A FORMAL GUARANTEE AT THE INDIVIDUAL DECISION

Anonymous authors
Paper under double-blind review

ABSTRACT

An accuracy rate says nothing about whether the decision in front of you is right. We show a construction that does by routing a probabilistic system's decisions through a knowledge property graph (KPG), a graph that holds an authority's own claims, each anchored to its source and marked with who stands behind it and whether it still binds. The KPG decides by descent and eliminates provisions where the authority licenses it. It takes a verdict on each surviving element in three values, and commits only where every way of settling the open elements reaches the same outcome. The guarantee is a theorem. If the KPG represents the authority and the settled verdicts are correct, every committed decision is the authority's decision. Every decision carries a witness, a certificate a checker verifies without the model or the system that produced it, and if a premise fails, the witness names the claim or the verdict, or bounds the claim the KPG lacks. A decision is therefore as right as the KPG and the authority it answers to, and every shortfall is located, so an organization corrects its KPG and every later decision inherits the correction. The construction is exercised on five authorities. On 700 variant classifications under the ACMG/AMP standard as one expert panel's four specifications refine it, with the KPG's conditions enforced on every recorded criterion, no commitment deviates from the authority's decision. Two disagree with the curator's label, and each is located in the label. On 537 curations from a second panel, with the same code, the same holds, and with the curator's criteria taken as settled, the one commitment that disagrees with a label is located in the label. Authority enters the record by two witnessed paths. Attestation adds or amends a claim to the KPG itself. Adjudication supplies a fact to one abstention and can change only that one. Without the KPG's certification, injected intake faults produce 615 silent wrong answers. With it, every fault is caught before it reaches a decision, and named so it can be corrected.

1 INTRODUCTION

Suppose a system benchmarks at 93% and has just produced an answer. What follows about that answer? At best, that there's a 7 in 100 chance the answer is wrong. The accuracy describes a distribution the benchmark sampled. The answer is one draw from the population. A probabilistic system cannot guarantee whether any single decision is right.

The tools built for this problem stop short of producing a guarantee. Selective prediction lets a system decline, but it declines on a learned confidence, a population statistic used as a proxy for an individual one (Geifman & El-Yaniv, 2017; Kamath et al., 2020). Conformal sets contain the truth at a stated rate over a sequence, not for the set in front of you. Neural-network verification proves a property of the network, not of the claim it made (Katz et al., 2017; Wang et al., 2021). Sending the problem to a logic solver returns a proof, but a network must first write the problem in logic, and when the translation fails the system answers anyway (Pan et al., 2023; Olausson et al., 2023). Each needs someone to write down what a correct answer is, and for most tasks nobody can (Seshia et al., 2022). Where an authority governs, what makes an answer correct is already recorded. Our work structures that record into a form a machine can reason over and check.

Prior work put authority into an ontology (Anonymous, 2026), a structure of ranked rules, scored evidence, and a stated order of precedence against which a model's verdicts were read. Where the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

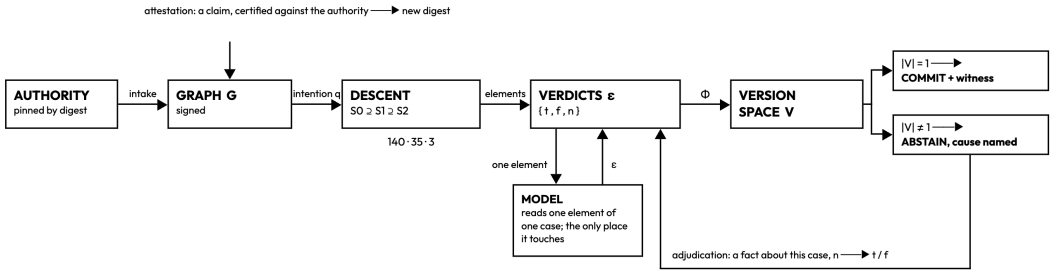


Figure 1: **The architecture.** An authority pinned by digest is read once into a signed KPG. A decision descends through the KPG to the provisions the intention reaches, asks a model one element at a time, and commits only where the version space of outcomes is a singleton. Otherwise it abstains and names the cause. Authority enters by two paths. Attestation adds a claim to the KPG, and every later decision inherits it. Adjudication supplies a fact to one abstained case, and only that decision changes. The model is used at one step, answering elements, and cannot commit a decision.

authority was in the structure the verdicts followed it and were right. Where it was absent, the model's own confidence carried nothing either way. So a model follows authority it is given, and following is not enforcing. Making a verdict answer to the authority, with no path around it, is a different act, and it is the subject here.

Our position is that a decision that carries the derivation establishing it is a different object from a decision that carries a probability. The derivation can be checked. The probability can only be trusted. Where the derivation exists, correctness is a property of that decision, established by replaying it. Where it does not, the system withholds the decision and names what is missing.

The object that produces such decisions is a **knowledge property graph** (KPG; Figure 1). It is a graph that holds an authority. Its nodes are the authority's provisions, carried verbatim. Its elements are the questions of fact each provision turns on. Its edges are the precedence the authority states between its provisions. Because the KPG holds the authority and nothing else, the ground truth for a decision is isolated and known before any case is read.

A **descent** is a chain of provision sets that only narrows, every elimination citing the claim in the KPG that licenses it, and it takes a decision from the whole authority to the governing provisions.

A **verdict** is the answer to one element, in one of three values. True, false, or undetermined. A model reads one element of one case and answers one reading. Readings that disagree leave the element undetermined, and every reading is kept in the record.

A **commitment** is a decision the system issues. It issues one only where the outcomes reachable under every way the open elements could be settled collapse to exactly one. A **witness** is the record a commitment emits, listing every premise it used, and it replays without the model.

An **abstention** is a decision the system withholds, naming the elements that blocked it.

Attestation adds a claim to the KPG. **Adjudication** supplies a fact to one abstention. These are the only two paths by which authority enters the record.

A model is used in two places. At intake it turns the authority's text into the claims of the KPG, and certification compares every claim it produced against the source text. At decision time it answers one question of fact about one case, and several answers are taken so that disagreement leaves the element open. It is not given the question being decided, a citation, or the outcome. Which provisions govern and what the decision is are computed from the KPG without a model, and replay repeats that computation without one.

Where a KPG is attested against its source, a *formal guarantee* attaches to the individual decisions of a probabilistic system that is otherwise unchanged (i.e., nothing is retrained, the system is as it was). The experiments count decisions on which the guarantee held, one at a time. Whether a decision is right is settled for that decision alone, never over a population.

Contributions. (1) The KPG and the descent, a chain in the lattice of provision sets whose replay is a zeta function, with verdicts in the three-valued Kleene domain and a commitment a singleton version space (§2). (2) The witness as a certificate, with soundness and localization theorems over the premises it names and a checker that verifies it without the producer (§2.6), and the proposition that separates the two paths by which authority enters the record (§2.5). (3) Experiments on five authorities, reported as right, wrong, and abstained at every stage. Every commitment is right on its record, every disagreement with a label is located, and 129 abstentions are discharged by the one fact each named, none wrongly (§5). (4) A test of the guarantee under 69 faults injected at intake and 87 composed plans, in which no wrong answer passes silently and every one becomes an abstention that names the fault (§5.3).

2 THE CONSTRUCTION

2.1 THE KPG, AND WHAT ATTESTS IT

An authority's text is decomposed into typed claims. A *provision* carries its text verbatim, a role, a combination rule over its elements, and the elements themselves. The role says whether the provision defines a term, states a condition, or excludes a case. An *element* is one question of fact, answerable from a case without deciding the outcome, and it travels with the phrase of the authority it decomposes and the place that phrase was taken from. A *precedence* claim says which provision yields to which, in the authority's own words.

A model performs the intake, reading the authority alone and never a case, against a schema that knows nothing about the corpus. Three checks then run on what it produced, and none is a matter of judgment. Every element must string-match the source, so an element whose text is not in the authority is refused at construction. Every provision of the source must contribute a claim, so a document that went silent fails certification, and text that states no provision, a preamble or a note, passes through unclaimed. And a structural audit names every provision whose rule, role, or scope is malformed before any case is decided. The KPG is signed by a hash over its provisions, so a decision cites the object that produced it.

A person may add a claim to the KPG. It passes the same three checks, is recorded with the name of who attested it, and changes the digest. A claim that no account states some fact is checked over every account, and a claim the check refutes is struck, and the striking recorded.

2.2 THE DESCENT IS A CHAIN, AND A WITNESS IS A ZETA FUNCTION ON IT

Let A be the set of provisions and order its subsets by inclusion. The descent is a chain in that lattice,

$$A = S_0 \supseteq S_1 \supseteq \cdots \supseteq S_n, \qquad S_{i+1} = S_i \setminus \{\, p \in S_i : \lambda_i(p) \in G \,\}, \tag{1}$$

where $\lambda_i(p)$ names the claim in G that puts p out of reach at step i, read against the intention and then against the case, one step for each reason the authority supplies. A provision leaves for one of three reasons. It decides a different question, it conditions on facts the deciding context does not carry, or a stated precedence subordinates it. A witness records the chain and every $\lambda_i(p)$. A checker holding only G and the record verifies at each step the inclusion $S_{i+1} \subseteq S_i$, on which the zeta function of the lattice is one, and that $\lambda_i(p)$ is a claim in G, both machine-checked.

Knuth and Skilling show that any consistent extension of the zeta function to a degree of inclusion on such a lattice must be a probability (Knuth & Skilling, 2012). The descent never takes that step. Every inclusion in its chain either holds or fails, and the checker verifies each one.

The model has no part in the descent. §5.5 measures what it costs to give it one, or to admit a degree.

2.3 VERDICTS TAKE THREE VALUES

An element verdict is one of $\{\mathbf{t}, \mathbf{f}, \mathbf{n}\}$, for true, false, and undetermined. Ordered by knowledge, $\mathbf{n} \sqsubseteq \mathbf{t}$ and $\mathbf{n} \sqsubseteq \mathbf{f}$ with $\mathbf{t}$ and $\mathbf{f}$ incomparable, this is the flat three-valued domain, and the strong Kleene connectives are monotone on it. An element asked in several readings takes their consensus. Agreement yields the common value, and anything else, a disagreement or an unanswered reading, yields $\mathbf{n}$, and the record keeps every reading.

2.4 A DECISION IS A VERSION SPACE, AND A COMMITMENT IS A SINGLETON

Let $R = \rho(G, q)$ be the region the descent reaches for intention q, $E(R)$ its elements, $\varepsilon : E(R) \to \{\mathbf{t}, \mathbf{f}, \mathbf{n}\}$ the verdicts, and Φ the authority's own combination of settled elements into an outcome. Call ε' a completion of ε when $\varepsilon' \sqsupseteq \varepsilon$ and ε' takes only $\mathbf{t}$ and $\mathbf{f}$. The version space of the decision is

$$V(\varepsilon) = \{\, \Phi(G|_R, \varepsilon') : \varepsilon' \text{ a completion of } \varepsilon \,\}, \tag{2}$$

the outcomes the authority reaches under every way the open elements could be settled (Mitchell, 1982). The system commits exactly when $|V(\varepsilon)| = 1$ and otherwise abstains, naming the open elements.

Proposition 1 (Supplying a fact cannot overturn a commitment)**.** *If $\varepsilon \sqsubseteq \varepsilon'$ then $V(\varepsilon') \subseteq V(\varepsilon)$.*

Proof. Every completion of ε' is a completion of ε, since completions are closed upward, and the image of a subset under Φ is a subset of the image.

Where the intention leaves a parameter of the authority unstated, V is the union over its values, and since a zeta valuation cannot sum over the parameter as a probability would, the system commits only if the outcome agrees under every value, and otherwise abstains naming the parameter.

2.5 TWO PATHS BRING AUTHORITY INTO THE RECORD

Proposition 1 is what separates them. **Adjudication** acts on an abstention. A person supplies a fact about one case, an element moves from $\mathbf{n}$ to $\mathbf{t}$ or $\mathbf{f}$, V can only narrow, and one decision changes. The witness of that decision records the supply. **Attestation** acts on the KPG. A person certifies a claim against the authority and it is written into G, whose digest changes. The descent under the new KPG may narrow differently than under the old, so earlier commitments are not preserved, which is why the KPG is versioned and why every witness names the digest it replayed against.

2.6 THE WITNESS IS A PROOF, AND WHAT IT PROVES

A **witness** W of a commitment is a certificate in the sense of proof-carrying code (Necula, 1997; Necula & Lee, 1996). The producer is untrusted, and a small checker that imports nothing from it verifies the artifact. W lists the digests of G and of the record, the chain of Equation 1 with each license $\lambda_i(p)$, every premise the derivation used with its source, the record, a stated rule over recorded readings, or the KPG, and the outcome o. The checker is handed the KPG and the record and verifies six things. The digests are theirs. Each step is a subset of the last and what leaves is what it lists. Each $\lambda_i(p)$ is a claim of G in G's words. Each premise lies in R and is what its readings or the record state, and one that neither settles is listed, so the certificate is conditional on it. And $V(\varepsilon)$, evaluated from G's own data, is the singleton $\{o\}$. A commitment is *right on its record* when the checker certifies it, and the checker is 334 lines of standard-library code (Appendix F). Write φ for the case's facts under the authority's elements, a total assignment of $\mathbf{t}$ or $\mathbf{f}$ to every element of R, and $\Phi^*(A, \varphi)$ for the authority's own decision.

Say that G *represents* A over R when every claim of $G|_R$ is a claim A makes, and $G|_R$ carries A's own combination over the region, its rules and its precedence, so that $\Phi(G|_R, \cdot) = \Phi^*(A, \cdot)$. The second part is completeness. A KPG that cites only true claims but lacks one A makes, an order of precedence, does not represent A, and §5.5 measures what that costs. A witness cites only claims of $G|_R$, so under representation each elimination rests on a license A gives.

Theorem 1 (Soundness)**.** *Suppose G represents A over R, φ is total over R, and the settled verdicts are correct, $\varepsilon \sqsubseteq \varphi$. Then $o = \Phi^*(A, \varphi)$.*

Theorem 2 (Localization)**.** *If $o \neq \Phi^*(A, \varphi)$, then some settled verdict in ε differs from φ, which W names, or $G|_R$ does not represent A, by a claim W cites that A does not make, which W names, or by a claim of A that $G|_R$ lacks, which W cannot name and bounds to the region R.*

Both proofs are in Appendix B, and both theorems are machine-checked in Lean.

Neither theorem depends on how the verdicts in ε were settled. Here a verdict is settled only when every reading of the element agrees. The theorems hold under any settling rule. Every committed decision that disagrees with a label is reported with the location of Theorem 2 (Appendix J).

3 REGISTERS AND PROTOCOL

Meridian (E1–E3) is a generated regulatory register of 111 documents with an amendment and supersession structure and 262 probes in five strata. E1 decides with no model at decision time, E2 injects faults into the intake, and E3 removes the authority's precedence and runs the same engine.

FRE 801–802 (E4) is the hearsay rule, 94 LegalBench cases (Guha et al., 2023), 15 for development and 79 evaluated, with a model answering the elements. **16 CFR 310.3 (E5)** is the Telemarketing Sales Rule, 47 LegalBench cases, 15 for development and 32 evaluated.

ACMG/AMP (E6) is the variant-classification standard (Richards et al., 2015) as four ClinGen Platelet Disorders expert panel specifications refine it, 715 curations the panel submitted to ClinVar, 15 for development and 700 evaluated. A model reads values and stated conditions out of the account, and every comparison and combination is the KPG's. **E6b** is E6's code on the Familial Hypercholesterolemia panel, 537 curations.

Evaluation reads label-free files, and labels are read only when scoring. Claude Haiku 4.5 reads for the KPG, and Claude Opus 5 answers the same cases without a KPG, as the control. The protocol, what is counted and what stops a run, is in Appendix Q. Abstentions are counted in their own column, not against errors. A discharge counts only if the supplied fact came from outside the scored label.

4 RELATED WORK

Selective prediction sets how the results are reported, the share of cases decided beside the share of those decided wrong (Geifman & El-Yaniv, 2017; Kamath et al., 2020), and differs in where the abstention comes from. A threshold declines when the model is unsure. A derivation declines when the authority does not settle the case, and names the element.

Three lines of work are closest to ours, each making a model answer to a written rule. Chain of Logic decomposes a rule into elements and recomposes them (Servantez et al., 2024). Conversational machine reading asks a follow-up question when a needed fact is missing (Saeidi et al., 2018). Neuro-symbolic retrieval synthesizes executable modules from text and detects the knowledge it lacks (Gann & Gertz, 2026). None enforces the authority's answer or abstains when it cannot reach one, none emits a witness that replays without the model, and the solver-based systems answer from the model when formalization fails (Pan et al., 2023; Olausson et al., 2023). Systems that resolve a claim against a theory (Toroghi et al., 2024), run a symbolic chain of thought (Xu et al., 2024), or formalize legal knowledge across agents (Sadowski & Chudziak, 2025) verify the model's own derivation, and the one that declines does so on a failed entailment rather than on a fact it can name (§5.5). Rule-based ACMG classification has a decade of tooling (Li & Wang, 2017), and the witness is a provenance record (Moreau & Missier, 2013) under the obligation of a proof-carrying code certificate (Necula, 1997), re-derived by a checker that trusts nothing the producer did. Retrieval-augmented generation (Lewis et al., 2020; Edge et al., 2024) selects by resemblance, and an authority's definitions are stipulated, which no similarity, lexical or learned, recovers (Appendix P).

5 RESULTS

Every decision the system commits is right, on every register and at every stage. Each is the authority's decision on its record, and the 4 that disagree with a recorded label are located in the label or in a fact the record carries, never in a verdict. Figure 2 shows the complete results.

Meridian, decided without a model. Each names a mismatch between two label vocabularies in the KPG, a fact about the KPG rather than about any one probe, so one attestation settles them. The model is asked once about the KPG's own label vocabulary, never a probe, and each of its 6 mappings is an attested claim that two vocabularies name the same class. Under them 85 of 89 abstentions across both splits convert to correct and none regresses. The test split is 128 of 130 with 2 abstained, and the remaining abstentions on both splits are cases where the operator's class is unstated and the authority states no outcome. Every probe's region contains its gold chain, including the 60 probes whose governing documents share no word with the query (Table 2).

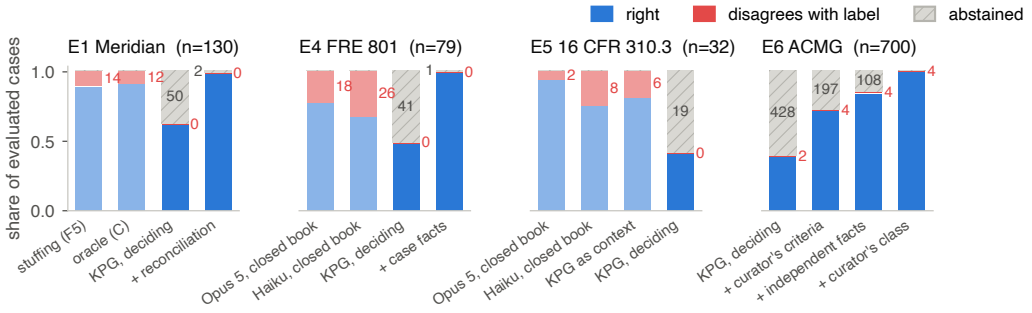


Figure 2: **Right, disagreeing with the label, and abstained on every register, at every stage.** KPG-routed arms in full color, reference arms muted, the count that disagrees with the recorded label beside every bar. An abstention that survives a supplied stage is one no fact from outside the label discharges. The supplied stage is an attested reconciliation on E1, the benchmark's slice annotation on E4, and on E6 the curator's recorded criteria taken as settled, then label-independent facts, then the curator's classification, reported separately because it is the label.

FRE 801, a model in the reading loop. Eleven attested claims record that no case in the register invokes the exclusions of 801(d) or Rule 802, each verified by reading all 94 accounts, so no descent reaches those provisions. The KPG commits 38 of 79 with 0 wrong and abstains on 41. Of the 26 the model gets wrong when answering unaided, the KPG commits on 7 and is right on all 7. Each abstention names one element, and on 39 of the 41 it is the decisive one, whether the statement is offered for its truth. The benchmark's own annotation of each case states the elements, so supplying the one named is adjudication in its plainest form. It discharges 40 of the 41 with none wrong, for 78 of 79, and 1 stays abstained, since its annotation does not state the element it named and nothing else is supplied, so the abstention stands as the right decision.

16 CFR 310.3, abstention as detector. The model answers only the elements and the KPG decides: 13 of 32 committed, 0 wrong, 19 abstained. The model alone is wrong on 8, and with the KPG deciding six of those abstain and two are right. This register has no adjudication stage, since its benchmark annotates nothing beyond the answer.

ACMG, two authorities and 700 decisions. Each record is a curator's written account of the evidence for one variant, the criteria the curator recorded from it, and the class the curator assigned, known as the label. The KPG decides from the account and the recorded criteria, and a model only reads values out of the account. The KPG is run two ways. In the first, a recorded criterion counts only where the account states the condition the panel attaches to it. The KPG then commits 270 right and 2 that differ from the label, and abstains on 428. In the second, every recorded criterion is accepted as recorded. The KPG then commits 499 right and 4 that differ, and abstains on 197. In all four that differ, the recorded criteria combine under the panel's own rules to the class the KPG committed, so the KPG is right on the record it was given, and Theorem 2 says where the difference exists (Appendix J). For two it is in the label: reaching the curator's class needs a criterion the account does not support. For the other two it is in a recorded criterion whose condition the account does not state, PP4 at Strong without full sequencing and BP4 without a REVEL score, and neither is committed when conditions are enforced. That is not a finding that the label is wrong, since a curator may know a fact the account does not state. Each abstention names one missing fact. In the accepted run, 89 of the 197 name a fact available without the label, the gene from ClinVar's gene column or a criterion the account meets under the panel's threshold, and supplying it discharges every one, none wrongly. The other 108 name a fact about the authority itself, which combination rule governs or which strength the panel licenses, an attestation by §2.5. One rule attested for all records was tested and rejected (Appendix I), so these are settled with the curator's own classification and reported as a separate stage, since what is supplied is the label.

ACMG, replicated on a second panel. The Familial Hypercholesterolemia panel (FH) has its own specification and 537 LDLR curations, and the platelet register's code runs on it unchanged (Appendix L). With conditions enforced, the KPG commits 161 right and 0 that differ from the

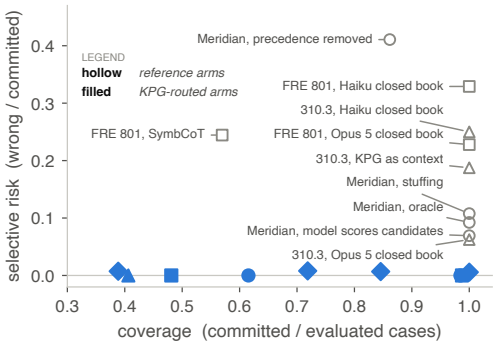


Figure 3: **Selective risk against coverage.** Every arm is one point: coverage is the share of cases it decides, risk is the share of those decisions that are wrong. KPG-routed arms are zero-risk at coverages from 0.39 to 1.

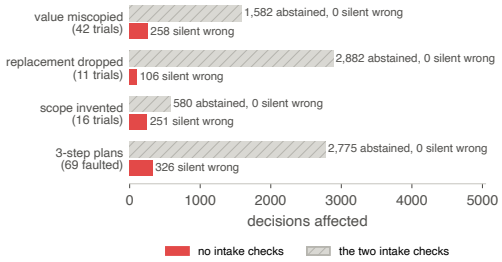


Figure 4: **Intake faults are caught, not absorbed.** 69 single-fault trials in three classes and 87 three-step plans. Unscreened, every affected decision is silently wrong. Screened, every one is an abstention naming the fault.

label, and abstains on 376. With every recorded criterion accepted as recorded, it commits 437 right and 1 that differs, right on its record and located in the label, and abstains on 99. No fact is supplied on this register, so these are its final counts.

5.1 On the selective-prediction plane

Figure 3 reports every arm by two numbers: coverage (the share of cases it decides), and risk (the share of those decisions that are wrong). Every KPG-routed arm has zero risk, whether it decides 0.39 of its cases or all of them. The two arms that decline without a derivation do not. SymbCoT declines on a failed entailment, and its commitments carry a risk of 0.24 (§5.5). The scored ablation lets a model rank the candidates the KPG has narrowed, and it commits 9 wrong decisions it cannot identify, a risk of 0.069. A graded score cannot say which of its decisions are wrong. A hard decline can.

5.2 The loop, measured

Turn zero is the KPG as first attested. Each later turn adds one attestation or one adjudication to the record (Appendix K), and nothing else changes: no model is retrained, and the only thing that differs between turns is the KPG. Coverage rises at every turn on every register. On Meridian and FRE 801 no turn adds a disagreement. On ACMG the one attestation, the panel's practice for BP4 and BP7 on synonymous and intronic variants, adds 29 commitments and one disagreement, and that disagreement is right on its record. Every commitment at every turn is the authority's own decision on its record (Figure 5).

5.3 Faults become abstentions

The guarantee rests on two conditions: that the KPG represents the authority, and that the record is intact. This section tests each by injecting faults against it, first into the KPG and then into the record, and reports what each fault becomes.

We tested the KPG for faults at intake (Figure 4) by injecting 69 of them, in three classes: a value miscopied, a superseded provision's replacement dropped, and a scope condition invented. Unchecked, these faults produce 615 silent wrong answers and no abstentions. With the two intake checks of §2 in place, the source match on every value and the requirement that every document contribute a claim, the silent wrong count falls to 0 and 5044 decisions abstain instead, each naming the claim that failed. The checks do not repair the faults. They convert undetectable errors into named refusals. The same holds when the answers are composed into 87 three-step plans: 326 plans are silently wrong unchecked, 0 checked.

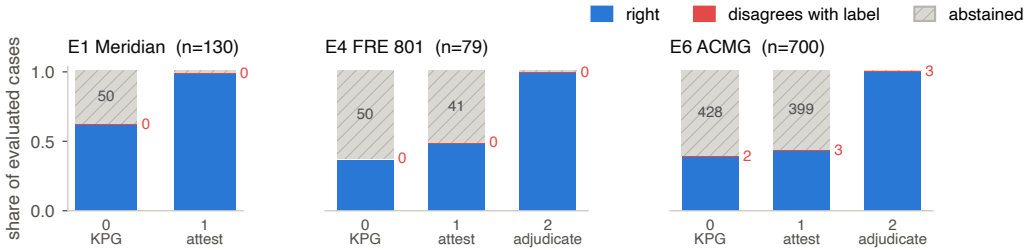


Figure 5: **The correction loop, one turn at a time.** Each turn is one named entry of authority into the record, an attestation to the KPG or an adjudication of one abstention, and each bar is the register after it, the count that disagrees with the label beside it. Appendix K lists what entered the record at each turn.

Second, the record. The checker of §2.6 certifies every commitment on all four registers, 1237 in total. 27 of these certificates are conditional on a single premise, which the certificate states (Appendix F). We then feed the checker tampered copies of every certified witness and every record. It rejects every dropped license, widened step, altered decision, renamed KPG, and altered record: 7418 of 7422 tamperings. The 4 it accepts all alter a premise the certificate had already declared conditional, which lies outside what the certificate claims.

5.4 What carries the guarantee where a model reads

On FRE 801 one element decides the outcome on 35 evaluated cases: whether the statement is offered to prove the truth of the matter asserted. The model reads it five ways, each phrased from a different part of the rule's structure, and each reading alone is right on between 20 and 30 of the 35. Two of the readings tend to err on the same cases, with a Cohen's κ on errors of 0.74 (Table 8). That is the one risk to a unanimity rule: readings that agree while wrong would settle a wrong verdict. It did not happen. On the 9 cases where all five agree they are right on all 9. On the 26 where they disagree the KPG abstains, and a majority vote would have been wrong on 9 of them. Over all 79 cases the readings are unanimous on 20, and there the KPG is right on 18, abstains on 2, and is wrong on none. These counts measure the settling rule, not the guarantee, which is per decision. The theorems hold under any rule, and unanimity is the rule under which the verdict premise held on every committed decision here.

5.5 Ablations, and the closed-book comparison

Precedence removed. With the authority's order of precedence absent from the Meridian corpus the KPG still replays, and the engine commits 46 wrong answers on 130 probes with 18 abstentions. A chain licensed by an incomplete KPG is a valid derivation from the wrong premises, which is why the first condition of the guarantee is attestation and not intake alone. **A model scoring candidates.** The KPG narrows the corpus as before, and a model then ranks the candidates and the top one decides, in place of the KPG's own combination. The KPG alone is right on 128 of 130 with none wrong. The ranked arm is right on 0.931 of the probes and wrong on 9: seven where the governing document shares no content word with the query or is lexically farther than the text it superseded, and the two where the authority states no outcome and the KPG abstains.

The closed-book arms, and a Chain of Logic arm that hands the same model the rule and the KPG's own element list to answer and recombine itself, show what a model alone lacks (Appendix M). The Chain of Logic arm is wrong on 19 of 79 hearsay cases where the KPG, deciding on the same elements, is wrong on none. A **SymbCoT** comparison (Xu et al., 2024) has the model translate the rule and the case into first-order logic and then derive the answer from that logic, all inside the model. On the 79 cases it answers 45 and returns unknown on 34, and of its answers 34 are right and 11 wrong. It is the only comparison that declines, but its unknown says only that the derivation failed, not which fact is missing, so nothing can be supplied to discharge it (Appendix N).

6 DISCUSSION

Correctness is bounded by the KPG, and the bound can be raised. A decision is as right as the ontology and the authority it answers to, and every shortfall is located. A commitment that disagrees with the world names the claim or the verdict that failed, an attestation corrects the KPG, and every later decision inherits the correction with its derivation. An organization converges on correctness by fixing named premises one attestation at a time, and the witness aims each correction.

Where the model is, and where the person is. A reader may ask where the AI is in a system whose decisions are computed without a model. It is in the reading, which no rule engine can do, and it is kept out of the deciding, which no organization can delegate to a rate, since a rate's own condition, exchangeability with the reference class, is usually uncheckable and often known to be false, because a case reaches a system for a reason. That is the relationship proposed here between a probabilistic system and an enterprise that answers for its decisions. Here an expert's judgment enters once, as an attestation or an adjudication, and every later decision resting on it executes without the expert and carries the derivation showing that it does. The organization stands over the loop, and delegation becomes defensible. Coverage before any supply is 62, 48, 41, and 39 percent on the four registers, and of the abstentions settled, 165 took a fact about the authority and 129 a fact about one case.

Authority and purpose. The guarantee certifies that an answer is the authority's decision on its record, not that it is true. A rule is correct relative to a purpose, and the right to name that purpose belongs to whoever answers for it. The converse follows. A value the organization hopes for is not a value it has; a value it has is one it has defined. No model can define it in the organization's place. What the system does is hold the organization to the definition once made: every commitment is checked against it, and where none exists the system abstains.

The KPG keeps its promises under attack. The construction applies wherever an authority states its own combination rules, the theorem's condition. The KPG makes three promises, and three results would break them. A wrong commitment that replays cleanly under a KPG that represents the authority. An abstention whose named fact, supplied, does not discharge it. The KPG makes three promises, and each was tested. First, a commitment that replays under a KPG representing the authority is the authority's decision. None of the 1237 commitments the checker certifies is otherwise, and when the KPG does not represent the authority, in the precedence-removed run of §5.5, 46 wrong commitments replay cleanly, so the condition is the KPG and nothing else. Second, an abstention names a fact that, once supplied, discharges it. Every one of the 129 abstentions discharged by its named fact was discharged rightly. Third, the guarantee holds because no model runs the descent. A model handed the descent turned zero wrong hearsay commitments into fourteen.

7 LIMITATIONS

The guarantee is bounded by the KPG. An authority that does not say how its parts relate yields a KPG that is incomplete rather than corrupted, the case the theorem's completeness condition names, and an incomplete KPG licenses a valid derivation from the wrong premises (§5.5). Where a model reads elements, or a curator's recorded criteria are taken as settled, a verdict is a premise the record carries, and ACMG reports both modes. Only attestation by a person with standing raises the bound. The witness binds the record it decided by digest, and a premise the record does not state and no reading settles stays conditional, which the checker lists. Unanimity settles an element only where the readings agree and does not protect against a bias every reading shares, so a reading-settled verdict is a premise the witness names, not a fact the KPG enforces. The legal registers are small, 79 and 32 cases.

REPRODUCIBILITY

Code and artifacts are at `https://codeberg.org/Josih9/0q9jd9`.

REFERENCES

Anonymous. Prior work by the authors, withheld for review. 2026.

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. Can GPT-3 perform statutory reasoning? In *Proceedings of the International Conference on Artificial Intelligence and Law*, 2023.

Darren Edge, Ha Trinh, Newman Cheng, and others. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Jonas Gann and Michael Gertz. NeSy-RAG: Neuro-symbolic RAG for explainable question answering. *arXiv preprint arXiv:2608.06292*, 2026.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 30, 2017.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, and others. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, 36:44123–44279, 2023.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. A dataset for statutory reasoning in tax law entailment and question answering. *arXiv preprint arXiv:2005.05257*, 2020.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of ACL*, 2020.

Guy Katz, Clark Barrett, David Dill, Kyle Julian, and Mykel Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, 2017.

Kevin H. Knuth and John Skilling. Foundations of inference. *Axioms*, 1(1):38–73, 2012.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, and others. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Quan Li and Kai Wang. InterVar: Clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. *The American Journal of Human Genetics*, 100(2):267–280, 2017.

Nelson F. Liu, Kevin Lin, John Hewitt, and others. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Tom M. Mitchell. Generalization as search. *Artificial Intelligence*, 18(2):203–226, 1982.

Luc Moreau and Paolo Missier. PROV-DM: The PROV data model. W3C Recommendation, 2013.

George C. Necula. Proof-carrying code. In *Proceedings of the 24th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, 1997.

George C. Necula and Peter Lee. Safe kernel extensions without run-time checking. In *Proceedings of the 2nd USENIX Symposium on Operating Systems Design and Implementation*, 1996.

Theo Olausson, Alex Gu, Ben Lipkin, and others. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Empirical Methods in Natural Language Processing*, 2023.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics*, 2023.

Sue Richards, Nazneen Aziz, Sherri Bale, and others. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the ACMG and the AMP. *Genetics in Medicine*, 17(5):405–424, 2015.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. Interpretation of natural language rules in conversational machine reading. In *Proceedings of EMNLP*, 2018.

Albert Sadowski and Jarosław A. Chudziak. On verifiable legal reasoning: A multi-agent framework with formalized knowledge representations. *arXiv preprint arXiv:2509.13032*, 2025.

Sergio Servantez, Joe Barrow, Kristian Hammond, and Rajiv Jain. Chain of logic: Rule-based reasoning with large language models. In *Findings of the Association for Computational Linguistics*, 2024.

Sanjit A. Seshia, Dorsa Sadigh, and S. Shankar Sastry. Toward verified artificial intelligence. *Communications of the ACM*, 65(7):46–55, 2022.

Paul Simpson, John Kozak, and Lisa Doake. Confidently wrong: Exception chain collapse in frontier LLM rule evaluation. *arXiv preprint arXiv:2607.23386*, 2026.

Armin Toroghi, Willis Guo, Ali Pesaranghader, and Scott Sanner. Verifiable, debuggable, and repairable commonsense logical reasoning via LLM-based theory resolution. In *Empirical Methods in Natural Language Processing*, 2024.

Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J. Zico Kolter. Beta-CROWN: Efficient bound propagation with per-neuron split constraints for neural network robustness verification. In *Advances in Neural Information Processing Systems*, 2021.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of the Association for Computational Linguistics*, 2024.

A THE CONSTRUCTION IN FULL

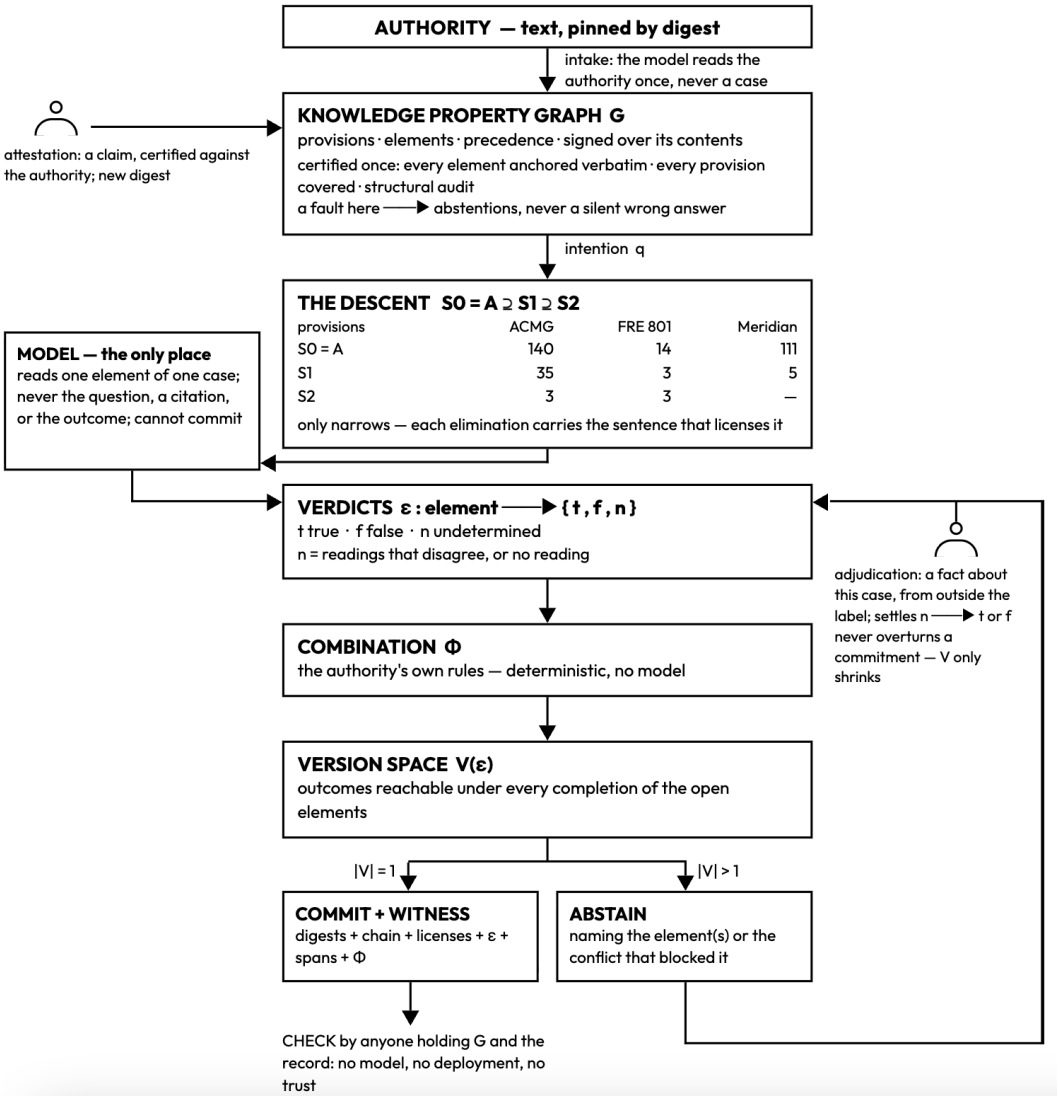


Figure 6: The architecture, in more detail. The certification screen runs once at intake. A fault there produces abstentions, never a silent wrong answer (§5.3). The two human paths enter at different places and have different reach. Attestation changes the KPG and its digest. Adjudication changes one decision and cannot overturn a commitment (Proposition 1).

Kinds of settlement. A vocabulary mismatch between two intake labels needs one reader and no standing. A claim that fails corroboration between independent readings needs a third reading, and the discharge is arithmetic. A determination that turns on a standard, whether something is material or whether a relationship is within scope, needs someone with authority to make it. A change to a precedence claim stands under every elimination in the KPG. The first is an attestation any reader can make. The last is an attestation that should require more than one.

No inference on the decision path. Where the KPG decides alone, no model runs when a decision is made. The model's participation is amortized over intake and confined to element reading, and none of it can commit, so a deployment pays for storage and a traversal rather than for inference per

decision. The record of a single decision replays against the authority as it stood at the time, after the model that read it has been retrained and the infrastructure that served it retired.

B PROOFS OF THE THEOREMS

Proof of Theorem 1. φ is a completion of ε, and every completion reaches o because $|V(\varepsilon)| = 1$. So $\Phi^*(A, \varphi) = \Phi(G|_R, \varphi) = o$.

Proof of Theorem 2. The contrapositive of Theorem 1. The premises the derivation consulted are exactly those W lists, so a failing cited premise is among them, and a failing absent one lies in A's claims over R.

Both theorems are machine-checked in `formal/Witness.lean`, Lean 4 core with no library, and `lean` on that file is part of `reproduce.sh`. The chain of Equation 1 is checked in the same file: each step is a subset of the one before it, the surviving region is S_0 less what any step eliminated, and no provision the authority holds to govern is eliminated at any step when every license cited is a claim of the authority.

C MERIDIAN STRATA, FLOORS, AND BOTH STAGES

Table 1: The five Meridian strata. Each probe is built so that one specific way of reading the register is required, and n is the test split.

stratum	n	what the probe requires
S1 near	30	The query shares three or more content words with a non-binding advisory that carries a wrong figure. The lexically nearest document is the trap.
S2 far	30	The query is in plain language and shares no content word with any governing document. Reaching the answer requires the register's own definitions.
S3 override	22	A bulletin overrides the section, and any amendment to it, for the operator's class. Rank and scope must be composed.
S4 currency	22	The query names a superseded section. Its printed figure is stale, and the amendment governs.
S5 conflict	26	An advisory and the register disagree. Every fifth probe names no operator class, and the authority states no outcome for it.

The majority floor is 0.092 and zero-shot scores 0.015, so the answers are not recoverable from the questions. On S2 both surface retrievers score exactly 0.000. The complete gold chain enters the top six for 0 of 60 probes under BM25 and 0 under TF-cosine. The oracle, given the gold documents, scores 0.667 on S2. Having the right documents is not sufficient, since the reader must still compose rank and scope, and the traversal composes structurally what a reader does by judgment.

D THE DESCENT, MEASURED

The region is the set of provisions the answer must come from, and it is not itself the answer. In the last column of Figure 7 the outcomes a region can reach collapse to exactly one on every committed decision, and to two or more on every abstention that leaves an element open. On ACMG the descent removes strengths more than provisions. A criterion can be applied at several strengths, and the standard admits 140 combinations of criterion and strength. The panel's main specification forbids 105 of them before any case is read, and a case then reaches a median of three combinations. Most of the narrowing is done by the authority before the case is read.

Table 2: Meridian test split ($n = 130$), the share of probes right in each stratum, with abstentions in parentheses where present and never counted as correct. Table 1 defines the strata. F0 is the majority-class floor, F1 zero-shot, F2 and F3 surface retrieval (Robertson & Zaragoza, 2009), F4 model reranking, F5 whole-corpus stuffing (Liu et al., 2024), and C an oracle given the gold documents. Mechanism is the KPG deciding with no model at decision time.

	S1	S2	S3	S4	S5	overall
F0 majority	0.100	0.067	0.182	0.000	0.115	0.092
F1 zero-shot	0.000	0.000	0.000	0.000	0.077	0.015
F2 BM25	0.933	0.000	1.000	0.545	0.346	0.546
F3 TF-cosine	0.933	0.000	0.636	0.727	0.423	0.531
F4 model rerank	0.933	0.700	0.636	0.909	0.538	0.746
F5 stuffing	0.933	0.967	1.000	0.818	0.731	0.892
C oracle	1.000	0.667	1.000	1.000	0.923	0.908
KPG, deciding (abstentions)	0.267 (22)	0.633 (11)	0.682 (7)	0.636 (8)	0.923 (2)	0.615 (50)
KPG, reconciliation supplied	1.000	1.000	1.000	1.000	0.923	0.985
ablation: precedence removed (abst.)	0.567 (4)	0.467 (6)	0.864 (3)	0.318 (3)	0.346 (2)	0.508 (18)
ablation: model scores candidates	1.000	0.867	1.000	0.864	0.923	0.931

Table 3: Right, disagreeing with the label, and abstained on every register and stage (the data of Figure 2).

		right	disagrees with label	abstained
E1 Meridian, KPG alone (n=130)	deciding	80	0	50
	reconciliation supplied	128	0	2
E3 Meridian, precedence removed	deciding	66	46	18
E4 FRE 801 (n=79)	deciding	38	0	41
	case facts supplied (benchmark slice)	78	0	1
E5 16 CFR 310.3 (n=32)	deciding	13	0	19
E6 ACMG, conditions enforced (n=700)	deciding	270	2	428
	curator's classification supplied	698	2	0
E6 ACMG, curator's criteria taken as settled (n=700)	deciding	499	4	197
	independent facts supplied	588	4	108
	curator's classification supplied	696	4	0

E THE CERTIFICATION ATTACK

Table 4: Injected intake faults and composed plans, unscreened and screened.

	silent wrong	abstained
69 injected intake faults, 262 decisions per trial		
no screen	615	0
verbatim anchoring + coverage	**0**	5,044
87 three-step plans, 69 faulted		
no screen	326	0
verbatim anchoring + coverage	**0**	2,775

The fault classes are value miscopies (42 trials), dropped replacement (11), and invented scope (16). Every outcome is scored against the gold answer the corpus generator derives for each probe. The clean KPG, with no fault injected, passes the checks with no claim flagged and all 262 decisions unchanged.

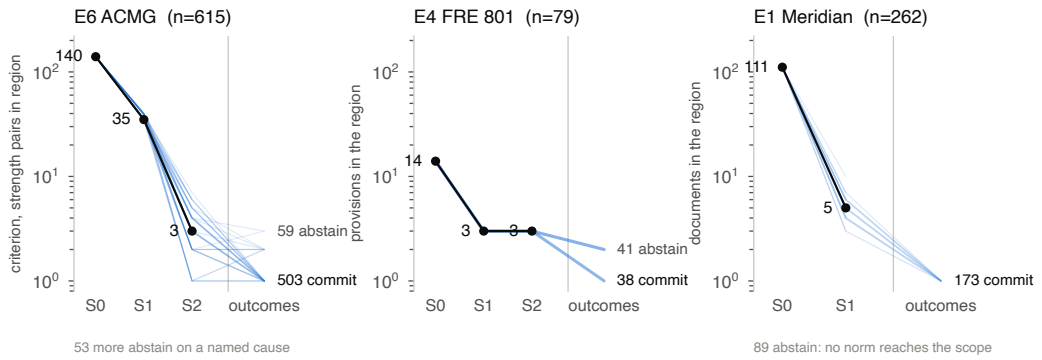


Figure 7: **The descent, ending where the decision is made.** Each line is one path, weighted by the share of decisions that took it, and the black line is the median. S_0 is the authority whole, S_1 what the specification or the intention licenses, S_2 what the case reaches, and the last column the number of outcomes the region can still reach. Every committed decision ends at one outcome and every abstention at two or more, or, on ACMG and Meridian, at a named cause with no region drawn, counted below the panel. Meridian is the KPG before reconciliation, FRE 801 the KPG deciding, and ACMG the curator's criteria taken as settled. The ACMG panel holds the 615 accounts that name a gene, since the 85 that name none reach no region.

F THE WITNESS AS A CERTIFICATE, AND THE CHECKER

The witness is one object per decision. It carries four things: the digest of the KPG bundle and of the record; the descent, as the ground set and a list of steps, each giving the set that remains and every eliminated provision with the license it cites in the KPG's words; the premises, each an element with its value and its source, the record, a reading, or the KPG; and the decision, or an abstention with its cause. An attested claim is recorded in the KPG with the name of who attested it, and the witness reaches it through the KPG digest it names.

The checker is standard-library code (`E9_witness_checker/check.py`) that imports nothing from any register and calls no model. It is handed the witness, the KPG bundle, and the record, and it evaluates the KPG's stated combination from the KPG's own data: on ACMG the counting rules over criteria at tiers, with the points framework's weights as the second rule; on FRE 801 the three-valued element logic over provisions with roles; and on Meridian elimination under stated precedence, where each elimination's trigger is a claim in the KPG. The checker verifies the six obligations §2.6 lists: the two digests, the chain, each license, each premise, and the decision. On the recorded decisions it certifies every commitment (Table 5). The 27 conditional certificates on FRE 801 rest on one element that was read from the case text with no reading recorded, whether the declarant was testifying, and the checker lists that element as the condition. Table 6 puts tampered copies of every certified witness, and of the record each decided, to the checker. The 16 CFR 310.3 register is absent because its witnesses record no descent.

Table 5: The checker on every recorded witness. The witnesses are each register's run of record: Meridian after reconciliation, FRE 801 the KPG deciding, and ACMG and FH with the curator's recorded criteria taken as settled, since those witnesses carry the recorded criteria as premises.

register	witnesses	commitments certified	conditional on a premise	abstentions
E1 Meridian	262	258 / 258	0	4
E4 FRE 801	79	38 / 38	27	41
E6 ACMG	700	503 / 503	0	197
E6b FH	537	438 / 438	0	99

Table 6: Tampered witnesses and records put to the checker, rejected over mutated. Each tampering alters one thing in one copy, and the premise tampering alters the last settled premise, so the four the checker accepts, on FRE 801, are the witnesses whose last settled premise is the one the certificate lists as conditional.

what is tampered		E1 Meridian	E4 FRE 801	E6 ACMG	E6b FH
descent	drop a license	258 / 258	38 / 38	503 / 503	438 / 438
descent	widen a step	258 / 258	38 / 38	503 / 503	438 / 438
descent	alter the decision	258 / 258	38 / 38	503 / 503	438 / 438
graph	name a different KPG	258 / 258	38 / 38	503 / 503	438 / 438
record	alter the record	258 / 258	38 / 38	503 / 503	438 / 438
premise	alter a premise	258 / 258	34 / 38	503 / 503	438 / 438

G FRE 801: READINGS, ABSTENTIONS, AND INDEPENDENT RUNS

The cases are LegalBench's hearsay task (Guha et al., 2023), one line each, and the question is the kind statutory reasoning benchmarks pose (Holzenberger et al., 2020; Blair-Stanek et al., 2023).

Table 7: The five readings of 801(c)(2) on the 35 evaluated cases it decides (the other three elements hold), and what the join buys.

reading of 801(c)(2), on the 35 cases it decides	correct
clause	30
full	29
purpose	23
restate	25
lookup	20
all five agree	9, all correct
readings disagree (abstain)	26; a majority vote is wrong on 9

Table 8: Whether the five readings fail on the same cases. For each pair, the share of cases both get wrong, the share expected if their errors were independent, and Cohen's κ on errors. A unanimity rule is at risk only where readings err together, and the largest overlap, purpose with restate, is the one §5.4 reports. Lookup errs independently of every other reading.

pair	both wrong	if independent	error κ
purpose–restate	0.26	0.10	0.74
full–purpose	0.14	0.06	0.42
full–restate	0.11	0.05	0.36
clause–full	0.06	0.02	0.25
clause–restate	0.06	0.04	0.09
clause–purpose	0.06	0.05	0.04
clause–lookup	0.06	0.06	-0.02
full–lookup	0.06	0.07	-0.07
purpose–lookup	0.11	0.15	-0.14
restate–lookup	0.09	0.12	-0.16

The discharge. The label's outcome lies within the region of every abstention: setting the blocked elements to the assignment the label implies reaches it on 79 of 79. That is a check on the region, not a discharge, since it uses the label. The rows above supply from LegalBench's slice annotation of the case, which names the failing element without stating the label. The annotations are not for truth, non-assertive conduct, in court, standard, and non-verbal. One case stays abstained. Its annotation names a different element from the one the KPG blocked on, so nothing is supplied for that element and the abstention stands, and the label's outcome is within its region like every other.

Table 9: What the 41 abstentions name. Each row is the exact set of elements an abstention named, and each case is in one row.

elements named	cases
801(c)(2) offered for the truth	26
801(a) intended as an assertion; 801(c)(2) offered for the truth	12
801(b) the declarant; 801(c)(1) not while testifying	1
801(a) intended as an assertion; 801(b) the declarant; 801(c)(1) not while testifying; 801(c)(2) offered for the truth	1
801(a) intended as an assertion	1

Table 10: Four independent runs of stage one under default sampling, and the same runs with the blocked elements supplied from the benchmark's slice annotation.

	deciding			case facts supplied		
	right	wrong	abst.	right	wrong	abst.
run of record	38	0	41	78	0	1
run 1	40	0	39	79	0	0
run 2	39	0	40	79	0	0
run 3	38	0	41	79	0	0

The descent on this register. The chain runs $14 \to 3 \to 3$, and the three provisions that survive, 801(a), 801(b), and 801(c), are the definition of hearsay and govern every case, so the region narrows no further. Every provision of 801(d) conditions on a trial-record fact these one-line accounts do not state, and Rule 802 states when hearsay is admissible, presupposing the question. The eliminations are one-directional. Every provision of 801(d) states when something is *not* hearsay, so their absence cannot turn a correct negative into a positive.

H 16 CFR 310.3

On the 32 evaluated cases, as right/wrong, Claude Haiku 4.5 closed-book is 24/8. Haiku with the KPG pasted into its prompt as text and left to decide itself, so that nothing can abstain, is 26/6. The KPG deciding from Haiku's element verdicts is 13/0 with 19 abstained. A null-graph control abstains on all 32, which refutes the reading that abstention is a dodge. Claude Opus 5 closed-book is 30/2.

I ACMG CAUSES, SUPPLY SOURCES, AND WHAT THE TRAVERSAL FOUND

Table 11: The 197 abstentions by cause, what discharges each, and whether the source is independent of the scored label.

cause	what is supplied	n	source
gene unresolved	the gene	85	ClinVar gene column
strength not licensed	which strength governs	44	the curator's classification
governing rule unstated	which combination rule governs	33	the curator's classification
criterion omitted	criteria the account establishes	31	stage-one extraction
criterion recorded twice	which reading of the repeated criterion	2	the curator's classification
stand-alone contradicted	which reading of BA1 applies	2	the curator's classification
independent source		89	
curator's classification		108	
27 records took an independent fact and then the curator's classification, and are counted under the latter			

170 abstentions are settled by one supplied fact. For 27 the case, decided again with that fact, abstains a second time naming a second fact, and the second supply settles it. A fact about the

authority as a whole is an attestation, not an adjudication. Attesting one combination rule for the whole register was tried: it turned 37 abstentions into commitments, 24 of which disagreed with the label, so it was rejected, and those abstentions are settled instead by the curator's classification, case by case.

The class spread is 235 pathogenic, 194 likely pathogenic, 178 uncertain, 47 benign, 46 likely benign. Every number in the paper is against the label the panel recorded.

J AUDIT OF EVERY COMMITMENT THAT DISAGREES WITH ITS LABEL

Table 12: The 4 ACMG commitments that disagree with the recorded label, audited against the record and the two authorities. Rules says whether the standard's combining table (Table 5 of Richards et al., 2015) and the points framework both reach the KPG's class on the recorded criteria. Located in names where the disagreement lies.

record	recorded criteria	curator	KPG	rules	located in
1691471	PM2_supporting, PM3_supporting, PM4, PP3, PP4_strong	Pathogenic	Likely pathogenic	both	label
the curator's class needs PS3, licensed *in a model organism or heterologous cell line*, and the account does not state it					
1691489	PM2_supporting, PM3_supporting, PP4_strong	Uncertain sig.	Likely pathogenic	both	recorded PP4 at strong
its condition, *full sequencing of both genes is required at this strength*, is not stated in the account, and the recorded criteria without it reach the curator's class; withdrawn by the run that reads conditions on recorded criteria					
701285	BP4, BP7, PM2_supporting	Uncertain sig.	Likely benign	both	recorded BP4 at supporting
its condition, *REVEL score of $<$ 0.25*, is not stated in the account, and the recorded criteria without it reach the curator's class; withdrawn by the run that reads conditions on recorded criteria					
953022	PM2_supporting	Likely benign	Uncertain sig.	both	label
the curator's class needs BP7, licensed *use with no specification, which is the standard's BP7 for a synonymous variant*, and the account does not state it; BP4, licensed *REVEL score of $<$ 0.25*, and the account does not state it					

Three checks are run per record, and none takes a judgement. (a) Every recorded criterion and strength is looked up in the governing specification for its applicability. (b) Both published combination rules are evaluated on the recorded criteria at the recorded strengths. (c) Every change of one or two edits to the recorded criteria, a drop, a change of strength, or an addition at a licensed strength, that reaches the curator's class is enumerated, and for each criterion a route relies on, the specification's sentence is quoted and the account is searched for the condition that sentence names. The terms searched are *model organism*, *heterologous*, *cell line*, *transgenic*, and *transfect* for PS3, *sequenc* for PP4 at Strong, *REVEL* for BP4, and *synonymous* for BP7. A commitment is right on its record when (a) is all licensed and (b) both rules reach the KPG's class. A disagreement is located in a recorded criterion when a route that only drops or re-strengths a recorded criterion exists and the account does not state that criterion's condition at its recorded strength. Otherwise every route adds a criterion the curator did not record, the KPG's own threshold check found none omitted on the account, and the disagreement is located in the label. All 4 are right on their record, 2 are located in the label and 2 in a recorded criterion.

Two routes are not settled by a citation. The first record in Table **??**, one of the two located in the label, also reaches Pathogenic if PM3 is read at Moderate, and the panel's PM3 defers to the SVI point recommendations, which the specification cites and does not contain. The KPG can neither license nor block that route and the curator's own PM3 at Supporting is their reading of it. For the two records located in the label, the additions the audit could not check are criteria the curator did not record and the account does not establish by any threshold the panel states.

Conditions read on the recorded criteria. In the recorded-criteria mode, the construction takes the curator's recorded criteria as settled verdicts and reads stated conditions only for criteria the record omits. That is why PP4 at Strong without sequencing and BP4 without a REVEL score are committed rather than withheld in that mode.

The enforced mode is a second stage-one run, `E6_acmg/stage1_conditions.py`, which applies the pipeline's own two checks to the recorded criteria at their recorded strength. A recorded PP4 or PS3 must have the condition for its strength stated in the account, read twice from the specification's sentence as before. A recorded criterion that a quantity establishes, PP3, BP4, BP7, or PS3, must have that quantity stated and satisfying the threshold. Pointers such as *use proposed SVI point recommendations* are not conditions and are not checked.

Under the condition check alone, 186 commitments become abstentions naming the condition, 185 of which had agreed with the label, and stage one is 314 right, 3 disagreeing, and 383 abstained. Most of the withdrawn commitments are PP4 at Strong or Moderate, where the account does not restate the clinical diagnosis and the sequencing that the curator's strength asserts. Of 402 condition readings, 69 disagreed between the two wordings and so count as not stated. Adding the value check withdraws 231 commitments, 229 of which had agreed with the label, and leaves stage one at 270 right, 2 disagreeing, and 428 abstained. Both remaining disagreements are located in the label.

A withdrawn commitment has no independent discharge, since the account does not state the fact. It stays abstained at the independent stage and is settled by the curator's classification at the final stage. The enforced mode is primary in the body because nothing is on the decision path but the KPG and the record. The recorded-criteria mode is reported as the first supply stage, the curator's recorded criteria taken as settled, with the coverage that buys.

What the traversal found in the records. 44 of 700 records apply a criterion at a strength their own specification declares inapplicable, and the KPG names both readings and asks which governs. Benign and uncertain calls end at a median of two criterion–strength pairs, the two pathogenic classes at four, because the standard's combination rules require more evidence at higher tiers.

K THE CORRECTION LOOP BY TURN

Table 13: The correction loop, turn by turn: what entered the record at each turn, and the register after it (the data of Figure 5). A turn of kind *attest* adds a claim to the KPG and changes every later decision. A turn of kind *adjudicate* supplies a fact to abstentions one case at a time.

register	turn	what entered the record	committed	right	disagree	abstained
E1 Meridian	0 (KPG)	graph from intake, certified	80	80	0	50
	1 (attestation)	6 label-vocabulary mappings attested (reconciliation)	128	128	0	2
E4 FRE 801	0 (KPG)	no region claim; all 14 provisions in region	29	29	0	50
	1 (attestation)	11 region claims attested	38	38	0	41
	2 (adjudication)	the decisive element supplied per case	78	78	0	1
E6 ACMG	0 (KPG)	conditions enforced on every recorded criterion	272	270	2	428
	1 (attestation)	The panel applies BP4 and BP7 to a synonymous or intronic variant where SpliceAI predicts no impact to splicing and the nucleotide is not highly conserved	301	298	3	399
	2 (adjudication)	the curator's classification supplied to every remaining abstention	700	697	3	0

On ACMG, turn zero is the run of Appendix J with the KPG's conditions enforced on every recorded criterion. Turn one attests the panel's practice for BP4 and BP7 on synonymous and intronic variants, that they apply where SpliceAI predicts no impact to splicing and the position is not conserved, which the sister specification GN079 states in words and which 17 of 18 identical criteria sets in the panel's own records apply. The claim's sentence is read against each withdrawn account in the same two wordings as every other condition, with the readings in `E8_correction_loop/readings.json`, and the record is re-decided only where both readings say the account states it. Turn two supplies the curator's classification to whatever is still abstained, and ends at 697 right and 3 disagreeing, where the enforced mode of Table 3 ends at 698 and 2. The difference is the one attested claim of turn one.

L THE FAMILIAL HYPERCHOLESTEROLEMIA REPLICATION

The Familial Hypercholesterolemia Variant Curation Expert Panel's specification, GN013 version 1.2, was fetched from the ClinGen registry and parsed by a parser first validated to reproduce the four platelet parses exactly. The panel's LDLR submissions were taken from the pinned ClinVar submission summary, 537 unique variants stating at least one criterion code and 6 narrative-only records set aside. The panel writes its criteria inline, so codes were parsed from the description and replaced in the account by a placeholder, the sentence stating the variant's classification was removed, and the variant's name was kept so the gene resolves. The class spread is 75 pathogenic, 131 likely pathogenic, and 294 uncertain. The decision code is imported from the platelet register unchanged, with LDLR mapped to its specification and no region claim attested. It was frozen before the register was built and did not change. The recorded criteria are the panel's own applied-codes clause, at the strength the account states for each code. The 1 disagreement is audited as in Appendix J, and it is right on its record and located in the label. The conditions-enforced run of Appendix J is made on this register with the same script. Reading the specification's condition on every recorded PP4 and PS3 withdraws 255 commitments and leaves 183 right, 0 disagreeing, and 354 abstained. Adding the value check on PP3, BP4, BP7, and PS3 withdraws 277 and leaves 161 right, 0 disagreeing, and 376 abstained, coverage 30 percent. Nearly every withdrawal is a recorded PP4, whose condition on this panel is a clinical diagnosis by validated criteria that the account rarely restates.

Table 14: FH stage one with the curator's recorded criteria taken as settled, abstentions by cause.

abstention cause	n
gene unresolved	43
governing rule unstated	37
criterion omitted	18
stand alone contradicted	1

Table 15: The FH commitment that disagrees with the recorded label, audited as in Appendix J.

record	recorded criteria	curator	KPG	rules	located in
689349	PM2, BP4, BS3, PP4	Uncertain sig.	Likely benign	both	label

every route to the curator's class drops the recorded BS3, licensed *Level 1 benign functional study criteria, the whole LDLR cycle in heterologous cells*, which the account states, or adds a criterion the curator did not record (PVS1, PP1, PS3, PS2, PS4, PS1, PM5); withdrawn by the run that reads conditions on recorded criteria

M CLOSED-BOOK AND CHAIN OF LOGIC ARMS

Table 16: The comparison arms on the two legal registers, with the KPG deciding beside them. Closed-book arms answer from the model alone. Chain of Logic hands the same model the rule verbatim and the KPG's own element list and lets it answer the elements and recombine them itself. In the closed-book, Chain of Logic, and KPG-as-context arms no abstention is possible, and their failure is the one Simpson et al. (2026) describe: the model answers confidently and misses the exception that governs. SymbCoT's unknowns and the KPG's abstentions are in parentheses.

register	arm	right / wrong
FRE 801 (n=79)	Haiku, closed book	53 / 26
	Opus 5, closed book	61 / 18
	Haiku, Chain of Logic over the KPG's elements	60 / 19
	Haiku, SymbCoT over the rule text	34 / 11 (34 unknown)
	Haiku + KPG, deciding	38 / 0 (41 abst.)
16 CFR 310.3 (n=32)	Haiku, closed book	24 / 8
	Opus 5, closed book	30 / 2
	Haiku, KPG as context, model decides	26 / 6
	Haiku + KPG, deciding	13 / 0 (19 abst.)

The prompts. The templates below are read from the scripts that ran the arms when the tables are generated, so what is printed here is what the runs used. Every arm is zero-shot with no system prompt, at the API's default sampling, one case per request. The element questions the KPG arm puts to the model on FRE 801 are in `E4_fre801/stage1.py`, one template per element, and on ACMG in `E6_acmg/extract.py` and `condition.py`.

Closed book, FRE 801. {`case`} is the case text. Haiku answers with 100 output tokens; Opus 5 thinks by default and answers within 4000.

```
You decide whether described evidence is hearsay under the Federal Rules
of Evidence.

{case}

Reply with ONLY: {"decision":"<Yes|No>"}
```

Closed book, 16 CFR 310.3. The same, over the Telemarketing Sales Rule cases.

```
You decide whether described conduct violates the FTC Telemarketing Sales
Rule, 16 CFR 310.

{case}

Reply with ONLY: {"decision":"<Yes|No>"}
```

Chain of Logic, FRE 801. {`rule`} is every provision of Rules 801 and 802 verbatim with its cite, {`elements`} the KPG's own element list, one line per element, and Haiku answers within 1500 tokens.

```
You decide whether described evidence is hearsay under the Federal Rules of
Evidence, by chain of logic.

RULE (Federal Rules of Evidence 801 and 802, verbatim by provision)
{rule}

ELEMENTS (each provision decomposed into the questions of fact it turns on)
{elements}
```

```
CASE
{case}

Do exactly this, in order.
Step 1. For every element above, answer HOLDS, FAILS, or UNKNOWN from the
case, with one sentence each.
Step 2. State the logical expression that combines the elements under the
rule: the evidence is hearsay when the elements of 801(c) hold and no
exclusion in 801(d) holds.
Step 3. Evaluate the expression on your Step 1 answers.
End with one line of JSON and nothing after it: {"decision":"Yes"} or
{"decision":"No"}
```

N THE SYMBCOT COMPARISON

SymbCoT (Xu et al., 2024) translates a problem into first-order logic, derives a plan, solves by named inference rules, and verifies its own execution, every stage inside the model. The arm runs SymbCoT's released prompts unchanged, with the problem set to Rules 801 and 802 verbatim by provision followed by the case and the question asking whether the described evidence is hearsay (`baselines_legalbench/run_symbcot.py`). Claude Haiku 4.5 answers every stage with 24951 tokens per case on average, and the verify stage changed 13 of the 79 answers. SymbCoT returns true, false, or unknown; true is Yes and false is No.

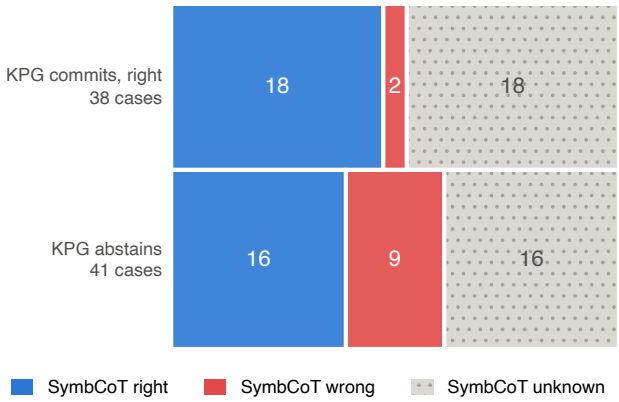


Figure 8: **SymbCoT against the KPG, case by case.** Each row is one outcome of the KPG on the 79 evaluated hearsay cases, sized by its share; each cell is what SymbCoT returned on those cases.

Table 17: SymbCoT and the KPG on the same 79 evaluated hearsay cases.

	SymbCoT	KPG, deciding
right	34	38
wrong	11	0
declined	34 unknown	41 abstained
selective risk	0.24	0.00
tokens per case	24,951	one element question per reading

Table 17 gives the totals and Figure 8 places every case. On the 38 cases the KPG commits, all of them right, SymbCoT is right on 18, wrong on 2, and unknown on 18. On the 41 the KPG withholds, SymbCoT commits wrongly on 9 and is right on 16. Its coverage is 0.57 and its selective risk 0.24.

SymbCoT declines, and the KPG declines, and the two declines are different objects. SymbCoT's unknown is a failed entailment: its translation of a case rarely asserts that a statement is offered for

its truth, so the solver cannot derive hearsay and stops, and what remains is a proof that did not close. It names no fact a person could supply. The KPG's abstention names the element, and §5 discharges 40 of 41 by supplying it.

The two kinds of answer are different objects too. SymbCoT's answers are derivations the model wrote and the model verified, and every one of its 11 wrong answers ends in a completed derivation that the verify stage accepted, two of them on cases the KPG decides correctly from the same rule. The KPG's commitments are certified by a checker that imports nothing from the model.

What SymbCoT lacks, then, is not formal reasoning or the ability to decline. It lacks a decline that says what is missing, and a check on its answers that does not come from the model that wrote them. Without those two, a wrong answer and a right one look the same, so formalizing inside the model does not produce a guarantee at the decision.

O COST

Table 18: Prompt characters per decision on Meridian, the register that decides with no model, and the region an auditor reads under the KPG. On the applied registers the KPG spends one model read per element. The scored ablation is the only KPG-based arm that issues a prompt at decision time.

	p10	median	p90	accuracy
C oracle (gold docs given)	493	697	886	0.908
F3 TF-cosine	868	1,316	1,429	0.531
F2 BM25	1,316	1,363	2,125	0.546
ablation: model scores candidates	1,483	1,764	2,894	0.931
F5 stuffing	17,589	17,602	17,619	0.892
KPG, deciding: region an auditor reads (documents of 111)		5	6	0.985
KPG, deciding: region an auditor reads (characters)	1,192	1,450	2,736	

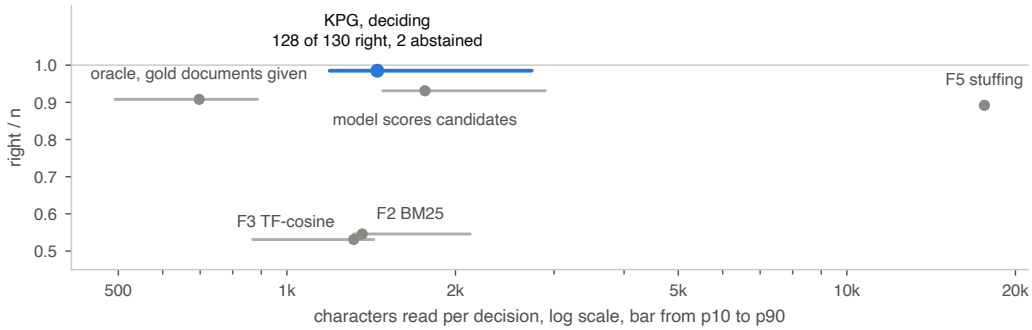


Figure 9: **What a decision costs to read, against how often it is right.** Every prompting arm at the characters of its prompt, and the KPG arm, which issues no prompt, at the characters of the region its witness names, which is what an auditor reads to check it. Stuffing reads ten times the region and carries 14 wrong answers with no derivation.

P WHY SELECTION CANNOT REST ON SIMILARITY

Proposition 2 (Surface-overlap degeneracy)**.** *Let a retriever score documents by* $s(q, d) = f(\mathrm{ov}(q, d))$ *with* f *monotone in the query–document content-word overlap, and let a probe satisfy* $\mathrm{ov}(q, g) = 0$ *for every governing document* g*. Then* $s(q, g)$ *is minimal over the corpus, and if at least* k *distractors share one content word with* q*, no governing document enters the top* k*.*

The proposition is about surface overlap. A learned matcher is not bound by it, and one may map a paraphrase to its target with no overlap at all.

Proposition 3 (Authority reach)**.** *If the KPG carries a definition claim mapping a phrase of the intention to a term, and provisions stated over that term, then the region derived by lookup over the KPG contains every governing document, independent of overlap.*

What separates the two is not the matcher but what licenses the reach. An authority stipulates its definitions. On Meridian a hull survey is checking the condition of equipment because the definitions document says so, and a stipulated alias resembles nothing. No similarity, lexical or learned, recovers it unless the authority's own definitions were put into the weights, where they are neither citable nor versioned. In the construction the mapping is a definition claim in G, the region is derived by lookup over it, the witness cites it, and the checker verifies that the record states the term or an alias G states. A learned matcher may propose the binding. It is then a reading, and its proposal is a premise the witness names, which the checker verifies against G or lists as unverifiable. Similarity can select, and only a claim can license.

Both propositions are measured on Meridian's S2 stratum (Appendix C): the gold chain enters the top six for 0 of 60 probes under either surface retriever, and the region contains it for every probe.

Q Evaluation protocol

The protocol is the set of rules the evaluation runs under, written before any scored run. It names the testbed pin, the primary metric (test-split accuracy on the stratum whose governing documents share no word with the query), the secondary metrics, the model under test, and four stopping rules: a corpus the judged model can guess zero-shot is regenerated; stuffing that matches the KPG on every stratum reframes the claim; a KPG below the best surface retriever on the primary stratum ships as a negative result; and a KPG that keeps planted bait alive more often than stuffing is reported as a defect. Intake is general, by the judged model against a schema that knows nothing of the corpus. A remediation counts only if it discharges the cause named. The hearsay and platelet ACMG pipelines were developed on the sets they are evaluated on, so those figures bound the construction rather than select it, and the Familial Hypercholesterolemia register is the held-out test.

Use of generative AI

All of the technical ideas in this paper, the construction, its claims, and its mathematical representation, originated with one author, who directed and reviewed every step. Language models appear in two roles. As the subject of the experiments, Claude Haiku 4.5 reads the authority at intake and one element of one case at decision time, and Claude Haiku 4.5 and Claude Opus 5 are the closed-book comparison, exactly as §3 describes. As a tool, a generative AI coding assistant, under that author's direction, implemented the code and analysis scripts, including the deterministic seeded generator that produces the Meridian corpus, drafted and edited prose and proofs from that author's arguments, encoded the theorems in Lean, produced the figures, and gave feedback on experimental design and on drafts, which the author accepted or rejected. We have reviewed all AI-assisted work. Every number regenerates from the run files by the scripts that ship with the paper, the theorems are machine-checked, and the authors read and verified the generated text, code, and outputs at each step. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.