The Reliability of Legal Artificial Intelligence

Seven measurable properties, the arithmetic of release,
and what a pre-registered benchmark shows

Research paper LL-REL-01
Version 1.0.0, 26 November 2026

Abstract

Legal artificial intelligence is sold on fluency and judged on anecdotes. Neither measures what a lawyer relies on. A lawyer relies on a statement because the official text it names supports it, at the right date and in the right legal order. This paper makes that standard measurable. It defines the reliability of a legal output as seven properties that can be tested separately: proposition-level support, citation existence and resolution, quotation fidelity, temporal applicability, jurisdictional applicability, abstention and calibration. Apart from resolution, which the others presuppose, passing one property does not establish another, and no single score can stand for them.

The analytical part derives five results. Answer-level reliability cannot exceed that of the weakest assertion, decays geometrically with the number of assertions under independence, and has a floor, one minus the sum of the per-assertion failure probabilities, that falls linearly under any dependence. The error rate among the outputs a verification gate releases is a closed function of the base error rate and the gate's false-accept and false-reject rates, so a gate that accepts every error of a class leaves that class's error rate untouched. The value of a gate therefore depends on base rates that accuracy figures hide. And a weighted average of verification stages is not a probability: it can be constant on correct and incorrect outputs alike, and nothing ties it to the evidence. A fifth explains why similarity measures cannot decide support: their verdict on a one-word change depends on its length and position, not on its legal effect.

The empirical part reports LLRB-1, a pre-registered benchmark of 4,053 test items built from the official text of eight EU acts [MEASURED]. Lawie's lexical grounding check, as implemented at the code basis, accepted every statement that carried one legally decisive edit and every statement cited to the wrong provision [MEASURED]. Calibrated containment, token F1 and Jaccard, each thresholded at the 5th percentile of supported development scores, still accepted 27.1%, 54.2% and 51.8% of decisive edits [MEASURED]. The pipeline's reported confidence was the same for supported and unsupported statements [MEASURED]. A typed research prototype rejected every edit of the kinds it models and accepted 88.5% of the kinds it does not [MEASURED]. A temporal index equated entry into force with application [MEASURED], a jurisdiction rule missed five foreign legal orders [MEASURED], and a citation-required rule passed fabricated citations [MEASURED]. The paper ends with requirements stated as testable properties, and with what was not measured: no model-generated answer was evaluated [PROPOSED], and no lawyer has reviewed the benchmark's labels [ABSENT].

Prepared by Lawie Labs. *Status:* First edition. *Distribution:* Public. *Basis:* source repository `main` @ `ffa13d4bce` (2026-11-24); experiments of 26 November 2026; EUR-Lex texts retrieved 26 November 2026. *Audience:* Senior lawyers, legal technologists, AI researchers, regulators and diligence teams.

Evidence labels. Every material statement about Lawie carries an evidence class. [IMPLEMENTED] in source at the basis commit; [TESTED] exercised by repository tests at that commit; [MEASURED] measured by an offline experiment of this suite on the stated public or synthetic data, never a production or customer measurement; [ANALYTICAL] follows by argument from stated assumptions; [SIMULATED] output of a stated model; [EXTERNAL] checked against a cited primary source; [PROPOSED] design or protocol not built or not run; [UNVALIDATED] a hypothesis, stated as such; [ABSENT] not implemented; [BLOCKED] awaits an external owner or decision. Statements describe the source repository at the basis commit, not a deployed service: no workflow described here has a dated production acceptance, nothing is independently audited, and no statement has been reviewed by an independent human expert. This is a technical publication, not legal advice, not an offer of services and not a statement that any use of Lawie complies with any law.

Position

Intelligence is not authority. A model's competence never confers the warrant to be believed. A legal statement earns reliance from the official text it names, read at the right date in the right legal order, and from a procedure that withholds what it cannot verify. This paper treats reliability as that procedure's measurable output, not as a quality of the model behind it.

Six positions

1. **Reliability is a vector, not a score.** Support, resolution, quotation, time, jurisdiction, abstention and calibration fail independently, so each is measured and reported on its own. *Cost:* there is no headline number; a buyer must read a table.
2. **Support is decided by decisive elements, not by resemblance.** Who must act, whether they must, by when, how much, under which provision and with which exception: a statement that changes any of these is unsupported, however similar it looks. *Cost:* every element type must be typed and tested, and types not yet modelled stay open; the prototype measured here leaves most of them open (section 4.2).
3. **A gate is judged by what it releases.** The quantity that matters to a reader is the error rate among released outputs, and it depends on the base rate of errors as much as on the gate. *Cost:* the released error of a real deployment cannot be stated without measuring the base rate of each error class in generated answers, which this programme has not done.
4. **A score is not a probability until outcomes make it one.** A number shown beside an answer will be read as a probability of correctness. Unless it has been calibrated against labelled outcomes from the setting in which it is shown, it must not be shown. *Cost:* calibration needs labelled data from each deployment, and until it exists the honest display is no number at all.
5. **Refusal is an output.** When support is not established, the correct answer is that it is not established. *Cost:* coverage falls, and users see more answers that end in "not established".
6. **Negative results are published.** Several of Lawie's own components fail the tests in this paper. The failures are reported as measured outcomes and rates, because a reader who cannot see what fails cannot judge what passes. *Cost:* the paper is evidence against its authors' product as it stands.

Contributions

K1 A definition of the reliability of a legal output as seven separately testable properties, with what establishes each and what does not (section 2).

K2 Closed-form results on compounding, on the released error of a gate, on base rates, on calibration and on the blind spots of similarity measures, each proved from stated assumptions (section 3).

K3 Pre-registered measurements of verification components on official EU legal text, including negative results on Lawie's components as implemented at the code basis and the mechanism of each (section 4).

K4 Requirements for reliable legal AI stated as testable properties with explicit parameters and costs (section 5).

K5 An explicit boundary of what the evidence does not show (section 6).

Contents

Abstract **1**

Position **2**

1 Introduction **4**

1.1 The failure that matters 4

1.2 Why competence is the wrong frame 4

1.3 What this paper does 4

2 Reliability as seven properties **5**

2.1 The object of measurement 5

2.2 The seven properties 5

2.3 Independence 6

3 Analytical results **7**

3.1 Compounding 7

3.2 The released error of a gate 8

3.3 Why base rates matter 9

3.4 Calibration, and why a stage score is not a probability 9

3.5 Why similarity cannot decide support 10

3.6 Exact and canonical matching 12

4 Evidence from a pre-registered benchmark **12**

4.1 Design in brief 12

4.2 Similarity is not support 13

4.3 Confidence carries no information 14

4.4 Quotation fidelity: blind, useless, brittle, exact 14

4.5 Citation resolution 15

4.6 Time and jurisdiction 15

4.7 Release discipline 16

5 Requirements for reliable legal AI **16**

6 What was not measured **18**

7 Conclusion **19**

References **20**

1 Introduction

1.1 The failure that matters

In 2023 a United States federal court sanctioned lawyers who had filed a brief citing judicial decisions that did not exist; a chatbot had produced them, complete with quotations and internal citations, and the lawyers had not checked them [1]. The episode is cited as a warning about careless lawyers. It is better read as a specification. Every element of the failure is checkable: whether the cited authority exists, whether it says what the brief attributes to it, whether the quotation is in it. None of those checks requires judgment about the merits of the case. They require access to the official text and a procedure that compares.

The professional rules have drawn the same conclusion. The American Bar Association's formal opinion on generative tools requires a lawyer to understand their capabilities and limits and to verify what they produce before relying on it [2]. The research record shows why. Dahl and colleagues found legal hallucinations pervasive across general-purpose models asked verifiable questions about case law, and found that the models often accepted a user's false legal premise and could not reliably tell when they were wrong [3]. Magesh and colleagues evaluated commercial legal research tools built on retrieval, including tools whose vendors claimed to avoid hallucination, and found that retrieval reduces hallucination but does not remove it; their definition counts an answer as a hallucination when it is incorrect or when it is *misgrounded*, that is, when it cites a source that does not support it [4]. The survey literature separates intrinsic hallucination, which contradicts the source, from extrinsic hallucination, which the source cannot confirm [5]. Both appear in legal practice as a wrong statement with a real citation attached.

The industry's response has been to place a verification layer between the model and the user: retrieve sources, compare the answer with them, attach citations, and report a confidence. This paper asks what would make such a layer reliable, and how one would know. The question is not rhetorical. A verification layer that is itself unreliable is worse than none, because it converts an unverified statement into a statement that appears verified.

1.2 Why competence is the wrong frame

The suite's doctrine is that intelligence is not authority. In law the point is old: what counts as law is fixed by criteria of recognition, not by the persuasiveness of a statement about it [6]. A model that writes an excellent paragraph about Article 33 of the General Data Protection Regulation has not thereby made the paragraph true of Article 33. Whether it is true is a relation between the paragraph and the official text, at a date, in a legal order. That relation can be checked without trusting the model at all.

This reframes reliability. It is not a property of the model, to be read off a leaderboard. It is a property of the relation between an output and the law it invokes, and of the procedure that decides whether to release the output. The NIST framework lists validity and reliability first among the characteristics of trustworthy AI [7]; for legal outputs the characteristic must be broken into parts that can be tested against the official text, because each part fails for different reasons and is repaired by different means.

1.3 What this paper does

Section 2 defines reliability for legal outputs as seven separately measurable properties and shows that, apart from resolution, none establishes another. Section 3 proves results about how per-assertion reliability compounds, what a release gate does to the error rate of what it releases, why base rates decide the value of a gate, what calibration means and why a weighted stage score is not a probability, and why similarity measures cannot decide support. Section 4 reports measurements from LLRB-1, a pre-registered benchmark built from the official text of eight EU acts, in which Lawie's own components, as implemented at the code basis, were run unchanged beside research baselines and a research prototype. Several of Lawie's

components fail; the failures are reported as measured outcomes and rates. Section 5 states requirements for reliable legal AI as testable properties. Section 6 states what was not measured, which includes the most obvious question: how often generated legal answers are wrong.

Three distinctions run through the paper. *Analytical* results are derived here from stated assumptions and carry proofs. *Empirical* results were measured by the suite's experiments on stated data and carry the label [MEASURED]; none is a production or customer measurement, because nothing described here is deployed in production [BLOCKED] and no Lawie product has a dated production acceptance [ABSENT]. *Literature* is cited. The benchmark's full method, its datasheet and its threats to validity are in the companion report, *Legal AI Grounding and Hallucination Benchmark* (LL-BEN-01).

2 Reliability as seven properties

2.1 The object of measurement

Definition 2.1 (Legal assertion)**.** A *legal assertion* is a tuple $a = (\varphi, c, q, t_q, J)$: a proposition φ about the law; a citation string c that names a provision; an optional quotation q presented as the provision's text; the date t_q the question is about; and the legal order J that governs the matter. A *legal output* is a finite set of assertions. A citation *resolves* when it denotes an address in $\mathscr{A}$ (instrument, article, paragraph, point); an address and a date determine a version of a provision s, whose official text is $\mathsf{text}(s)$.

The definition is deliberately atomic. A memorandum, a contract review or an answer to a research question is a set of such assertions, and its reliability is the reliability of each of them (section 3.1). The decomposition of long text into atomic claims is itself a research problem; FActScore decomposes generations into atomic facts and scores the share supported by a knowledge source [8], and that step is assumed here, not solved.

2.2 The seven properties

Definition 2.2 (Proposition-level support, P1)**.** $\mathsf{Supp}(s, \varphi)$ holds when a competent reader of $\mathsf{text}(s)$ would accept φ as stated by s, with every *decisive element* preserved: the actor who bears the duty or holds the power, the deontic modality and its polarity, quantities and their comparators, dates and time units, cross-references, and the scope of the rule including its exceptions. An assertion is *misattributed* when φ is supported by some provision but not by the one c names.

Support is a semantic relation, and nothing in this paper reduces it to a string operation. What can be reduced is a *sufficient* test: the provision's own words, and typographic variants of them, are supported; a statement that changes one decisive element is not. The attribution literature defines related notions: a statement is attributable to a source when a generic reader would agree that, according to the source, the statement holds [9]; citation recall and precision ask whether the cited passages entail the statement and whether each citation is needed [10]; faithfulness asks whether a generated answer is inferable from the retrieved context [11]. Legal support is narrower in one respect and stricter in another. It is narrower because the source is fixed by the citation: support by some other passage is a misattribution, not a success. It is stricter because a statement that is nearly entailed is wrong: "may" for "shall" is a small difference in wording and the whole difference in law.

Definition 2.3 (Citation existence and resolution, P2)**.** A citation *exists* when it resolves to an address in $\mathscr{A}$ that the instrument contains. Resolution is *format-invariant* when every official and customary rendering of the same address resolves to the same address, and *refusing* when a rendering of a non-existent address resolves to nothing.

Definition 2.4 (Quotation fidelity, P3)**.** For a declared canonicalisation κ that folds only characters without legal meaning, a quotation q is *faithful* to s when $\kappa(q)$ is a substring of $\kappa(\mathsf{text}(s))$.

Fidelity is relative to κ, and κ is a legal choice as much as a technical one. Folding a no-break space into a space changes nothing in law; folding case can merge a defined term with a common word; folding a superscript digit into a digit can merge two numbers (section 3.6). A quotation check that does not declare its κ cannot be audited.

Definition 2.5 (Temporal applicability, P4)**.** Four dates bound the life of a provision: publication t_{pub}, entry into force t_{eif}, the date from which the provision applies $t_{\text{app}}(s)$, which may differ by provision and by category of addressee, and repeal or expiry t_{rep}. A fifth date, the knowledge date t_k, fixes which version of the schedule of dates is in use. $\mathsf{App}(s, t_q)$ holds when $t_{\text{app}}(s) \le t_q < t_{\text{rep}}$ under the schedule known at t_k.

The separation of the time at which a fact is true from the time at which it is recorded is standard in temporal databases (valid time and transaction time) [12], and legal informatics models several temporal dimensions of a legal rule [13]. EU law makes the distinction concrete: the General Data Protection Regulation entered into force in May 2016 and applied from May 2018 [EXTERNAL]; the Artificial Intelligence Act applied in stages from 2025 to 2027 as published [EXTERNAL], and an amending regulation of 2026 moved several of those dates while they were pending [EXTERNAL]. A statement can be true of a provision and still be wrong about the law on the date in question.

Definition 2.6 (Jurisdictional applicability, P5)**.** s is *admitted* for the legal order J when its instrument is part of J's law, directly or by transposition, or when a conflict rule of J makes it applicable. An assertion is *jurisdictionally applicable* when the provision it cites is admitted for J.

Definition 2.7 (Abstention, P6)**.** A system *abstains* on an assertion when it withholds it, or marks it as not established, instead of releasing it as established. Abstention is *correct* when the assertion fails one of P1–P5 or cannot be shown to meet them.

Definition 2.8 (Calibration, P7)**.** A confidence $C \in [0, 1]$ reported with an output is *calibrated* for a population of outputs when $\Pr(Y = 1 \mid C = c) = c$ for every reported value c, where $Y = 1$ when the output is correct.

Abstention and calibration are properties of the procedure, not of a single assertion, and both are measured over a population. Chow's analysis of the error–reject trade-off [14] and the later theory of selective classification [15] treat abstention as a decision with a cost, to be traded against coverage. Calibration is assessed by proper scores such as Brier's mean squared error of probability forecasts [16] and by binned calibration error [17].

2.3 Independence

Observation 2.1 (No property establishes another)**.** *Apart from resolution, which the others presuppose, passing one property does not establish another. A faithful quotation of Article 33(2) of the General Data Protection Regulation ("The processor shall notify the controller without undue delay after becoming aware of a personal data breach."), offered for the proposition that the processor must notify the supervisory authority, satisfies P3 and fails P1. A supported statement of a provision of the Cyber Resilience Act that applies from the act's general date of application, 11 December 2027, made about a date in 2026, satisfies P1 and fails P4 [EXTERNAL]. A correctly resolved citation of a German statute in an answer for a Czech matter satisfies P2 and fails P5 unless a conflict rule makes the statute applicable. A system that abstains correctly and reports the same confidence c for every answer, with c different from its accuracy, satisfies P6 and fails P7.*

Table 1: The seven properties: what establishes each, what does not, and where LLRB-1 measures it.

Property	What can establish it	What does not establish it	LLRB-1 family
P1 Support	Comparison of every decisive element with the cited provision's text	Word overlap; embedding similarity; the presence of a citation	Support (SUP)
P2 Resolution	Resolution of each rendering to an address that exists in the instrument	A string that looks like a citation	Citation (CIT)
P3 Quotation	Canonical substring match against the official text under a declared κ	Similarity above a threshold; a match against a title or a summary	Support (SUP)
P4 Time	Per-provision application dates under the schedule known at t_k	An entry-into-force date; a publication date	Temporal (TMP)
P5 Jurisdiction	Admission of the cited instrument for the matter's legal order	The absence of a recognised foreign marker	Jurisdiction (JUR)
P6 Abstention	A release rule that withholds when P1–P5 are not shown	A warning attached to a released answer	Release (GRD)
P7 Calibration	Agreement of stated confidence with observed frequency on labelled outcomes	A weighted sum of stage results	Support (SUP)

The observation has a direct consequence. A single accuracy figure aggregates properties that fail for different reasons and are repaired by different means; it cannot say which one failed, and a gain in one can hide a loss in another. Reliability for legal outputs is therefore reported as a vector, one entry per property, with the population and the interval stated for each.

3 Analytical results

This section derives what follows from the definitions alone. Each result states its assumptions; none depends on the measurements of section 4, which test some of them.

3.1 Compounding

Assumption 3.1. *An output of k assertions is relied on as a whole: it is reliable when every assertion is. Let A_i be the event that assertion i satisfies the properties required of it, with $r_i = \Pr(A_i)$.*

Proposition 3.1 (Compounding)**.** *Under assumption 3.1,*

$$1 - \sum_{i=1}^{k} (1 - r_i) \;\le\; \Pr\Big(\bigcap_{i=1}^{k} A_i \Big) \;\le\; \min_i r_i,$$

for any dependence between the events, and $\Pr(\bigcap_i A_i) = \prod_i r_i$ when they are independent. The upper bound is attained; the lower bound is attained when $\sum_i (1 - r_i) \le 1$ (in general the bound is $\max\{0, 1 - \sum_i (1 - r_i)\}$).

Proof. The intersection is contained in each A_j, which gives the upper bound. The complement of the intersection is the union of the complements, and Boole's inequality bounds its probability by $\sum_i (1 - r_i)$, which gives the lower bound. The product form is the definition of independence. The upper bound is attained by nested events, the lower bound by failure events that are pairwise disjoint when $\sum_i (1 - r_i) \le 1$. □

Corollary 3.1. *With equal per-assertion reliability r, an output of k assertions is reliable with probability r^k under independence and at least $1 - k(1 - r)$ under any dependence. A target R for the whole output is met under any dependence when $1 - r \le (1 - R)/k$; under independence it requires $r \ge R^{1/k}$.*

Worked example 3.1. A memorandum of 20 assertions, each reliable with probability 0.99, is reliable as a whole with probability $0.99^{20} \approx 0.818$ if the assertions fail independently, and with probability at least 0.80 whatever their dependence: under independence, roughly one memorandum in five contains a wrong assertion. To make the memorandum reliable with probability 0.99, a per-assertion failure probability of at most 0.0005 secures it under any dependence. Per-assertion reliability that sounds high is not high enough.

The error budget divides twice. Each assertion is itself a conjunction of the properties of section 2, so with $\varepsilon_{i,p}$ the probability that assertion i fails property p, $1 - r_i \leq \sum_p \varepsilon_{i,p}$ by the same inequality. A system that is excellent on resolution and quotation and weak on time is at most as reliable as its weakest property. The conjunctive structure is what makes legal outputs expensive to check by hand: the reader cannot know which assertion failed without checking all of them. Verification therefore has to be done, and its outcome recorded, per assertion.

3.2 The released error of a gate

Definition 3.1 (Gate)**.** A *gate* G decides for each output whether to release it (ACCEPT) or withhold it (REJECT). Let E be the event that the output is erroneous, $e = \Pr(E)$ the *base error rate*, $F = \Pr(\mathsf{ACCEPT} \mid E)$ the *false-accept rate* (FAR) and $R = \Pr(\mathsf{REJECT} \mid \neg E)$ the *false-reject rate* (FRR). The *coverage* is $\mathrm{cov} = \Pr(\mathsf{ACCEPT})$ and the *released error* is $\mathrm{risk} = \Pr(E \mid \mathsf{ACCEPT})$.

Proposition 3.2 (Released error)**.** *For* $\mathrm{cov} > 0$*,*

$$\mathrm{cov} = eF + (1-e)(1-R), \qquad \mathrm{risk} = \frac{eF}{eF + (1-e)(1-R)}.$$

Proof. By total probability, $\Pr(\mathsf{ACCEPT}) = \Pr(\mathsf{ACCEPT} \mid E)\Pr(E) + \Pr(\mathsf{ACCEPT} \mid \neg E)\Pr(\neg E)$, which is the first identity. Bayes' rule gives $\Pr(E \mid \mathsf{ACCEPT}) = \Pr(\mathsf{ACCEPT} \mid E)\Pr(E)/\Pr(\mathsf{ACCEPT})$, which is the second. □

The formula [ANALYTICAL] is elementary, and it is routinely ignored: verification layers are described by their accuracy on a test set, which mixes F and R in proportions set by whoever built the test set, and says nothing about risk.

Corollary 3.2 (When a gate helps)**.** *For* $0 < e < 1$ *and* $\mathrm{cov} > 0$*:* $\mathrm{risk} < e$ *if and only if* $F < 1 - R$*, and* $\mathrm{risk} = e$ *if and only if* $F = 1 - R$*. A gate that accepts every error of a class (*$F = 1$*) cannot lower the released error for that class, and raises it when it rejects any correct output (*$R > 0$*).*

Proof. $\mathrm{risk} < e$ is equivalent to $F < eF + (1-e)(1-R)$, that is to $F(1-e) < (1-e)(1-R)$, that is to $F < 1 - R$. The equality case is the same computation. With $F = 1$ the condition $F < 1 - R$ fails for every $R \geq 0$, and $\mathrm{risk} > e$ when $R > 0$. □

Corollary 3.3 (The false-accept rate sets the divisor)**.** *For* $F > 0$*,* $e/\mathrm{risk} = e + (1-e)(1-R)/F$*. When* $F \leq 1 - R$ *the gate therefore divides the base error rate by a factor between* $(1-e)(1-R)/F$ *and* $(1-R)/F$*; for a small base rate the factor is close to* $(1-R)/F$*.*

Proof. By proposition 3.2, $e/\mathrm{risk} = \mathrm{cov}/F = \big(eF + (1-e)(1-R)\big)/F = e + (1-e)(1-R)/F$, which is at least $(1-e)(1-R)/F$. When $F \leq 1 - R$, $e \leq e(1-R)/F$, so $e/\mathrm{risk} \leq e(1-R)/F + (1-e)(1-R)/F = (1-R)/F$. □

Errors come in classes, and gates differ by class. Let the erroneous outputs be partitioned into classes $m = 1, \ldots, M$ with base rates e_m, $\sum_m e_m = e$, and class false-accept rates F_m. Then

$$\mathrm{risk} = \frac{\sum_m e_m F_m}{\sum_m e_m F_m + (1-e)(1-R)}, \tag{1}$$

and every class with $F_m = 1$ contributes its whole base rate e_m to the numerator. A gate that is strong on the error classes its designers anticipated and blind to the others releases, once the anticipated errors are filtered, mostly the errors nobody anticipated. Section 4.2 reports measured false-accept rates.

3.3 Why base rates matter

Proposition 3.3 (Dependence on the base rate)**.** *For fixed F and R with $0 < F < 1 - R$ (so that $F < 1$ and $R < 1$), the released error* risk *is strictly increasing in e. The share of withheld outputs that are in fact erroneous,*

$$\Pr(E \mid \mathit{REJECT}) = \frac{e(1-F)}{e(1-F) + (1-e)R},$$

is, when moreover $R > 0$, also strictly increasing in e, and tends to 0 as $e \to 0$.

Proof. risk $= 1/(1 + \frac{1-e}{e} \cdot \frac{1-R}{F})$ and $(1-e)/e$ is strictly decreasing on $(0,1)$. The expression for $\Pr(E \mid \mathsf{REJECT})$ follows from Bayes' rule as in proposition 3.2, and the same argument applies with $R/(1-F)$ in place of $(1-R)/F$. □

Worked example 3.2. Take a gate with $F = R = 0.05$. At a base error rate of 1% it releases outputs with error rate about 0.05%, and 84% of the outputs it withholds are correct. At a base error rate of 30% the same gate releases outputs with error rate about 2.2%, and 11% of the outputs it withholds are correct. The gate has not changed. What it means to a reader, and what it costs the reviewers who handle withheld outputs, changed by an order of magnitude.

Three consequences follow. First, a verifier's quality cannot be summarised without the base rate of the errors it faces; a benchmark that reports one accuracy figure has fixed a base rate by its own construction and reported a number about that construction. LLRB-1 fixes its mixture deliberately and reports rates per error class for that reason: its test split holds 1,396 supported items among 4,053, a proportion chosen by design and meaningless for any deployment. Second, the released error of a real deployment needs the base rate e_m of each error class in real generated answers, which LLRB-1 does not measure [PROPOSED]; per-class F_m and R transfer to a deployment only to the extent that the benchmark's constructed errors resemble the errors generators make. Third, the same arithmetic bounds the value of an alert: a screen that triggers on a share t of cases when the prevalence of true anomalies is p can be right on at most p/t of its alerts [ANALYTICAL].

The base rate is not a constant of nature either. Kalai and colleagues argue that evaluations which grade answers as simply right or wrong give an abstention the same score as an error, and so reward a model for guessing where it should abstain [18]. A generator trained and selected under such grading arrives at the gate with a higher e on exactly the questions it cannot answer. Scoring abstention separately from error is therefore a requirement on the evaluation of generators, not only on gates (section 5).

3.4 Calibration, and why a stage score is not a probability

Definition 3.2 (Calibration metrics)**.** For a confidence C and correctness $Y \in \{0,1\}$ on a population of outputs: the Brier score is $\mathbb{E}[(C-Y)^2]$ [16]; the expected calibration error over bins $B_1, \ldots, B_K$ of the confidence range is $\sum_k \frac{|B_k|}{n} |\mathrm{acc}(B_k) - \mathrm{conf}(B_k)|$ [17]; the area under the receiver operating characteristic, $\mathrm{AUROC} = \Pr(C^+ > C^-) + \frac{1}{2}\Pr(C^+ = C^-)$ with C^+ and C^- the confidences of a random correct and a random incorrect output, measures discrimination [19].

Calibration and discrimination are different properties. A calibrated confidence says how often outputs like this one are correct; a discriminating confidence separates the correct outputs from the others. A score can have either without the other.

Proposition 3.4 (Constant confidence)**.** *If $C \equiv c$ on a population in which a share ρ of outputs is correct, then* AUROC $= \frac{1}{2}$*, the expected calibration error is $|c - \rho|$, the Brier score is $\rho(1-c)^2 + (1-\rho)c^2$, and C is calibrated if and only if $c = \rho$.*

Proof. Every pair is tied, so $\Pr(C^+ = C^-) = 1$. All outputs fall in one bin, whose accuracy is ρ and whose mean confidence is c. The Brier score is the average of $(1-c)^2$ over correct and c^2 over incorrect outputs. The calibration condition has the single value c to check. □

A constant equal to the base rate is perfectly calibrated and useless. Calibration is necessary for a number to be honest; it is not sufficient for the number to be informative, so both must be reported.

Definition 3.3 (Stage score)**.** A verification pipeline runs a set J of stages. Stage j returns a value $v_j \in [0, 1]$, and the pipeline reports $S = \sum_{j \in J} w_j v_j / \sum_{j \in J} w_j$ with fixed weights $w_j > 0$.

Proposition 3.5 (Stage scores)**.** *(i) If each v_j depends only on the outcome of stage j, then S is constant on any set of outputs on which J and every stage outcome coincide, and its AUROC between correct and incorrect outputs in that set is $\frac{1}{2}$. (ii) If $k \in J$ and $v_k < S$, removing stage k from J strictly increases S. (iii) Consequently, if a stage that cannot run is left out of J rather than scored as a failure, an output for which less was verified can receive a higher score than the same output with more verified.*

Proof. (i) S is a function of J and the outcomes; the AUROC follows from proposition 3.4. (ii) Write $W = \sum_{j \in J} w_j$. After removing k, $S' = (WS - w_k v_k)/(W - w_k)$, and $S' - S = w_k(S - v_k)/(W - w_k) > 0$. (iii) Apply (ii) to the stage that did not run. □

A weighted stage score is therefore not a probability of correctness. It becomes one only when it is calibrated against labelled outcomes from the population in which it is shown, and the calibration must be redone when that population changes. Distribution-free methods help but do not remove this requirement: conformal prediction controls coverage in finite samples on the assumption that calibration and test data are exchangeable [20], and a change in the law breaks that assumption overnight. When the AI Act's application dates were moved in 2026 [EXTERNAL], every confidence calibrated on questions about those dates lost its warrant.

Nor is calibration of the generator a substitute for verification. Kalai and Vempala prove a lower bound: a language model that is calibrated in a natural generative sense hallucinates arbitrary facts, those whose truth cannot be inferred from its training data, at a rate close to the share of such facts seen exactly once in training [21]. The bound does not cover facts that appear more than once, a category in which the authors place references to publications. It is our inference, not theirs, that citations to rarely discussed authorities, docket numbers and the exact wording of rarely quoted provisions fall within it. Calibration makes a generator's error rate honest; it does not make it zero, and it does not say which output is wrong. Selective prediction states the right object to report for a gate: the risk–coverage curve, the released error as a function of the share of outputs released [14, 15], per error class.

3.5 Why similarity cannot decide support

Many grounding checks compare words. This subsection shows that their verdict on a decisive edit depends on the edit's length, position and vocabulary, not on its legal effect.

Definition 3.4 (Overlap measures)**.** Fix a tokeniser. Let φ have counted tokens $t_1, \ldots, t_n$ (a multiset) and let $T(s)$ be the set of tokens of $\mathsf{text}(s)$. *Containment* is $\mathrm{cont}(\varphi, s) = |\{i : t_i \in T(s)\}|/n$. For a window w of the provision, *token F1* is $2\,|\varphi \cap w|/(|\varphi| + |w|)$ with multiset intersection, and *Jaccard* similarity is $|\{\varphi\} \cap \{w\}|/|\{\varphi\} \cup \{w\}|$ on token sets. A verifier *accepts* when the measure reaches a threshold τ (for F1 and Jaccard, in the best window).

Proposition 3.6 (Blind spots of containment)**.** *Let* $\mathrm{cont}(\varphi, s) = 1$. *Then* $\mathrm{cont}(\varphi', s) = 1$ *for every* φ' *obtained from* φ *by (i) deleting counted tokens, keeping at least one [ANALYTICAL]; (ii) changing only tokens that the tokeniser discards; (iii) inserting tokens that occur anywhere in* $\mathsf{text}(s)$. *Moreover (iv) at threshold* τ, φ' *is accepted when it replaces at most* $(1-\tau)n$ *counted tokens or inserts at most* $n(1-\tau)/\tau$ *counted tokens [ANALYTICAL].*

Proof. (i) Every remaining token lies in $T(s)$. (ii) The sequence of counted tokens is unchanged. (iii) Each inserted token lies in $T(s)$. (iv) After m replacements at least $n-m$ of n counted tokens are still found, so $\mathrm{cont} \geq (n-m)/n \geq \tau$ when $m \leq (1-\tau)n$; a replacement by a discarded token shortens the denominator and can only raise the measure. After k insertions the n original tokens are found among $n+k$, so $\mathrm{cont} \geq n/(n+k) \geq \tau$ when $k \leq n(1-\tau)/\tau$. □

Corollary 3.4 (Low thresholds accept every single edit)**.** *At any threshold* $\tau \leq \frac{1}{2}$, *every single-token substitution or insertion in a fully contained statement with at least two counted tokens is accepted, since* $(n-1)/n \geq \frac{1}{2}$ *and* $n/(n+1) \geq \frac{2}{3}$*; a statement keeps passing after up to a share* $1-\tau$ *of its counted tokens are replaced. Every statement of LLRB-1's support family has between 12 and 70 words [IMPLEMENTED], so a containment verifier with such a threshold accepts every single decisive edit whatever its legal effect, and every misattribution to a provision that contains a share* τ *of the statement's counted tokens.*

Raising the threshold does not repair containment, because (i)–(iii) hold at every threshold: deletions, short tokens and vocabulary already present in the provision are invisible to it. Symmetric measures see deletions, and fail elsewhere.

Proposition 3.7 (Symmetric overlap under one edit)**.** *Let* φ *coincide token for token with a window* w *of* n *tokens. Against* w, *token F1 equals* $2n/(2n+1)$ *after one inserted token,* $(n-1)/n$ *after one substituted token, and* $2(n-k)/(2n-k)$ *after* k *deleted tokens. When the tokens of* w *are distinct and inserted or substituted tokens are new to* w, *Jaccard similarity equals* $n/(n+1)$, $(n-1)/(n+1)$ *and* $(n-k)/n$ *in the same three cases.*

Proof. For F1 the intersection of multisets has size n, $n-1$ and $n-k$, and the statement has $n+1$, n and $n-k$ tokens, against n for the window; these counts do not depend on whether the edited tokens occur elsewhere in w. Substitute in $2|\varphi' \cap w|/(|\varphi'| + |w|)$. For Jaccard, under the stated conditions, the intersection and union of the token sets have sizes $(n, n+1)$, $(n-1, n+1)$ and $(n-k, n)$. □

Corollary 3.5 (Length decides, not law)**.** *At the token-F1 threshold calibrated in LLRB-1,* $\theta = 0.971429$, *just above* $34/35$ *[MEASURED], a one-token insertion is accepted against the statement's own window once* $2n/(2n+1) \geq \theta$, *that is once* $n \geq 18$*; a one-token substitution once* $(n-1)/n \geq \theta$, *that is once* $n \geq 36$*; and a deletion of* k *tokens only once* $n > 18k$. *Inserting* not *into a long sentence passes; replacing* shall *by* may *passes only in very long ones; deleting a three-word exception clause fails in any sentence shorter than 55 tokens.*

The measured scores take the best window, which can only raise them, so the first two conditions are sufficient for acceptance. A post hoc check on LLRB-1's test split, not part of the registered analysis, finds the prediction holding as a boundary: token F1 accepted no negation edit of a statement shorter than 18 tokens and no modality edit of one shorter than 36, and accepted 481 of 484 negation edits and 242 of 245 modality edits at or above those lengths; each of the six exceptions follows a point label in its provision, which adds a token to its window [MEASURED]. The corollary explains the measured pattern of table 3: symmetric measures accept most negation edits, fewer than half of the modality edits and almost no deleted exceptions; containment does the reverse. In both cases the verdict is a function of how many tokens changed and where, which bears no relation to whether the change reverses an obligation. McCoy and colleagues showed the same failure in natural language inference, where models that rely on lexical

overlap accept hypotheses that share the premise's words without following from it [22]. In legal text the failure is sharper, because the words that decide the law are short, frequent and repeated: *not*, *may*, *shall*, a number, a point letter, the name of the other party.

3.6 Exact and canonical matching

Proposition 3.8 (Exact and canonical quotation checks)**.** *(i) A check that accepts a quotation only when its bytes equal a receipted span of the official text accepts no altered quotation and rejects every typographic variant. (ii) A canonical substring check under κ accepts every variant that differs from a substring of the text only within the character classes κ folds, and rejects every edit that changes a character outside them unless the edited string occurs elsewhere in the text* [ANALYTICAL]*. (iii) Neither accepts a paraphrase.*

Proof. (i) Byte equality, or equality of cryptographic hashes, admits only the identical string. (ii) κ maps the variant and the original to the same string; an edit outside the folded classes survives κ, so the canonical edited string differs from the canonical original. (iii) A paraphrase is not a substring of the text, before or after κ. □

Canonicalisation buys tolerance of typography at the price of blindness inside the folded classes, and the price must be stated. Compatibility normalisation maps a superscript digit to the ordinary digit, so "10^2" and "102" share a canonical form; case folding maps a capitalised defined term to the common word, so "the Board" and "the board" coincide; neither folds look-alike letters across scripts [ANALYTICAL]. A canonicalisation for legal text must therefore be declared, versioned and chosen so that what it folds carries no legal meaning in the corpus it is applied to. Part (iii) is the hard limit: a check that only accepts quotations rejects every correct paraphrase, and a statement of law in a memorandum is usually a paraphrase.

4 Evidence from a pre-registered benchmark

4.1 Design in brief

LLRB-1 measures verification components, not language models. Its corpus is 3,028 provisions of eight EU acts retrieved from EUR-Lex on 26 November 2026, with per-act article counts equal to the acts' own [MEASURED]. Its support family pairs a statement with the provision it cites: 5,820 items built from 1,007 statements, 4,053 of them in the test split, with every label fixed by construction and none assigned by a model [MEASURED]. A statement is a sentence or clause of 12 to 70 words taken from a provision; each yields its verbatim form, at most one typographic variant, one edit for each family of decisive edit that applies to it, and one misattribution to the most lexically similar provision of another article of the same act [IMPLEMENTED]. Items are split by article, so that no article contributes to both the development and the test split [IMPLEMENTED]. Further families test citation resolution, temporal applicability, jurisdictional scope and release discipline [MEASURED].

Lawie's components were imported unchanged from the code basis and run offline; each run records the commit and the SHA-256 of every exercised source file and refuses to run if one differs [MEASURED]. They were compared with lexical baselines whose thresholds were calibrated on the development split, and with a typed reference prototype that is a research instrument of this suite, not a Lawie component [PROPOSED]. The hypotheses, thresholds and test-split checksums were committed before the confirmatory run, which ran once [MEASURED]. No human legal expert has reviewed the labels [ABSENT]. The companion benchmark report (LL-BEN-01) gives the full method, the statistics and the threats to validity; this section reports what the measurements mean for the properties of section 2.

4.2 Similarity is not support

Table 2: Confirmatory hypotheses of LLRB-1 on the test split. Intervals are exact (Clopper–Pearson); p-values are from exact binomial and McNemar tests, Holm-corrected within H1 and within H2.

Hypothesis and quantity	Count	95% CI (%)	Test	p (Holm)	Result
H1 decisive edits accepted by Lawie grounding	1,959/1,959	99.8–100.0	binomial > 0.5	$< 10^{-6}$	supported
H1 misattributions accepted by Lawie grounding	698/698	99.5–100.0	binomial > 0.5	$< 10^{-6}$	supported
H2 decisive edits accepted, calibrated containment	531/1,959	25.1–29.1	binomial > 0.05	$< 10^{-6}$	supported
H2 decisive edits accepted, calibrated token F1	1,062/1,959	52.0–56.4	binomial > 0.05	$< 10^{-6}$	supported
H2 decisive edits accepted, calibrated Jaccard	1,015/1,959	49.6–54.0	binomial > 0.05	$< 10^{-6}$	supported
H3a modelled edits accepted, typed prototype vs. containment	0 vs. 392 of 1,444	FRR 0% (0.00–0.26)	McNemar; binomial < 0.05	$< 10^{-6}$	supported
H3b held-out edits accepted, typed prototype	456/515	85.5–91.2	binomial > 0.5	$< 10^{-6}$	supported

Four registered hypotheses were tested once on the test split (table 2). All four were supported.

Lawie's grounding check accepted every decisive edit. Lawie's grounding check, as implemented at the code basis, accepted 1,959 of 1,959 statements carrying one legally decisive edit and 698 of 698 statements cited to the wrong provision [MEASURED]. To check the harness, its decisions were reproduced independently on all 4,053 test items [MEASURED]. In the setting measured, the cited provision was the only retrieved source. For these error classes the check's false-accept rate is one, and by corollary 3.2 it passes the base error rate of decisive edits and misattributions through unchanged.

Calibration does not repair overlap. Calibrated containment, token F1 and Jaccard, each thresholded at the 5th percentile of supported development scores (token F1 kept 94.8% of them [MEASURED]), still accepted 27.1% (95% CI 25.1–29.1), 54.2% (52.0–56.4) and 51.8% (49.6–54.0) of decisively edited statements on the test split [MEASURED]. The best of them divides the base error rate of decisive edits by less than four (corollary 3.3). Table 3 gives the rates per family; their pattern is the one propositions 3.6 and 3.7 predict. Calibrated containment accepted every deleted exception and 96% of cross-reference edits, which change short tokens, and rejected nearly every date and time-unit edit, which introduce new words. Token F1 and Jaccard accepted 89% and 90% of negation edits, each a single inserted or deleted *not*, against 45% and 36% of modality edits, which substitute one word, and 0% and 7% of deleted exceptions. No measure's verdict followed the legal effect of the edit.

Typing the decisive elements works where it types. The reference prototype requires that a window of up to four clauses of the cited provision contain at least 90% of the statement's words and that the multiset of decisive elements it models (deontic operator with polarity, quantity, date, time unit, cross-reference) be equal in statement and window. It accepted 0 of 1,444 edits of the types it models and rejected 0 of 1,396 supported statements, with a 95% upper bound of 0.26% on the false-reject rate [MEASURED]. It accepted 456 of 515 edits of the types it does not model (88.5%, 95% CI 85.5–91.2) [MEASURED]: a swapped actor, a reversed comparator, a deleted exception. The first result is less than it seems, because the prototype types exactly the element classes that the modelled edits alter, and both were defined by the same authors [ANALYTICAL]. The second result is the informative one. Typing protects the element types it types and no

Table 3: Acceptance (%) by system and item family, LLRB-1 test split. The first block is supported (higher is better); the others are unsupported (lower is better). Modelled families (second block) are the decisive-edit types the typed prototype checks; held-out families (third block) are types it does not model. [a]Quote matcher given the provision text. [b]Exact span with the one admitted derivation.

Family	n	Lawie grounding	Lawie quote matcher[a]	Lawie exact span[b]	Canonical match	Containment (cal.)	Token F1 (cal.)	Jaccard (cal.)	Typed prototype
Verbatim	698	100	100	100	100	100	96	96	100
Spacing	242	100	100	0	100	100	97	99	100
Initial capital	347	100	100	0	100	100	100	100	100
Final punctuation	93	100	100	0	100	100	76	75	100
Quotation marks	16	100	100	0	100	100	94	94	100
Negation	538	100	100	0	0	21	89	90	0
Modality	538	100	100	0	0	20	45	36	0
Quantity	131	100	100	0	0	45	51	45	0
Date	41	100	100	0	0	0	20	17	0
Time unit	85	100	100	0	0	4	62	54	0
Cross-reference	111	100	100	0	0	96	54	48	0
Actor	355	100	99	0	0	25	23	31	93
Comparative/quantifier	118	100	100	0	0	7	59	49	83
Deleted exception	42	100	100	0	0	100	0	7	64
Misattributed	698	100	52	0	0	8	1	2	7

others. Each decisive element of law, actors and exceptions included, has to be modelled and tested before a verifier may be trusted with it.

Misattribution survives typing. The prototype also accepted 7% of misattributed statements (table 3). It is anchored: it compares the statement only with the provision cited. But EU acts repeat formulae from article to article, and its acceptances are near-identical clauses of the wrong article, some repeating the statement's words and others differing in a word or two that it does not type. Fuzzy comparison at 90% containment admits them. Only exact and canonical matching against the cited provision rejected all 698 (section 4.4).

4.3 Confidence carries no information

The verification pipeline reported a confidence of 0.973 with the label HIGH for every supported and every unsupported statement of the test split: AUROC 0.5 and expected calibration error 0.63 [MEASURED]. A score that does not vary between correct and incorrect outputs cannot discriminate between them, whatever its form (proposition 3.5(i)). The calibration error is the distance between 0.973 and the share of supported items in the test split, a property of the benchmark's mixture and not of any deployment [ANALYTICAL]; the absence of discrimination does not depend on the mixture.

No component of the system produces a calibrated probability of correctness [ABSENT], and the number shown is not one.

4.4 Quotation fidelity: blind, useless, brittle, exact

Four quotation checks were measured, and they fail in four different ways.

Blind. Lawie's quoted-holding matcher, given the provision's text as its reference, accepted 1,956 of 1,959 decisively edited quotations and 51.6% of misattributed ones [MEASURED]. Given the right text, it still

could not tell a quotation from a decisively edited one; proposition 3.6(iv) shows, in another form, how any fuzzy comparison with a threshold below one can fail this way.

Useless. Given only the citation's metadata instead of the provision text, the same matcher accepted 1.0% of authentic quotations. It rejects true quotations because it is never shown the text they quote.

Brittle. Lawie's citation-authority predicate treats a statute version as authoritative only when its text is an exact byte span of an official-source observation retrieved under an approved source contract, still current and in force at the citation's date, and never takes model output as an input [TESTED]. It accepted every authentic quotation and rejected every typographic variant and every edit; with the span derivation that EU texts would need, it rejected everything [MEASURED], because exactly one derivation, for Czech statute paragraphs, is admitted [IMPLEMENTED]. Exactness here is correct and brittle: a quotation re-typed with a straight apostrophe is as unverifiable as a forged one, and no EU text can be verified at all.

Exact. Canonical substring matching accepted every authentic quotation and typographic variant and rejected every edit and misattribution [MEASURED]. The result holds by construction for the edit families used (proposition 3.8(ii)): the benchmark discarded edited statements that occur in their provision after canonicalisation, and every edit changes a character that the canonicalisation does not fold. It is the right check for quotations and says nothing about paraphrase (proposition 3.8(iii)).

The lesson is about the comparison as much as the anchor. Given the cited provision's text, the matcher is anchored and still fails, because its comparison is fuzzy; given only the citation's metadata, it is not anchored and fails the other way. Only exact-span and canonical matching against the cited provision rejected all 698 misattributions.

4.5 Citation resolution

Resolution failed on the form of citation an EU lawyer is most likely to write. Lawie's citation extractor resolved the instrument for none of the citations written in the official style "Regulation (EU) 2016/679" (0 of 114), for all 49 written without "(EU)", for 37.7% of citations by short name, and found no article in citations abbreviated "Art." or written in the Czech format [MEASURED]. Property P2 requires format invariance; this extractor is format-sensitive in a way that inverts the expected ordering, resolving the informal rendering and failing the official one.

4.6 Time and jurisdiction

Table 4: Temporal errors on the LLRB-1 test split by the true state of the provision on the date asked. An error is a date on which the component treated the provision as applicable when it was not, or the reverse.

True state on the date	n	Index (entry into force)	Authority in-force check	Index (publication)
Not yet in force	1,298	0 (0%)	0 (0%)	791 (61%)
In force, not yet applicable	781	781 (100%)	781 (100%)	781 (100%)
Applicable	456	0 (0%)	0 (0%)	0 (0%)

Lawie's temporal index and the in-force check of its citation-authority predicate reported a provision as in force on 781 of 781 test dates on which it was in force but not yet applicable, and were correct on every other date; a variant keyed on publication also erred on 791 of 1,298 dates before entry into force (table 4) [MEASURED]. The error is structural. The parser starts a version at entry into force and leaves its end open [IMPLEMENTED]; staggered dates of application, transitional provisions and consolidated versions are not modelled [ABSENT]; and an index keyed on entry into force answers "applicable" wrongly on every day between entry into force and application [ANALYTICAL]. The exposure is not small. Over 2016–2028,

treating entry into force as applicability contradicts the official schedules on 7–20% of provision-days, depending on the act; ignoring provision-specific dates contradicts them on up to 6%; and an AI Act schedule frozen before Regulation (EU) 2026/1744 errs on 2.0% of provision-days for systems in Annex III [MEASURED]. The Cyber Resilience Act shows why per-provision dates matter. Its final provision reads: "This Regulation shall apply from 11 December 2027. However, Article 14 shall apply from 11 September 2026 and Chapter IV (Articles 35 to 51) shall apply from 11 June 2026." [23] A single date per act is wrong for this act on any reading.

Jurisdictional scope failed by omission. The gateway's scope rule treats Czech, Slovak and EU law as one compatible bloc and recognises foreign primary-law markers only for a short list of legal orders [IMPLEMENTED]. In answers for a Czech matter it flagged citations of US law (8 of 8) and never flagged citations of Slovak, German, Austrian, UK or Polish law (0 of 32) [MEASURED]. A rule that blocks the foreign orders it knows and admits the rest fails exactly where the risk is highest: neighbouring legal orders whose statutes look like domestic ones. Retrieval, by contrast, routes jurisdiction along a declared chain and drops untagged documents from jurisdiction-routed results [TESTED], which is the admission principle of property P5 applied at the right place.

4.7 Release discipline

One release control was measured. The gateway's citation-required rule blocked every legal conclusion without a citation (4 of 4) and none with a fabricated one (0 of 8) [MEASURED]. A rule that behaves this way rewards fabrication: an answer that invents a citation passes where an honest answer without one is stopped.

Some of Lawie's controls withhold correctly by construction. A research answer with any citation not backed by a retrieved source is withheld, blocks reliance and requires human review [TESTED]. Agent outputs pass a citation gate that fails the run when a required factuality or citation-accuracy metric is missing or failing [TESTED]. A good-law check on case law cannot pass, because the validity service returns no determination and the gate requires a positive one [IMPLEMENTED]; the control fails closed, and as a result case law cannot currently be relied on through it at all.

5 Requirements for reliable legal AI

The analysis and the measurements support a short list of requirements. Each is stated as a property that an independent evaluator can test, with the parameters the deploying institution must choose made explicit. None is a threshold this paper sets on its own authority: how much error a firm, a court or a regulator accepts is their decision. What can be fixed is the form of the test, so that the decision is about a measured quantity. Table 5 summarises the requirements and the benchmark family that measures each; the text gives the test and its cost. The requirements are stated for any system; the measurements of section 4 show how far a system can be from them.

R1 Typed support. For every decisive element type the system claims to check, the evaluator builds single-edit contrast items from the official text and reports the false-accept rate per type with an exact upper bound, together with the false-reject rate on verbatim and typographically varied statements. Pooling across types is not permitted, because a pooled rate hides the blind types (equation (1)). The research prototype of section 4.2 meets it on the modelled types for any ε above their per-type upper bounds (largest 8.6%, dates) [MEASURED] and fails it on the held-out types [MEASURED]. *Cost:* each element type is engineering work, the list of types is open, and a verifier is trustworthy only on the types listed.

R2 Anchored support. The verifier compares the assertion only with the text of the provision its citation names, at the version in force on the date asked. Similarity to other provisions is irrelevant. R2 is what

Table 5: Requirements as testable properties, and where LLRB-1 measures each.

	Requirement	Testable property	Measured in LLRB-1
R1	Typed support	Per decisive element type, the 95% upper bound on the false-accept rate over single-edit contrast items is below ε, at a false-reject rate below δ	Support family, decisive edits (table 3)
R2	Anchored support	Each assertion is compared with the text of the address its citation names, at the version in force on t_q, and with nothing else	Support family, misattributions
R3	Format-invariant resolution	Every official and customary rendering of an address resolves to that address; non-existent addresses resolve to nothing	Citation family
R4	Declared canonical quotation	Accept a quotation iff it is a canonical substring of the official text under a declared, versioned κ	Support family, typographic variants and edits
R5	Applicability by date	Zero errors on the day before and the day of every transition, under the schedule known at t_k	Temporal family (table 4)
R6	Jurisdiction by admission	Any citation not resolved to an instrument admitted for J is flagged	Jurisdiction family
R7	Fail-closed release	No output is released with a support stage that did not run or failed; presence of a citation is never evidence	Release family
R8	Honest confidence	Either a probability calibrated on labelled outcomes from the deployment, reported with ECE, AUROC and intervals, or no number	Confidence analysis
R9	Scored abstention	Risk–coverage per error class; abstention scored apart from error	Deferred generated-answer study

defeats misattribution, which typing alone did not (section 4.2). *Cost:* the system must resolve citations correctly (R3) and hold versioned official text, so a citation it cannot resolve leaves the assertion unverified.

R3 Format-invariant resolution. The evaluator renders each address in every official and customary format, including the Official Journal's own, and fabricated addresses in the same formats. *Cost:* formats differ by legal order and language, and each must be enumerated and tested.

R4 Declared canonical quotation. A quotation is accepted if and only if its canonical form is a substring of the canonical official text, and the canonicalisation is published with the list of character classes it folds. *Cost:* the check accepts no paraphrase, so paraphrased statements must pass R1 and R2 instead; and what κ folds is invisible to it (section 3.6).

R5 Applicability by date. The evaluator asks about every provision on the day before and the day of publication, entry into force and application, per category of addressee where the act distinguishes categories, and under each version of the schedule. The pass criterion is zero errors on these boundary dates, because an error there is not noise but a wrong model of time. *Cost:* schedules must be transcribed per provision, with the knowledge date of each version, and provisions whose application depends on facts must be refused rather than dated.

R6 Jurisdiction by admission. The check admits instruments for a legal order and flags everything else, instead of blocking the foreign orders it recognises. *Cost:* false flags on legitimately applicable foreign law, which only a human with the conflict rules in view can clear.

R7 Fail-closed release. An output is released only if every support stage ran and passed. A stage that could not run blocks release, and the presence of a citation string is never evidence of support. *Cost:* lower coverage and a queue of withheld outputs whose composition depends on the base rate (proposition 3.3).

R8 Honest confidence. A confidence is shown only if it has been calibrated on labelled outcomes from the population it is shown to, and its calibration error and discrimination are published with intervals. Otherwise the system shows no number, and says which checks passed. No Lawie component produces a calibrated probability [ABSENT]. *Cost:* labelled outcomes per deployment, and recalibration whenever the law or the population changes.

R9 Scored abstention. Evaluations report the risk–coverage curve per error class and score an abstention separately from an error, so that a system is not rewarded for guessing [18]. The generated-answer study that would measure this in LLRB-1 is defined and not run [PROPOSED]. *Cost:* a system that abstains honestly looks worse on accuracy-only leaderboards, and the evaluation must explain why that is the better system.

These requirements do not make a system correct about the law. They make its errors measurable, attributable to a property, and bounded at release. That is what a lawyer needs before relying on a machine: not that it is intelligent, but that what it releases has been checked against the source in a way the lawyer could check again.

6 What was not measured

The measurements of section 4 are narrow by design, and the narrowness bounds what they show. A reader should weigh this section as heavily as the results.

No generated answer was evaluated. LLRB-1 measures verification components on constructed statements. The study of model-generated answers, which would measure how often generated legal answers fabricate citations, misquote, err about time or fail to abstain, and how much a verification layer lowers the released error, has a defined protocol and has not been run [PROPOSED]. The base rates e_m of equation (1) are therefore unknown, and no statement in this paper says how often a Lawie answer is wrong. This programme ran no model.

No lawyer has reviewed the labels. Labels are fixed by construction from the official text: the provision's own words are supported, and one decisive edit makes a statement unsupported [MEASURED]. Construction makes labels mechanical, not infallible. An edit can produce a statement that another sentence of the provision supports, which the construction screens only for edits that occur verbatim, after canonicalisation, in the provision [IMPLEMENTED], and an edit can produce a sentence no drafter would write. No human legal expert has reviewed the labels [ABSENT].

Supported items are quotations. Every supported item is verbatim or a typographic variant. The benchmark does not test whether a verifier accepts a correct paraphrase, so a verifier that accepts only quotations scores perfectly here and would reject every legitimate paraphrase in use (proposition 3.8(iii)). The false-reject rates reported are lower bounds on the false-reject rates a paraphrasing generator would meet.

One edit, one language, one kind of law. Each unsupported item carries exactly one edit, and generated errors can carry several. The corpus is English-language EU legislation; Czech and other national law, case law, contracts and other languages are not covered. The temporal ground truth excludes provisions whose application depends on facts, on interpretation or on sub-paragraph granularity [IMPLEMENTED].

Components, not services. Lawie's components were run in isolation with inputs constructed from the official text, not through a deployed service. The grounding check was given the cited provision as its only retrieved source. Results describe the code basis only.

The system's state. Nothing described here is deployed in production [BLOCKED]; no Lawie product has a dated production acceptance [ABSENT]; and no independent audit, certification or human expert review of the components is recorded [ABSENT]. The typed prototype that meets requirement R1 on the modelled types,

for any ε above their per-type upper bounds, is a research instrument of this suite, not a Lawie component [PROPOSED]. This paper has not been peer reviewed.

7 Conclusion

The reliability of a legal output is not one quantity. It is support, resolution, quotation, time, jurisdiction, abstention and calibration, and each fails for its own reasons. Measured separately, they show where a system can be trusted and where it cannot; averaged, they hide both.

Three analytical facts organise the rest. Reliability compounds downward with the number of assertions, so a per-claim error rate that sounds small is not small for a memorandum. A gate's effect on what it releases is fixed by its false-accept rate, its false-reject rate and the base rate of errors, so a gate that accepts every error of a class does nothing for that class whatever its accuracy elsewhere. And a confidence assembled from stage results is not a probability: it can be constant on right and wrong answers alike, and nothing ties it to the evidence.

The measurements bear these facts out on Lawie's own components. Its grounding check, as implemented at the code basis, accepted every decisive edit and every misattribution; calibrated similarity measures accepted 27–54% of decisive edits, in a pattern set by the length and position of the edit rather than its legal effect; its reported confidence carried no information; its temporal index equated entry into force with application; its jurisdiction rule admitted the foreign law it did not know; and a citation-required rule passed fabricated citations. A typed prototype removed the errors it was built to see and few of the others. Canonical matching against the cited provision verified quotations exactly and cannot verify a paraphrase.

None of this measures how often a generated legal answer is wrong; that study is defined and not run. What it does establish is the form a reliable legal system must take: support decided on the decisive elements of the provision a citation names, quotations checked against the official text under a declared canonical form, applicability computed from each provision's own dates, jurisdiction by admission, release that fails closed, confidence that is calibrated or absent, and abstention that is scored rather than punished. Each of these is a property an evaluator can test. A machine that meets them still has no authority. It has something more useful to a lawyer: outputs whose warrant can be checked.

References

[1] United States District Court for the Southern District of New York. Mata v. Avianca, Inc., No. 22-cv-1461 (PKC), Opinion and Order on Sanctions, 2023. URL https://www.courtlistener.com/docket/63107798/mata-v-avianca-inc/. 22 June 2023.

[2] American Bar Association Standing Committee on Ethics and Professional Responsibility. Formal Opinion 512: Generative Artificial Intelligence Tools, 2024. URL https://www.americanbar.org/content/dam/aba/administrative/professional_responsibility/ethics-opinions/aba-formal-opinion-512.pdf. 29 July 2024.

[3] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis*, 16(1):64–93, 2024. doi: 10.1093/jla/laae003.

[4] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *Journal of Empirical Legal Studies*, 2025. doi: 10.1111/jels.12413.

[5] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):248:1–248:38, 2023. doi: 10.1145/3571730.

[6] H. L. A. Hart. *The Concept of Law*. Clarendon Press, Oxford, 1961.

[7] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical Report NIST AI 100-1, U.S. Department of Commerce, 2023.

[8] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, 2023. doi: 10.18653/v1/2023.emnlp-main.741.

[9] Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring Attribution in Natural Language Generation Models. *Computational Linguistics*, 49(4):777–840, 2023. doi: 10.1162/coli_a_00486.

[10] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, 2023. doi: 10.18653/v1/2023.emnlp-main.398.

[11] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, 2024. doi: 10.18653/v1/2024.eacl-demo.16.

[12] Christian S. Jensen, Curtis E. Dyreson, et al. The Consensus Glossary of Temporal Database Concepts — February 1998 Version. In *Temporal Databases: Research and Practice*, volume 1399 of *Lecture Notes in Computer Science*, pages 367–405. Springer, 1998. doi: 10.1007/BFb0053710.

[13] Monica Palmirani, Guido Governatori, and Giuseppe Contissa. Modelling Temporal Legal Rules. In *Proceedings of the 13th International Conference on Artificial Intelligence and Law*, pages 131–135, 2011. doi: 10.1145/2018358.2018378.

[14] C. K. Chow. On Optimum Recognition Error and Reject Tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

[15] Yonatan Geifman and Ran El-Yaniv. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems 30*, pages 4878–4887, 2017.

[16] Glenn W. Brier. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(1): 1–3, 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330, 2017.

[18] Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why Language Models Hallucinate, 2025. arXiv:2509.04664.

[19] James A. Hanley and Barbara J. McNeil. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747.

[20] Anastasios N. Angelopoulos and Stephen Bates. Conformal Prediction: A Gentle Introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

[21] Adam Tauman Kalai and Santosh S. Vempala. Calibrated Language Models Must Hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 160–171, 2024. doi: 10.1145/3618260.3649777.

[22] R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, 2019. doi: 10.18653/v1/P19-1334.

[23] European Parliament and Council of the European Union. Regulation (EU) 2024/2847 of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements (Cyber Resilience Act). Official Journal of the European Union, L, 2024/2847, 20.11.2024, 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32024R2847.