pi}) \\ &\ge \widehat{V}(\pi') - w_n(\pi') \\ &\ge V(\pi') - 2w_n(\pi'). \end{aligned}$$

The first and third inequalities hold on $\mathcal{G}$, and the second follows from the definition of $\widehat{\pi}$. Maximizing the right hand side over $\pi' \in \Pi$ proves (29).

For the stated trajectory importance sampling example, each summand lies in $[0, C_\infty^H]$ by Lemma 6, the summands are i.i.d. across logged episodes, and their mean is $V(\pi)$. The two-sided Hoeffding bound at level $\delta/|\Pi|$ therefore yields

$$w_n(\pi) = C_\infty^H \sqrt{\frac{\log(2|\Pi|/\delta)}{2n}}.$$

□

Proof of Lemma 7

Proof. For fixed N, write

$$F_q(x) := \mathbb{P}_q(\mathrm{Bin}(N, q) \le x)\,.$$

For $0 \le x < N$, the function $q \mapsto F_q(x)$ is continuous and strictly decreasing on $(0, 1)$. Consequently,

$$U_\alpha(X; N) < p \quad \Longleftrightarrow \quad F_p(X) < \alpha.$$

For $X = N$, both sides are false because $U_\alpha(N; N) = 1$ and $F_p(N) = 1$.

Let

$$x_\alpha := \min\{x \in \{0, \dots, N\} : F_p(x) \ge \alpha\}.$$

Then

$$\{F_p(X) < \alpha\} = \{X < x_\alpha\} = \{X \le x_\alpha - 1\}.$$

If $x_\alpha = 0$, this event is empty. Otherwise its probability is $F_p(x_\alpha - 1) < \alpha$ by the minimality of x_α. Thus,

$$\mathbb{P}_p(U_\alpha(X; N) < p) \le \alpha,$$

which is equivalent to the claimed coverage. The case $N = 0$ follows from the convention $U_\alpha(0; 0) = 1$. □

Proof of Proposition 12

Proof. Fix $\tau \in \Theta$ and condition on the full acceptance pattern $\boldsymbol{A}_\tau = (A_{\tau,1}, \dots, A_{\tau,K})$. Because the episode pairs are i.i.d. and the fitted model and gate are fixed, the miss indicators at the accepted indices are conditionally independent and each has distribution

$$M_{\tau,i} \mid \{A_{\tau,i} = 1\} \sim \mathrm{Bernoulli}(R(\tau)).$$

It follows that

$$X_\tau \mid \boldsymbol{A}_\tau \sim \mathrm{Bin}(N_\tau, R(\tau)).$$

The conditional distribution depends on the pattern only through N_τ, so the same binomial law holds conditionally on N_τ alone.

For fixed τ, define the failure event

$$\mathcal{B}_\tau := \left\{U_{\delta/|\Theta|}(X_\tau; N_\tau) < R(\tau)\right\}.$$

Lemma 7, applied conditionally on N_τ, gives

$$\mathbb{P}(\mathcal{B}_\tau \mid N_\tau) \le \frac{\delta}{|\Theta|}$$

for every value of N_τ. The tower property therefore yields $\mathbb{P}(\mathcal{B}_\tau) \le \delta/|\Theta|$. A union bound now gives

$$\mathbb{P}\left(\bigcap_{\tau\in\Theta} \mathcal{B}_\tau^{\mathsf{c}}\right) \ge 1-\delta.$$

On this event the upper limit covers $R(\tau)$ simultaneously for every $\tau \in \Theta$, and hence also at the data-dependent choice $\widehat{\tau}$. This proves (31).

Finally, for each fixed τ,

$$J(\tau) = \mathbb{P}(M_\tau = 1, A_\tau = 1) = \mathbb{P}(A_\tau = 1)R(\tau) \le R(\tau).$$

Conditionally on the calibration sample, the selected threshold $\widehat{\tau}$ is fixed and the fresh deployment episode is independent. Thus, on the simultaneous coverage event and whenever deployment is permitted,

$$J(\widehat{\tau}) \le R(\widehat{\tau}) \le \overline{R}(\widehat{\tau}) \le \varepsilon.$$

□

Proof of Theorem 6

Proof. Right continuity, compactness of Λ, and $\ell_i(\lambda_{\max}) = 0$ ensure that the minimum in (32) exists. Define the full sample empirical loss

$$\widetilde{L}(\lambda) := \frac{1}{K+1}\sum_{i=1}^{K+1} \ell_i(\lambda)$$

and the symmetric oracle threshold

$$\lambda^\star := \min\left\{\lambda \in \Lambda : \widetilde{L}(\lambda) \le \varepsilon\right\}.$$

The same assumptions ensure that $\lambda^\star$ exists. Since $\ell_{K+1}(\lambda) \le 1$ for every λ,

$$\frac{\sum_{i=1}^{K} \ell_i(\lambda) + 1}{K+1} \ge \widetilde{L}(\lambda).$$

Hence every parameter feasible for the criterion defining $\widehat{\lambda}$ is feasible for the oracle criterion, so $\widehat{\lambda} \ge \lambda^\star$. Because ℓ_{K+1} is nonincreasing,

$$\ell_{K+1}(\widehat{\lambda}) \le \ell_{K+1}(\lambda^\star). \tag{56}$$

The map from the $K+1$ loss functions to $\lambda^\star$ is symmetric. Exchangeability therefore implies that the random variables $\ell_i(\lambda^\star)$ have the same expectation. Consequently,

$$\begin{aligned}\mathbb{E}[\ell_{K+1}(\lambda^\star)] &= \frac{1}{K+1}\sum_{i=1}^{K+1} \mathbb{E}[\ell_i(\lambda^\star)] \\ &= \mathbb{E}\left[\widetilde{L}(\lambda^\star)\right] \le \varepsilon.\end{aligned}$$

Taking expectations in (56) proves (33). □

Proof of Proposition 13

Proof. Let E denote the event of any material miss. By hypothesis,

$$E \subseteq E_{\mathrm{M}} \cup \bigcup_{j=1}^{\infty} \big(\{j \le N_{\mathrm{H}}\} \cap H_j\big).$$

Boole's inequality gives

$$\mathbb{P}(E) \le \mathbb{P}(E_{\mathrm{M}}) + \mathbb{P}\left(\bigcup_{j=1}^{\infty} \big(\{j \le N_{\mathrm{H}}\} \cap H_j\big)\right). \tag{57}$$

For the human review term,

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{j=1}^{\infty} \big(\{j \le N_{\mathrm{H}}\} \cap H_j\big)\right) &\le \mathbb{E}\left[\sum_{j=1}^{\infty} \mathbf{1}\{j \le N_{\mathrm{H}}\} \mathbf{1}_{H_j}\right] \\
&= \sum_{j=1}^{\infty} \mathbb{E}\big[\mathbf{1}\{j \le N_{\mathrm{H}}\} \mathbb{E}\big[\mathbf{1}_{H_j} \mid \mathcal{F}_{j-1}\big]\big] \\
&\le \sum_{j=1}^{\infty} \mathbb{E}[\mathbf{1}\{j \le N_{\mathrm{H}}\} p_j] \\
&= \mathbb{E}\left[\sum_{j=1}^{N_{\mathrm{H}}} p_j\right].
\end{aligned}$$

The second line uses $\{j \le N_{\mathrm{H}}\} \in \mathcal{F}_{j-1}$ and the tower property. The interchange of the non-negative sum and expectation follows from Tonelli's theorem. Substituting this bound and $\mathbb{P}(E_{\mathrm{M}}) \le \varepsilon_{\mathrm{M}}$ into (57) proves (34). The deterministic form (35) follows by taking $N_{\mathrm{H}} = m'$ and $p_j = p_{h_j}$. Replacing E_{M} by a union of individually bounded machine miss events gives the per repair variant. □

Proof of Proposition 14

Proof. Let

$$Z := \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}.$$

Since $B \subseteq A$ and $Z \le \Lambda$ holds $\mathbb{P}$ almost surely on A,

$$\mathbb{Q}(B) = \int \mathbf{1}_B Z \,\mathrm{d}\mathbb{P} \le \Lambda \mathbb{P}(B) \le \Lambda \varepsilon_{\mathrm{M}},$$

which proves (36). Applying Proposition 13 under $\mathbb{Q}$, using the human review bounds assumed valid under that law, gives (37).

For the conditional risk claim, $B = \{M = 1\} \cap A$ and therefore

$$\begin{aligned}
\mathbb{Q}(M = 1 \mid A) &= \frac{\mathbb{Q}(B)}{\mathbb{Q}(A)} \\
&\le \Lambda \frac{\mathbb{P}(B)}{\mathbb{Q}(A)} \\
&= \Lambda \mathbb{P}(M = 1 \mid A) \frac{\mathbb{P}(A)}{\mathbb{Q}(A)} \\
&\le \Lambda \rho \frac{\mathbb{P}(A)}{\mathbb{Q}(A)},
\end{aligned}$$

which proves (38). If the density ratio bound is imposed directly on the conditional laws, the same change of measure calculation within A gives

$$\mathbb{Q}(M = 1 \mid A) \leq \Lambda \mathbb{P}(M = 1 \mid A) \leq \Lambda \rho.$$

Finally, when a valid monitor certifies $Z \leq \overline{\Lambda} \leq \Lambda_{\max}$ on A, the first calculation applies with $\Lambda_{\max}$. When certification fails and machine acceptance is suspended, A and hence B are empty, so the machine term is zero. □

Proof of Proposition 15

Proof. Put $q_i = 1 - p_i$ and $s_{12} = \sqrt{p_1 q_1 p_2 q_2}$. Since $0 < p_i < 1$, both variances are positive and the correlation is well defined. By the definition of correlation,

$$\mathrm{Cov}(M_1, M_2) = \varrho s_{12}.$$

On the other hand,

$$\mathrm{Cov}(M_1, M_2) = \mathbb{E}[M_1 M_2] - p_1 p_2 = \mathbb{P}(M_1 = 1, M_2 = 1) - p_1 p_2,$$

because $M_1 M_2$ is the indicator that both experts miss. Rearranging gives (40).

The Fréchet bounds for two events give

$$\max\{0, p_1 + p_2 - 1\} \leq \mathbb{P}(M_1 = 1, M_2 = 1) \leq \min\{p_1, p_2\}.$$

Substituting (40), subtracting $p_1 p_2$, and dividing by $s_{12} > 0$ yields (42).

With expert 1 alone, the materiality weighted error term equals $w(x)p_1$. After adding expert 2, the corresponding term equals

$$w(x)\mathbb{P}(M_1 = 1, M_2 = 1) + \widetilde{c}_2.$$

The second review strictly reduces the objective exactly when

$$w(x)\left[p_1 - \mathbb{P}(M_1 = 1, M_2 = 1)\right] > \widetilde{c}_2.$$

Using (40) gives (41).

Finally, $\varrho = 1$ is the equality case of the Cauchy Schwarz inequality for the centered miss indicators. Since the indicators are nondegenerate Bernoulli variables, this equality implies $M_1 = M_2$ almost surely and hence $p_1 = p_2$. If $p_1 = p_2 = p$, then

$$\mathbb{P}(M_1 = 1, M_2 = 1) = p^2 + p(1 - p) = p.$$

The risk reduction from the second opinion is therefore zero. □

Proof of Theorem 7

Proof. Define the coverage failure events

$$B_R = \{D_R > U_R\}, \qquad B_C = \{D_C < L_C\}.$$

By (44), $\mathbb{P}(B_R) \leq \delta_R$ and $\mathbb{P}(B_C) \leq \delta_C$.

Suppose that π_1 is adopted and $D_R > \Delta_{\text{safe}}$. Adoption implies $U_R \leq \Delta_{\text{safe}}$. Hence,

$$U_R \leq \Delta_{\text{safe}} < D_R,$$

which means that B_R occurred. The unsafe adoption event is therefore contained in B_R, which proves (45).

Next, suppose that π_1 is adopted and $D_C \leq 0$. Adoption implies $L_C > 0$. Hence,

$$D_C \leq 0 < L_C,$$

which means that B_C occurred. The cost ineffective adoption event is therefore contained in B_C, which proves (46).

The event in (47) is contained in $B_R \cup B_C$. Boole's inequality gives

$$\mathbb{P}(B_R \cup B_C) \leq \mathbb{P}(B_R) + \mathbb{P}(B_C) \leq \delta_R + \delta_C.$$

No independence between the two confidence bounds is required. Setting $\delta_R = \delta_C = \delta/2$ completes the proof. □

Proof of Theorem 8

Proof. For $j \geq 1$, put

$$C_j := \{j \leq N_{\mathrm{H}}\} \cap H_j, \qquad C := \bigcup_{j \geq 1} C_j.$$

The explicit failure-accounting assumption implies

$$M \subseteq B \cup E_{\mathrm{M}} \cup C.$$

Hence, Boole's inequality gives

$$\mathbb{Q}(M) \leq \mathbb{Q}(B) + \mathbb{Q}(E_{\mathrm{M}}) + \mathbb{Q}(C). \tag{58}$$

Because $E_{\mathrm{M}} \subseteq A$, the density-ratio bound and the machine certificate yield

$$\begin{aligned} \mathbb{Q}(E_{\mathrm{M}}) &= \int_{E_{\mathrm{M}}} \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} \, \mathrm{d}\mathbb{P} \\ &\leq \Lambda \mathbb{P}(E_{\mathrm{M}}) \\ &\leq \Lambda \varepsilon_{\mathrm{M}}. \end{aligned}$$

For the human channel, countable subadditivity, conditional expectation, and the predictability of $\{j \leq N_{\mathrm{H}}\}$ give

$$\begin{aligned} \mathbb{Q}(C) &\leq \sum_{j \geq 1} \mathbb{Q}(C_j) \\ &= \sum_{j \geq 1} \mathbb{E}_{\mathbb{Q}}\big[\mathbf{1}_{\{j \leq N_{\mathrm{H}}\}} \mathbb{Q}(H_j \mid \mathcal{F}_{j-1})\big] \\ &\leq \sum_{j \geq 1} \mathbb{E}_{\mathbb{Q}}\big[\mathbf{1}_{\{j \leq N_{\mathrm{H}}\}} p_j\big] \\ &= \mathbb{E}_{\mathbb{Q}}\left[\sum_{j=1}^{N_{\mathrm{H}}} p_j\right]. \end{aligned}$$

The final interchange follows from Tonelli's theorem because all summands are nonnegative. Combining these bounds with $\mathbb{Q}(B) \leq \rho_{\mathrm{graph}}$ in (58), and using the trivial bound $\mathbb{Q}(M) \leq 1$, proves (48). The deterministic form (49) follows by setting $N_{\mathrm{H}} = m'$ and $p_j = \bar{p}_{h_j}$. No independence assumption is used. □

Proof of Proposition 16

Proof. By assumption, the variables $D_1, \dots, D_E$ are independent and identically distributed and take values in $[-B, B]$, an interval of length $2B$. Hoeffding's inequality therefore gives, for every $s > 0$,

$$\begin{aligned}\mathbb{P}\big(\overline{D}_E \le \mathbb{E}[D_e] - s\big) &\le \exp\left(-\frac{2Es^2}{(2B)^2}\right) \\ &= \exp\left(-\frac{Es^2}{2B^2}\right).\end{aligned}$$

Setting the right-hand side equal to δ_H gives

$$s = B\sqrt{\frac{2\ln(1/\delta_H)}{E}}.$$

Since $\mathbb{E}[D_e] = \Delta_H(\pi_0, \pi_1)$, it follows with probability at least $1 - \delta_H$ that

$$\Delta_H(\pi_0, \pi_1) \ge \overline{D}_E - B\sqrt{\frac{2\ln(1/\delta_H)}{E}},$$

which is (54).

Let C_H be the event on which this workload bound holds, and let C_R be the event on which

$$R(\pi_1) \le \widehat{U}^{\mathrm{e2e}}_{\delta_R}(\pi_1).$$

The assumed coverage of the two certificates gives $\mathbb{P}(C_H^c) \le \delta_H$ and $\mathbb{P}(C_R^c) \le \delta_R$. Boole's inequality yields

$$\mathbb{P}(C_H \cap C_R) \ge 1 - \delta_H - \delta_R.$$

No independence between C_H and C_R is required. On their intersection, a positive lower confidence bound in (54) proves $\Delta_H(\pi_0, \pi_1) > 0$, while (51) and the coverage event C_R imply $R(\pi_1) \le \varepsilon$. This proves the joint claim. □