Legal AI Grounding and Hallucination Benchmark

LLRB-1: a pre-registered benchmark of legal verification components on the official text of eight EU acts

Benchmark report LL-BEN-01
Version 1.0.0, 26 November 2026

Abstract

Despite its registered title, LLRB-1 measures detection by verification components, not how often models hallucinate. A verification layer compares a model's answer with the sources it cites and decides whether to release it. LLRB-1 asks whether such components detect the errors that change legal meaning: a statement that alters one decisive element of the provision it cites, a correct statement cited to the wrong provision, a citation in the Official Journal's own format, a provision that is in force but not yet applicable, a foreign statute in a domestic answer, and a conclusion that carries a fabricated citation.

The benchmark is built from the official text of eight EU acts retrieved from EUR-Lex, 3,028 provisions in all [MEASURED]. Its labels are fixed by construction from that text and none is assigned by a model [MEASURED]. It has five families: support (5,820 items), citation resolution, temporal applicability, jurisdictional scope and release discipline; 4,053 support items form the test split. Its hypotheses, thresholds and the checksums of every data file were committed before the confirmatory run, which ran once [MEASURED]. The systems are Lawie's components as implemented at the code basis, imported unchanged and run offline, five lexical baselines and a typed reference prototype that is a research instrument, not a product [PROPOSED].

All four registered hypotheses were supported. Lawie's grounding check accepted 1,959 of 1,959 decisively edited statements and 698 of 698 misattributed ones [MEASURED]. Containment, token F1 and Jaccard baselines, each thresholded at the 5th percentile of supported development scores, accepted 27.1%, 54.2% and 51.8% of decisive edits [MEASURED]. The typed prototype accepted none of 1,444 modelled edits, at no measured cost, and 88.5% of 515 unmodelled ones [MEASURED]. The pipeline's reported confidence was constant (AUROC 0.5) [MEASURED]. Lawie's citation extractor resolved the instrument of no citation in the official "Regulation (EU)" format [MEASURED]; its temporal components treated every in-force but not yet applicable provision as applicable [MEASURED]; its jurisdiction rule flagged US law and never Slovak, German, Austrian, UK or Polish law [MEASURED]; and its citation-required rule passed every fabricated citation [MEASURED]. Limits: supported items are quotations, edits are single, the corpus is English EU legislation, no lawyer has reviewed the labels [ABSENT], and no generated answer was studied [PROPOSED].

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:* AI and NLP researchers, legal technologists, evaluation and assurance teams, 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.

Contents

Abstract **1**

1 Purpose and scope **3**

1.1 Why benchmark verifiers 3

1.2 Research questions 3

1.3 Scope 3

2 Corpus and provenance **3**

2.1 The acts 3

2.2 Schedules of application 4

3 Item families and their construction **5**

3.1 Support (SUP) 5

3.2 Citation resolution (CIT) 5

3.3 Temporal applicability (TMP) 7

3.4 Jurisdictional scope (JUR) 7

3.5 Release discipline (GRD) 7

4 Labels, safeguards and splits **7**

4.1 Labels by construction 7

4.2 Validity safeguards 7

4.3 Splits 8

5 Systems under test **8**

5.1 Lawie components 8

5.2 Baselines 9

5.3 The reference prototype 9

6 Metrics, statistics and pre-registration **9**

6.1 Metrics 9

6.2 Statistics 10

6.3 Pre-registration 10

7 Results **11**

7.1 Confirmatory hypotheses 11

7.2 Support by family 12

7.3 Citations and time 13

7.4 Jurisdiction and release 14

8 Analysis by family **14**

8.1 Three mechanisms 14

8.2 Family by family 15

9 Threats to validity **16**

9.1 Construct validity 16

9.2 Label validity 17

9.3 Realism, design and sampling 17

9.4 External and statistical validity, and conflict of interest 17

10 Datasheet **18**

11 Relation to prior benchmarks **19**

12 The deferred generated-answer study, and reproduction **20**

References **21**

1 Purpose and scope

1.1 Why benchmark verifiers

Studies of legal hallucination measure generators: how often a model or a research tool produces a wrong or misgrounded answer [1, 2]. That rate is the base error rate e that a verification layer faces. What reaches a reader is different. If a gate releases outputs, the error rate among released outputs is $eF/(eF + (1-e)(1-R))$, where F is the gate's false-accept rate and R its false-reject rate [ANALYTICAL]. A generator benchmark measures e. A verifier benchmark measures F and R, and must measure them per error class, because a gate that is strong on one class and blind to another releases the second class untouched. The companion paper *The Reliability of Legal Artificial Intelligence* (LL-REL-01) derives these relations; this report documents the benchmark that measures the second half.

LLRB-1 is therefore a benchmark of *components*, not of language models. It asks of each verification component the question a lawyer would ask of a junior's citation check: does it notice when the statement and the cited provision differ in a way that changes the law?

1.2 Research questions

RQ1 Support. Does a component that compares a statement with its cited provision accept the provision's own words and their typographic variants, and reject a statement that changes one decisive element, or that is correct but cited to the wrong provision?

RQ2 Resolution. Does a citation extractor resolve the instrument and the article in each format in which EU citations are written?

RQ3 Time. Do temporal components distinguish a provision that is in force from one that applies?

RQ4 Jurisdiction. Does a scope rule flag foreign primary law cited in an answer for a domestic matter?

RQ5 Release. Do release rules withhold conclusions that carry fabricated citations?

1.3 Scope

The benchmark covers English-language EU legislation, eight acts. It does not cover case law, national law, contracts or other languages; it does not test paraphrase; and it does not evaluate model-generated answers, whose study has a defined protocol and has not been run [PROPOSED] (section 12). Every result describes the code basis named on the first page, run offline; nothing here is a production or customer measurement, and no Lawie component described is deployed in production [BLOCKED].

2 Corpus and provenance

2.1 The acts

The corpus is eight EU acts in their English versions, retrieved from EUR-Lex on 26 November 2026: the General Data Protection Regulation [3], the Digital Operational Resilience Act [4], the NIS 2 Directive [5], the Markets in Crypto-Assets Regulation [6], the Digital Services Act [7], the Data Act [8], the Cyber Resilience Act [9] and the Artificial Intelligence Act [10]. The consolidated text of the AI Act of 27 July 2026 [11] and the amending Regulation (EU) 2026/1744 [12] were retrieved with them to fix the AI Act's current schedule of dates. The acts were chosen because they are recent and long, and because several of them apply different provisions from different dates, which makes them hard cases for temporal reasoning.

A parser splits each act into provisions, one per paragraph of each article (an article without numbered paragraphs is one provision). It yields 3,028 provisions, and the number of articles it finds equals each act's own count [MEASURED] (table 1). The SHA-256 digest of every retrieved rendering is recorded in a manifest, so that a later retrieval can be compared byte for byte.

Table 1: The LLRB-1 corpus. Dates are from each act's final provisions; "Applies (general)" is the general date of application, from which individual provisions may depart.

Act	CELEX	Published	In force	Applies (general)	Articles	Provisions
GDPR	32016R0679	2016-05-04	2016-05-24	2018-05-25	99	389
DORA	32022R2554	2022-12-27	2023-01-16	2025-01-17	64	267
NIS2	32022L2555	2022-12-27	2023-01-16	2024-10-18	46	199
MiCA	32023R1114	2023-06-09	2023-06-29	2024-12-30	149	751
DSA	32022R2065	2022-10-27	2022-11-16	2024-02-17	93	363
Data Act	32023R2854	2023-12-22	2024-01-11	2025-09-12	50	235
CRA	32024R2847	2024-11-20	2024-12-10	2027-12-11	71	305
AI Act	32024R1689	2024-07-12	2024-08-01	2026-08-02	113	519

Only the Official Journal of the European Union is authentic. The corpus is a research copy made under the Commission's reuse decision [13], and nothing in the benchmark treats it as anything else.

2.2 Schedules of application

Temporal labels need more than the dates in table 1. Every act in the corpus separates publication, entry into force and application, and several apply different provisions from different dates [EXTERNAL]. The benchmark transcribes each act's final provisions into a schedule: for every rule of application, the date, the provisions it covers and a verbatim quote of the sentence it comes from. All 26 quotes were found in the official text by an automatic check [MEASURED], so every date in the ground truth can be traced to the words that set it.

The AI Act is the hardest case. As published, it applied from 2 August 2026, with Chapters I and II from 2 February 2025, the governance, general-purpose model and penalty provisions from 2 August 2025, and Article 6(1) with its corresponding obligations from 2 August 2027 [EXTERNAL]. Regulation (EU) 2026/1744, in force since 27 July 2026, moved the application of Chapter III Sections 1–3 to 2 December 2027 for systems listed in Annex III and to 2 August 2028 for systems covered by Annex I, and set new dates for Articles 102–110 and for new prohibitions [EXTERNAL] (table 2). The schedule therefore carries versions, each with the date from which it was the law, and the temporal labels use the version in force on the knowledge date of the benchmark.

Table 2: Dates of application of the AI Act under Article 113 as published and as amended by Regulation (EU) 2026/1744. *The original date for Annex I systems rests on reading "the corresponding obligations" of the original Article 113(c); items that depend on it are excluded from date-only labels.

Provisions	As published (2024)	As amended by Reg. (EU) 2026/1744
Chapters I and II (definitions, prohibited practices)	2 February 2025	2 February 2025; new Art. 5 prohibitions 2 December 2026
Chapter III Section 4, Chapters V, VII, XII (except Art. 101), Art. 78	2 August 2025	2 August 2025
Articles 102–110 (amendments to other acts)	2 August 2026	27 July 2026
Chapter III Sections 1–3 (except Art. 6(5)), Annex III high-risk systems	2 August 2026	2 December 2027
Chapter III Sections 1–3 (except Art. 6(5)), Annex I high-risk systems	2 August 2027*	2 August 2028
All other provisions	2 August 2026	2 August 2026

3 Item families and their construction

3.1 Support (SUP)

A support item pairs a statement with the provision it cites, and asks whether the provision supports it. Statements are sentences or clauses of 12 to 70 words taken from the provisions of the corpus [IMPLEMENTED]. They are a seeded random sample per act, enriched with statements to which the rarer kinds of edit apply, so that every family has enough items; family sizes are therefore a design choice, and results are reported per family as well as in the registered pools; no pooled rate is a population rate [IMPLEMENTED]. The 1,007 statements yield 5,820 items [MEASURED]. From each statement the generator builds:

The verbatim statement, labelled supported.

At most one typographic variant, labelled supported: the first that applies of a no-break space replaced by an ordinary space, typographic quotation marks replaced by straight ones, the initial capital lowered, and the final punctuation removed. Each changes the bytes and none changes the law.

One decisive edit per applicable family, labelled unsupported. Each edit changes exactly one element of the statement and changes its legal effect; none is a paraphrase.

One misattribution, labelled unsupported: the verbatim statement cited to the most lexically similar provision of another article of the same act, skipping any provision that contains the statement.

The decisive edits fall into nine families in two groups. The *modelled* families are the six whose element types the reference prototype checks (section 5); the *held-out* families are three that it does not, kept out of its design to test whether typed checking generalises.

NEG Negation. Insert or remove *not* after the first deontic modal.

MOD Modality. Replace *shall* by *may* or the reverse, or *must* by *may*.

NUM Quantity. Change one number that is not part of a cross-reference.

DATE Date. Shift the year of a date by one.

UNIT Time unit. Replace a unit of time by another (*days* and *months*, *working days* and *calendar days*, and similar).

XREF Cross-reference. Change the number or letter of a referenced article, paragraph, point, chapter, section or annex.

ACT Actor (held out). Swap a legal role for another role of the same act, such as controller and processor.

QNT Comparative or quantifier (held out). Reverse *at least* and *at most*, *higher* and *lower*, and similar, or weaken *all* to *some*.

SCOPE Exception (held out). Delete an exception clause introduced by *unless*, *except*, *other than*, *provided that* and similar.

Table 3 shows one test-split item of each family, quoted exactly from the item file. The original is the statement as it appears in the act; the item is what the systems were asked to verify against the cited provision.

3.2 Citation resolution (CIT)

A citation item is a string that cites a provision in one of six formats: the Official Journal style (“Article 5(1) of Regulation (EU) 2016/679”), the same with a point, the same without the “(EU)” marker, the short name of the act (“Article 5(1) GDPR”), an abbreviated form (“Art. 5(1) GDPR”) and the Czech form (“čl. 5 odst. 1 nařízení (EU) 2016/679”). Every fourth item cites an article beyond the act’s last, so that it names a provision that does not exist [IMPLEMENTED]. The test split holds 325 citation items [MEASURED]. A system under test must report the article and resolve the instrument.

Table 3: One test-split item per family, quoted exactly. Changed words are in italics. In the spacing variant the no-break space after "a" is replaced by an ordinary space, which is invisible in print. The misattributed statement is cited to the definitions article of the same act.

Family	Statement in the act	Item	Source
Spacing	This shall include a risk assessment of the incident, and corrective action.	This shall include a risk assessment of the incident, and corrective action.	AI Act 73(6)
Quotation marks	infringing party's annual turnover in the preceding financial year in the Union.	infringing party′s annual turnover in the preceding financial year in the Union.	Data Act 40(3)
Initial capital	*T*ime shall begin to run on the day on which the decision becomes final.	*t*ime shall begin to run on the day on which the decision becomes final.	DSA 78(2)
Final punctuation	a list of the third parties consulted as part of the audit*;*	a list of the third parties consulted as part of the audit	DSA 37(4)
Negation	Testing in real world conditions shall *not* be covered by that exclusion.	Testing in real world conditions shall be covered by that exclusion.	AI Act 2(8)
Modality	Time *shall* begin to run on the day on which the decision becomes final.	Time *may* begin to run on the day on which the decision becomes final.	DSA 78(2)
Quantity	That decision shall be implemented by the issuer within *40* working days of the date of notification of that decision.	… within *20* working days …	MiCA 46(2)
Date	By way of derogation from paragraph 2, Titles III and IV shall apply from 30 June *2024*.	… shall apply from 30 June *2023*.	MiCA 149(3)
Time unit	… within 40 *working* days of the date of notification of that decision.	… within 40 *calendar* days …	MiCA 46(2)
Cross-reference	to issue an opinion and approve draft codes of conduct pursuant to Article *40*(5);	… pursuant to Article *41*(5);	GDPR 58(3)
Actor	The *controller* shall provide the information referred to in paragraphs 1 and 2:	The *processor* shall provide the information referred to in paragraphs 1 and 2:	GDPR 14(3)
Comparative/ quantifier	Without prejudice to the second subparagraph, the support period shall be at *least* five years.	… shall be at *most* five years.	CRA 13(8)
Deleted exception	Financial entities, *other than microenterprises,* shall maintain redundant ICT capacities …	Financial entities, shall maintain redundant ICT capacities …	DORA 12(4)
Misattribution	the protection of the data subject or the rights and freedoms of others;	The same words, cited to Article 4	GDPR 23(1)

3.3 Temporal applicability (TMP)

A temporal item asks whether a provision applies on a date. Its label is one of three states, computed from the schedules of section 2: *not yet in force*, *in force but not yet applicable*, and *applicable*. Items sample 60 provisions per act and ask about the day before and the day of publication, of entry into force and of application, plus a uniformly drawn date for about half of them; AI Act provisions of Chapter III are labelled per category of system [IMPLEMENTED]. Provisions whose application depends on facts rather than a date (such as Data Act Article 3(1)), on interpretation (the "corresponding obligations" of the original AI Act Article 113(c)) or on sub-paragraph granularity are excluded. The test split holds 2,535 temporal items. Besides the items, the run counts exactly, for every provision of the corpus and every day from 1 January 2016 to 31 December 2028, on how many provision-days four heuristics contradict the schedule.

3.4 Jurisdictional scope (JUR)

A jurisdiction item is a sentence from an answer for a Czech matter that cites primary law of one of eight legal orders: Czech, EU, Slovak, German, Austrian, Polish, US or UK. The citations are placed into four answer templates [IMPLEMENTED], for 60 items; for example, "Under § 823 BGB, the defendant is liable for the damage caused." Czech and EU law are labelled domestic, the other six foreign.

3.5 Release discipline (GRD)

A release item is a legal conclusion with no citation, a fabricated citation or a real one, for 20 items; for example, "The controller is liable and must pay compensation to the data subject. See Article 412(3) of Regulation (EU) 2016/679.", which cites an article the regulation does not have.

4 Labels, safeguards and splits

4.1 Labels by construction

Every label in LLRB-1 is fixed by the way the item was made from the official text, and none is assigned or reviewed by a model [MEASURED]. The provision's own words support a statement made of them; a typographic variant changes no word; one decisive edit makes the statement say something the provision does not; a statement cited to an article that does not contain it is misattributed. Temporal labels are computed from the transcribed schedules, whose every date is tied to a verbatim quote of the act [MEASURED]. Jurisdiction and release labels follow from the legal order cited and from whether the cited article exists.

This choice trades breadth for exactness. A model-assigned or crowd-assigned label can cover paraphrase and inference; it also brings the labeller's errors into the ground truth, and those errors correlate with the errors of the systems being measured when the labeller is itself a model. Labels by construction cover only what construction can decide. Within that range, they are exact up to the safeguards below.

4.2 Validity safeguards

Construction makes labels mechanical, not infallible. Four safeguards address the known ways it can go wrong.

1. *An edit that the provision happens to support.* A perturbed statement could, by coincidence, occur elsewhere in the cited provision. Any edited statement whose canonical form occurs in its provision is discarded.
2. *A misattribution that is in fact correct.* The provision chosen as the wrong citation could contain the statement. Candidate provisions that contain it are skipped.

3. *A date that is not in the act.* Every rule of the temporal ground truth carries a verbatim quote of the final provision it comes from, and an automatic check found all 26 quotes in the official text [MEASURED].
4. *A date that a date cannot decide.* Provisions whose application turns on facts, on interpretation or on sub-paragraph granularity are excluded from temporal items rather than given a guessed label [IMPLEMENTED].

The safeguards do not make a lawyer's review unnecessary. No human legal expert has reviewed the labels [ABSENT]. Section 9 lists the residual threats, including items whose wording no drafter would produce and edits whose family assignment is approximate.

4.3 Splits

Items are assigned to the development or the test split by a hash of the act and the article, about three articles in ten to development, so that no article contributes items to both splits [IMPLEMENTED]. An article is the natural unit of leakage in legal text: its paragraphs share defined terms, actors and phrasing, and a threshold tuned on one paragraph would be tested on its neighbour if the split were by item. The development split holds 1,767 support items, 155 citation items and 1,144 temporal items; the test split holds 4,053, 325 and 2,535, and the jurisdiction and release items, which have no development counterpart, are all in the test split [MEASURED].

The development split was used to build the harness, set every threshold and choose the hypotheses. The test split was evaluated once, after registration (section 6.3).

5 Systems under test

5.1 Lawie components

Lawie's components were imported unchanged from the code basis and executed offline, without a database, network, model or secret. Each run records the code commit and the SHA-256 digest of every exercised source file, and refuses to run if a file differs from the commit [MEASURED]. Inputs that a production path would read from storage, such as an approved official-source contract, an observation whose payload is the official provision text and a receipt for the byte span of a sentence, are constructed from the official text, and every decision recorded is the return value of the Lawie function [IMPLEMENTED]. The components are described here by function.

Lawie grounding. Lawie's lexical grounding check, the grounding stage of its verification pipeline. It was given the cited provision as the only retrieved source and the statement as the only claim.

Lawie quote matcher. The quoted-holding matcher of Lawie's research module. It was run twice: with the provision's text as its reference, its most favourable input, and given only the citation's metadata instead of the provision text.

Lawie exact span. Lawie's citation-authority predicate. It treats a statute version as authoritative only when its text is an exact byte span of an official-source observation retrieved under an approved, active source contract, still current and in force at the citation's date [TESTED]. The candidate statement stood in for the version text, with a receipt for the authentic sentence. It was run with the one admitted span derivation, which isolates its byte-span logic, and with an EU derivation, which shows the behaviour of Lawie's code path on EU text, since exactly one derivation, for Czech statute paragraphs, is admitted [IMPLEMENTED].

Lawie temporal components. A temporal index populated as Lawie's EUR-Lex parser populates versions, from the entry-into-force date with an open end [IMPLEMENTED], and a variant populated from the publication date, as the parser's metadata path does when no effective date is present; and the in-force check of the citation-authority predicate.

Lawie gateway rules. The jurisdiction-scope rule, which treats Czech, Slovak and EU law as one compatible bloc and recognises foreign markers for a short list of legal orders [IMPLEMENTED], and the citation-required rule, which is meant to stop legal conclusions that carry no citation.

To check the harness, the component's decisions were reproduced independently on all 4,053 test items [MEASURED] and all 1,767 development items.

5.2 Baselines

Exact match. The statement is a byte substring of the provision.

Canonical match. The statement is a substring of the provision after both are canonicalised: compatibility normalisation, case folding, and folding of quotation marks, dashes, spaces and final punctuation.

Containment (calibrated). Token containment as defined for this benchmark: the share of the statement's tokens of three or more characters that occur anywhere in the provision.

Token F1 and Jaccard (calibrated). Token F1 and Jaccard similarity between the statement and the best window of one to four consecutive clauses of the provision.

The three similarity baselines were calibrated on the development split, each at the score of the fifth percentile of the 618 supported development items. The thresholds were 1.0 for containment, 0.971429 for token F1, just above 34/35, and 0.941176 for Jaccard; token F1 kept 586 of the 618 (94.8%) [MEASURED]. Calibration fixes the false-reject rate and leaves the false-accept rate to fall where it falls, which is the point of hypothesis H2.

5.3 The reference prototype

The typed reference prototype is a research instrument of this suite, not a Lawie component [PROPOSED]. It accepts a statement when some window of one to four clauses of the cited provision contains at least 90% of the statement's tokens and the multiset of the decisive elements it models is identical in statement and window. The elements are the deontic operator with its polarity, quantities, dates, time units and cross-references; the containment threshold of 0.9 is a design constant, not a calibrated value. The prototype deliberately does not model actors, comparatives and quantifiers, or exception clauses. Those are the held-out families, kept out of its design so that the benchmark can test whether typing one set of elements protects against edits of another.

The prototype and the edit families were defined by the same authors, and the prototype types exactly the element classes that the modelled edits alter; its rejection of every modelled edit is expected by design, and the held-out families are the informative test [ANALYTICAL]. It is included to show what typed verification buys and what it does not, not as a candidate product.

6 Metrics, statistics and pre-registration

6.1 Metrics

Acceptance rate. For each system and family, the share of items accepted. On supported families it is one minus the false-reject rate (FRR); on unsupported families it is the false-accept rate (FAR).

Confidence. For the verification pipeline's reported confidence, the area under the receiver operating characteristic between supported and unsupported items, with ties counted one half [14], and the expected calibration error over ten equal-width bins [15], treating "supported" as correct.

Resolution. Per citation format, the share of items for which the extractor reported the right article, and the share for which it also resolved the right instrument.

Temporal error. Per component and per true state, the share of dates on which the component treated the provision as applicable when it was not, or the reverse; and the exact number of provision-days on which each of four heuristics contradicts the schedule between 1 January 2016 and 31 December 2028.

Scope and release. The share of answers flagged per cited legal order; and the share of conclusions blocked per kind of citation.

6.2 Statistics

Most observed rates in this benchmark are at or near 0 or 1, where normal approximations to the binomial fail. Every interval and test is therefore exact or nearly so.

- Descriptive rates carry Wilson score intervals at 95% [16], which behave far better than the normal approximation near the boundaries and never leave $[0, 1]$.
- The quantities of the registered hypotheses carry Clopper–Pearson intervals [17], which are exact and conservative: when all n items are accepted, the lower bound is $(0.025)^{1/n}$.
- One-sided hypotheses about a rate are tested by exact binomial tests against the registered value.
- The comparison of two systems on the same items (H3a) uses the exact two-sided McNemar test on discordant pairs [18], which conditions on the items where the systems disagree and ignores the rest.
- Within each registered family of tests (the two tests of H1 and the three of H2) p-values are adjusted by Holm's sequentially rejective procedure [19], which controls the family-wise error rate without assuming independence.
- The significance level is $\alpha = 0.05$. A p-value below 10^{-6} is reported as such. Descriptive results carry no test and are not corrected for multiplicity.

6.3 Pre-registration

The hypotheses, their tests and decision rules, the thresholds, the rule for choosing the H3a comparator, the significance level, the checksums of every data file and the digest of the exploratory run were committed before the confirmatory run, which ran once [MEASURED]. The run refuses to start if a data file differs from its registered checksum, if the registration is not committed, or if an exercised Lawie source file differs from the code basis.

Hypothesis H1. *Lawie's grounding check, as implemented at the code basis, accepts more than half of the statements that carry one decisive edit, and more than half of the misattributed statements. Two exact one-sided binomial tests against 0.5, Holm-corrected.*

Hypothesis H2. *Each calibrated lexical baseline (containment, token F1, Jaccard) accepts more than 5% of decisively edited statements. Three exact one-sided binomial tests against 0.05, Holm-corrected, reported per baseline.*

Hypothesis H3a. *The reference prototype accepts fewer modelled-family edits than the comparator (exact two-sided McNemar test) and rejects fewer than 5% of supported statements (exact one-sided binomial test against 0.05).*

Hypothesis H3b. *On the held-out families the reference prototype accepts more than half of the edits. Exact one-sided binomial test against 0.5.*

H1 and H2 state that similarity is not support; H3a that typing the decisive elements removes the modelled errors at no measured cost to supported statements; H3b that it does not remove errors of types it does not model. Under the registration a refuted hypothesis would have been reported as refuted.

Exploration came first, and is reported. The development split was used to build the harness, set the thresholds and choose the hypotheses. Two choices made there should be visible to a reader. First, the exploratory run used token F1 as the H3a comparator; after seeing the development results, the registration switched to the calibrated baseline with the lowest development acceptance of modelled-family edits, and so the hardest to beat: containment, at 31.5%, against 63.0% for token F1 and 55.5% for Jaccard [MEASURED]. Second, an earlier draft of H1 stated that every lexical baseline accepts more than half of the decisive edits. On the development split the calibrated baselines accepted 32.3%, 53.9% and 49.3%, so the draft was not supported for two of the three, and the protocol records that it was replaced by the present H1 and H2 before registration; the registration lists no deviations [MEASURED]. The present hypotheses were thus chosen with knowledge of development results. What the registration secures is narrower and still essential: that the test split, which no choice saw, was evaluated once against hypotheses fixed in advance.

Pre-registration separates prediction from postdiction, so that a confirmatory result cannot be a story told after the data were seen [20]. The REFORMS recommendations for machine-learning-based science ask in particular for evidence against leakage between the data used to make choices and the data used to evaluate them [21]; the split by article (section 4) and the checksum guard are LLRB-1's answer. They do not protect against a design that favours a system: the prototype and the edit families come from the same hands, which is why H3b, and not H3a, carries the weight.

7 Results

7.1 Confirmatory hypotheses

Table 4: Confirmatory hypotheses on the test split. Intervals are Clopper–Pearson; p-values are exact and 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

All four registered hypotheses were supported (table 4).

- **H1.** Lawie's grounding check, as implemented at the code basis, accepted 1,959 of 1,959 decisively edited statements and 698 of 698 misattributed statements; the 95% lower bounds are 0.998 and 0.995 [MEASURED].
- **H2.** Calibrated containment, token F1 and Jaccard 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 [MEASURED].

- **H3a.** The prototype accepted 0 of 1,444 modelled-family edits, against 392 for the comparator, and rejected 0 of 1,396 supported statements, with a 95% upper bound of 0.26% [MEASURED].
- **H3b.** The prototype accepted 456 of 515 held-out edits (88.5%, 95% CI 85.5–91.2) [MEASURED].

7.2 Support by family

Table 5 gives the acceptance rate of every system on every family of the test split, and figure 1 plots the acceptance of unsupported items for Lawie's grounding check, the three calibrated baselines and the prototype.

Table 5: Acceptance (%) by system and family, test split. First block supported (higher is better); the others unsupported (lower is better): modelled families, held-out families, misattribution. [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

Three patterns stand out. Lawie's check's bars are full in every family: it accepted every unsupported item. Only on date edits (0–20%) and misattributions (1–8%) are all three calibrated baselines low, and each accepts most edits of some other family: containment accepts every deleted exception and almost every cross-reference edit, token F1 and Jaccard most negation edits. The prototype is at zero on every modelled family and high on every held-out family. Section 8 explains each pattern by mechanism.

Typographic variants separate the exact checks from the rest. Lawie's exact-span predicate accepted every authentic quotation and rejected every typographic variant and every edit; with the derivation that EU text would need it rejected everything [MEASURED]. Canonical matching accepted every authentic quotation and every variant and rejected every edit and every misattribution [MEASURED], a result that holds by construction for the edit families used (section 9). Lawie's quote matcher, given the provision text, accepted 1,956 of 1,959 (99.8%) decisively edited quotations and 51.6% of misattributed ones; given only the citation's metadata instead of the provision text, it accepted 1.0% of authentic quotations [MEASURED].

Confidence. The verification pipeline reported a confidence of 0.973 with the label HIGH for every supported and every unsupported statement: AUROC 0.5, expected calibration error 0.63 [MEASURED]. A score that does not vary between supported and unsupported items cannot discriminate between them, whatever

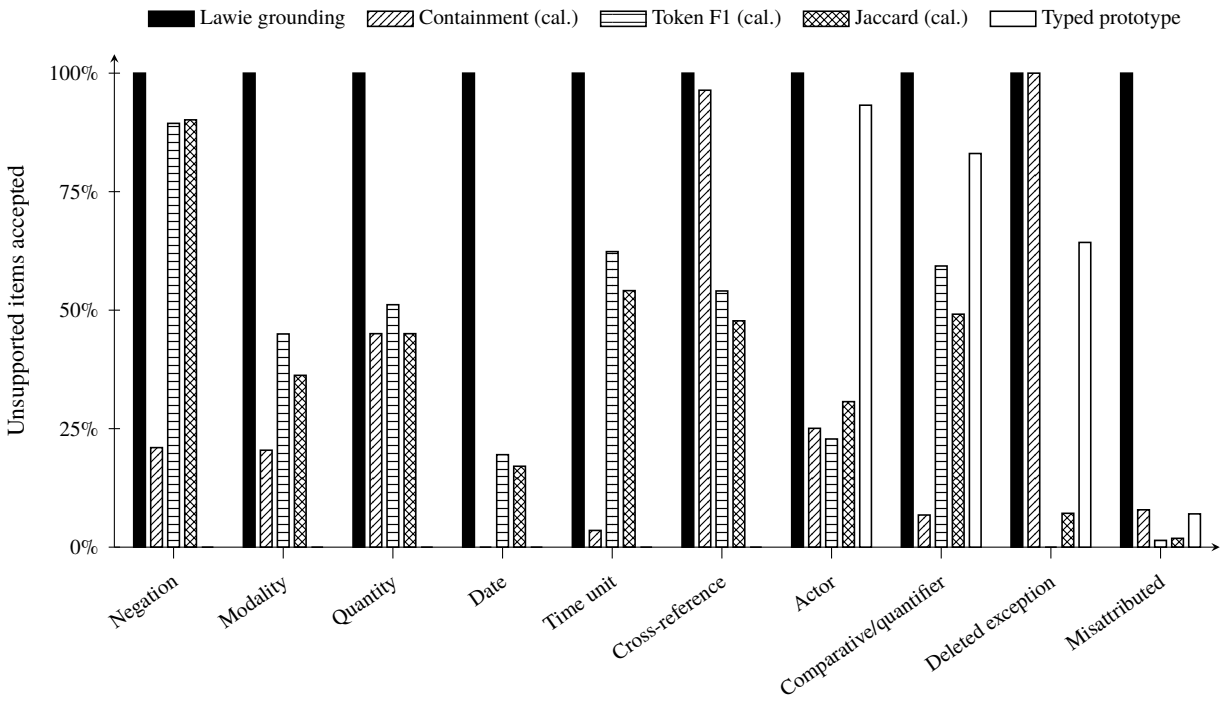

Figure 1: Share of unsupported test-split items accepted, by family and system. The first six families are the edit types the prototype models; actor, comparative/quantifier and deleted exception are held out from its design. A missing bar is zero. Lawie's grounding check accepts every item of every family.

its form. The calibration error is the distance between 0.973 and the share of supported items in the test split, 1,396 of 4,053, and is a property of the benchmark's mixture rather than of any deployment [ANALYTICAL].

7.3 Citations and time

Table 6 gives citation resolution by format; tables 7 and 8 give temporal errors on the test dates and the exact exposure of four temporal heuristics over 2016–2028.

Table 6: Citation resolution by format, test split. "Article found": the extractor reported the cited article; "instrument resolved": it also identified the right act. Every fourth item cites a non-existent article, and all of those fall in the point, abbreviated and Czech formats, so the table does not measure refusal of fabricated provisions [MEASURED].

Citation format (example)	n	Article found (%)	Instrument resolved (%)
Article 5(1) of Regulation (EU) 2016/679	58	100	0
Article 5(1), point (a), of Regulation (EU) …	56	100	0
Article 5(1) of Regulation 2016/679	49	100	100
Article 5(1) GDPR	53	100	38
Art. 5(1) GDPR	53	0	0
čl. 5 odst. 1 nařízení (EU) 2016/679	56	0	0

Lawie's citation extractor resolved the instrument for none of the citations in the official "Regulation (EU)" style, with or without a point (0 of 114), for all 49 written without "(EU)", for 37.7% of short-name citations, and found no article in abbreviated or Czech-format citations [MEASURED] (table 6).

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; the variant keyed on publication also erred on 791 of 1,298 dates before entry into force

Table 7: Temporal errors by true state on the date asked, test split. 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%)

Table 8: Share of provision-days (%) from 1 January 2016 to 31 December 2028 on which a heuristic contradicts the official schedules, counted exactly over every provision of the corpus. "From publication" and "from entry into force" treat validity from that date as applicability; "general date only" ignores provision-specific dates; "frozen" uses the AI Act schedule as known on 1 July 2026, before Regulation (EU) 2026/1744.

Act	From publication	From entry into force	General date only	Frozen before 2026/1744
AI Act	12.4	12.0	3.4	–
AI Act (Annex I systems)	16.1	15.7	5.7	–
AI Act (Annex III systems)	15.1	14.7	4.7	2.0
CRA	20.9	20.5	2.6	–
DORA	15.8	15.4	0.0	–
DSA	7.3	6.9	2.7	–
Data Act	13.3	12.8	0.0	–
GDPR	15.8	15.4	0.0	–
MiCA	9.9	9.4	2.1	–
NIS2	13.9	13.5	0.0	–

[MEASURED] (table 7). 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 Annex III provision-days [MEASURED] (table 8).

7.4 Jurisdiction and release

In answers for a Czech matter the gateway's scope rule flagged citations of US law (8 of 8) and never flagged citations of Slovak, German, Austrian, UK or Polish law (0 of 32) [MEASURED]. Its citation-required rule blocked every legal conclusion without a citation (4 of 4) and none with a fabricated citation (0 of 8) [MEASURED].

8 Analysis by family

Every pattern in table 5 has a mechanism that can be read off the definition of each baseline and of the prototype. This section gives them family by family. The proofs of the general statements are in the companion paper (LL-REL-01); here they are applied. The section is descriptive: its checks were made after the confirmatory run, explain registered results and test no hypothesis.

8.1 Three mechanisms

Lawie's grounding check. Lawie's grounding check accepted every unsupported item of every family [MEASURED]. LL-REL-01 shows analytically that any overlap measure with a threshold below one tolerates single edits.

Containment sees only new vocabulary. At its calibrated threshold of 1.0, containment accepts a statement exactly when every token of more than two characters occurs somewhere in the provision. Its acceptance rate in each family is therefore the share of edits that introduce no such new token: every deletion, every inserted word the provision already uses elsewhere, and every change confined to tokens of one or two characters, such as a paragraph number, a point letter or a two-digit quantity, which the benchmark's containment does not count [ANALYTICAL].

Symmetric measures see length. Against a window that coincides with the statement, token F1 is $2n/(2n+1)$ after one inserted token, $(n-1)/n$ after one substituted token and $2(n-k)/(2n-k)$ after k deleted tokens. At the calibrated threshold of 0.971429, just above 34/35, an insertion passes in statements of at least 18 tokens, a substitution in statements of at least 36, and a deletion of three tokens only in statements of at least 55 (LL-REL-01). A post hoc check, not part of the registered analysis, finds the prediction holding on the test split: 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 [MEASURED]. Jaccard similarity behaves alike.

8.2 Family by family

Typographic variants. Lawie's grounding check, its quote matcher given the provision text, containment, canonical matching and the prototype accepted every typographic variant; the exact-span predicate rejected them all, which is correct for a byte-exact check and fatal for its use on text that has been retyped. Token F1 and Jaccard rejected a few verbatim statements and spacing or quotation-mark variants, and about a quarter of final-punctuation variants: a post hoc check shows that every final-punctuation variant that token F1 rejected is a statement that follows a point label in its provision [MEASURED], so its best window carries an extra token, and a short statement cannot absorb one.

Negation. Containment accepted 21%: every removal of *not*, which is a deletion, and every insertion into a provision that uses *not* elsewhere. Token F1 and Jaccard accepted about nine in ten, because a one-word insertion barely moves a symmetric measure in a long sentence. The prototype rejected every negation edit, because it types the deontic operator with its polarity.

Modality. *Shall* and *may* are both common in EU legislation, so containment accepted the 20% of modality edits whose new modal occurs elsewhere in the provision. Symmetric measures accepted fewer than half, since a substitution costs twice what an insertion does. The prototype rejected all.

Quantity. Containment accepted 45%. A change of "40 working days" to "20 working days" is invisible to it, because its tokeniser discards one- and two-digit numbers; a change of "two" to "one" is visible unless the provision says "one" elsewhere.

Date. Containment rejected every date edit, because a four-digit year that the provision does not contain is new vocabulary. Here a lexical check looks competent, for a reason unrelated to the meaning of dates.

Time unit. Containment rejected almost all (4% accepted), for the same reason: "calendar" and "months" are usually new words. Token F1 and Jaccard accepted more than half.

Cross-reference. Containment accepted 96%: most article, paragraph and point numbers are tokens of one or two characters. A statement that sends the reader to Article 41(5) instead of Article 40(5) is, to a containment check, the same statement. The prototype, which types cross-references, rejected all.

Actor. Containment accepted 25%, the edits whose substituted role already occurs in the provision, as it often does when a provision regulates the relation between two parties. The prototype accepted 93%, because it does not model actors; nothing in its multiset of decisive elements changes when "controller" becomes "processor".

Comparative or quantifier. Containment accepted 7% and the prototype 83%. "At least five years" and "at most five years" contain the same number, the same unit and the same modal.

Deleted exception. Containment accepted every one: deletion cannot introduce a new token. Token F1 accepted none and Jaccard 7%, because a multi-word deletion is a large change to a symmetric measure. The prototype accepted 64%: it accepted exactly the deletions that left its modelled elements unchanged and rejected those whose deleted clause contained a cross-reference, a number or a modal, which is a by-product of typing, not a model of exceptions.

Misattribution. Lawie's grounding check accepted every misattributed statement and the quote matcher, given the provision text, about half. Calibrated containment accepted 8%, token F1 1% and Jaccard 2%. The prototype accepted 7%: EU acts repeat formulae, and its acceptances are near-identical clauses of the wrong article, some with the statement's words and others differing in a word or two that it does not type. Only an exact check, the exact-span predicate or canonical matching against the cited provision, rejected them all.

What the families show together. No lexical system has a family profile that tracks legal effect. Each accepts the edits that its arithmetic cannot see, and the arithmetic is unrelated to the law: containment cannot see deletions and short tokens, symmetric measures cannot see single-word changes in long sentences. The typed prototype shows the alternative and its limit. Typing the elements that carry legal effect removes the errors in those elements entirely, and leaves the rest to its 90% containment test, which accepts more actor and comparative edits than the calibrated baselines. The held-out families were not chosen to embarrass it: actors, comparatives and exceptions are ordinary elements of legal rules, and a verifier that ignores them ignores whole classes of legal error.

9 Threats to validity

9.1 Construct validity

Support is tested through a sufficient condition. LLRB-1 labels the provision's own words and their typographic variants as supported and a single decisive edit or a wrong citation as unsupported. It does not test paraphrase. A verifier that accepts only quotations therefore scores perfectly, and canonical matching did. That row is a property of the design: the benchmark discarded every edited statement whose canonical form occurs in its provision, and every edit changes a character that the canonicalisation does not fold. The same verifier would reject every legitimate paraphrase in use. The false-reject rates reported are rates on quotations and are lower bounds on what a paraphrasing generator would meet.

Constructed errors are not generated errors. A false-accept rate measured on single decisive edits transfers to deployment only to the extent that generated errors resemble them. Generated legal errors also include invented provisions, merged provisions, changes to several elements at once and gradual drift in paraphrase. LLRB-1 measures the verifier's response to one well-defined kind of error; the base rates of the kinds that occur in practice are the subject of the deferred study [PROPOSED].

The prototype and the edits share authors. The reference prototype types exactly the element classes the modelled edits alter, and both were defined by the same authors [ANALYTICAL]. Its rejection of every modelled edit was expected by design. H3b, on families held out from its design, is the informative test.

Misattributions are hard ones. Each misattribution cites the most lexically similar provision of another article of the same act. A random wrong citation would be easier to detect. The misattribution rates are rates for near-misses, which are the misattributions most likely to survive a human reader's glance.

9.2 Label validity

Mechanical is not infallible. The safeguards of section 4 remove edits that happen to occur in the provision and misattribution targets that contain the statement. They do not make family membership exact. The quantity family is meant to exclude numbers that are part of a cross-reference or a date, and in some test items it does not: "Paragraph 1, point (b), shall not apply to microenterprises and small enterprises." became "Paragraph 11, point (b), shall not apply …", and "… Titles III and IV shall apply from 30 June 2024." became "… from 30 June 1012." Both items are correctly labelled unsupported; their family is approximate, so per-family rates carry a small and unmeasured admixture of neighbouring families.

No lawyer has reviewed the labels. No human legal expert has reviewed the labels [ABSENT]. A review should sample items per family, have at least two lawyers label them independently against the cited provision, and report their agreement with a chance-corrected coefficient such as Krippendorff's α [22] before any label is changed.

Temporal labels inherit the schedules. Temporal labels are only as good as the transcription of each act's final provisions. Every rule carries a verbatim quote that was found in the official text [MEASURED], and provisions whose application a date cannot decide are excluded [IMPLEMENTED]. The labels use one knowledge date; a later amendment would change some of them, as Regulation (EU) 2026/1744 changed the AI Act's.

9.3 Realism, design and sampling

Perturbation realism. Some edits produce text that no drafter and few models would write: "The term of office of the co-chairs shall be one years, renewable once."; "Competent authorities shall cooperate with one other for the purposes of this Regulation."; "Financial entities, shall maintain redundant ICT capacities …"; "Where the data subject makes the request by electronic means, and, the information shall be provided in a commonly used electronic form." For lexical systems, grammar is invisible and these artefacts change no decision. For a verifier built on a language model, disfluency is a cue that a fluent generated error would not offer, and it could inflate apparent detection. Model-based verifiers should not be compared on these items without a control for fluency.

Design of the citation family. Citation formats and fabrication were both assigned by cycling through the items. As a result, every fabricated citation in the test split fell in one of the three formats in which the extractor never resolved an instrument (official with a point, abbreviated and Czech) [MEASURED]. Whether the extractor refuses a non-existent article of an instrument it does resolve was therefore not tested, and the citation results say nothing about refusal of fabricated provisions. A citation benchmark tests refusal only if it assigns fabrication independently of format.

Single edits, sampled statements. Each unsupported item carries exactly one edit. For symmetric measures a single edit is the hardest case, since more edits move the score further; for containment, deletions of any size remain invisible. Statements were sampled per act and enriched for rare families [IMPLEMENTED], so no pooled rate in this report estimates a rate in any population of real statements, and none is offered.

9.4 External and statistical validity, and conflict of interest

External validity. The corpus is English-language EU legislation, eight acts. Czech and other national law, case law, contracts and other languages are not covered, and verification of national statute text faces different typography and citation forms. Lawie's components were run in isolation, with inputs constructed from the official text [IMPLEMENTED]; the grounding check saw only the cited provision as its retrieved source, and a retrieval that returns more text is a different setting. The jurisdiction and release families are small

template sets of 60 and 20 items: they demonstrate behaviours and do not estimate rates beyond their templates.

Statistical conclusion validity. The registered intervals and tests are exact, the confirmatory run was executed once, and the registered families of tests are Holm-corrected. The H3a decision combines two tests, both of which had to pass. The descriptive analyses are many and uncorrected, and are offered as descriptions. With 16 quotation-mark variants, 41 date items and 42 deleted-exception items in the test split, per-family intervals for the rarest families are wide, and differences of a few points between systems on those families carry little weight.

Conflict of interest. The benchmark was built by the company whose components it measures. Pre-registration, the checksum guard, the publication of every negative result and the recorded digests of every exercised source file [MEASURED] reduce the room for a favourable reading. They do not remove the conflict, and no independent audit or review of the components is recorded [ABSENT]. This report has not been peer reviewed. The registration is a commit in the company's own repository, not an entry in an independent registry; its precedence over the run is attested by the run record, which embeds the registration commit, and by our commit history, and that the run was executed once is attested by our record.

10 Datasheet

This section answers the questions of the datasheet for datasets [23], in their order.

Motivation. *For what purpose was the dataset created?* To measure whether verification components of legal AI systems detect errors that change legal meaning, per class of error, so that the released error of a gate can be reasoned about (section 1). *Who created it?* Lawie Labs, the research function of the company whose components it measures; the conflict of interest is stated in section 9.

Composition. *What do the instances represent?* Statements paired with the provision they cite (SUP), citation strings (CIT), provision–date pairs (TMP), answer sentences citing primary law (JUR) and legal conclusions with or without citations (GRD). *How many?* 5,820 support items from 1,007 statements [MEASURED]; in the confirmatory test split 4,053 support items, 325 citation items, 2,535 temporal items, 60 jurisdiction items and 20 release items, and in the development split 1,767, 155 and 1,144 items of the first three families [MEASURED]. *Is it a sample?* The corpus is complete for the eight acts; support statements are a seeded sample per act enriched for rare edit families [IMPLEMENTED]; temporal items sample 60 provisions per act [IMPLEMENTED]. *What are the labels?* Supported or unsupported with the edit family (SUP); article and instrument (CIT); not in force, in force but not applicable, or applicable (TMP); domestic or foreign (JUR); kind of citation (GRD). *Are there errors or noise?* Labels are fixed by construction and safeguarded as described in section 4; family membership is approximate in a few items (section 9). *Is it self-contained?* Yes: the corpus file holds the text of every provision, and the item files refer to it by identifier. *Does it contain personal or confidential data?* No. It contains public legislation and constructed sentences; no person, client or case is named.

Collection process. The acts were retrieved from EUR-Lex on 26 November 2026 and parsed into provisions; the SHA-256 digest of each rendering is recorded [MEASURED]. Items were generated by a deterministic, seeded program from the corpus and the transcribed schedules. No model generated, filtered or labelled any item. No human annotator was involved, and no lawyer has reviewed the labels [ABSENT].

Preprocessing and labelling. Provisions are paragraphs of articles. Statements are sentences or clauses of 12 to 70 words; leading point labels are dropped from statements and kept in the provision text. Temporal labels are computed from the schedules [MEASURED]; provisions whose application a date cannot decide are excluded [IMPLEMENTED].

Uses. *Intended:* measuring the false-accept and false-reject rates of verification components per class of error, and the behaviour of citation, temporal, jurisdiction and release components on EU legislation. *Not intended:* estimating how often generated answers are wrong; ranking language models; certifying a system as reliable; any claim about paraphrase, case law, national law or languages other than English. A perfect score on LLRB-1 bounds the rates of the measured failure modes on its items; it does not show reliability.

Distribution. The benchmark's harness and data are held with the eval lab and may be made available to qualified reviewers on request. The corpus is derived from EUR-Lex under the Commission's reuse decision [13]; only the Official Journal is authentic.

Maintenance. LLRB-1 is versioned with its data checksums. A corrected or extended release will carry a new version and a new registration; the results in this report remain tied to the registered version. The limitations that any later version must face are those of section 9.

11 Relation to prior benchmarks

Legal language benchmarks. LegalBench collects legal reasoning tasks built with legal professionals and scores the answers of language models [24]. CaseHOLD asks a model to identify the holding of a cited case among candidates [25]. LexGLUE gathers English legal language-understanding tasks, among them the classification of EU legislation [26]. These benchmarks measure what a model knows or can infer. LLRB-1 measures something else: whether a component that checks an answer against the official text notices when the two differ. A system can score well on the first kind of benchmark and its verification layer still fail the second.

Studies of legal hallucination. Dahl and colleagues [1] and Magesh and colleagues [2] measure how often models and research tools produce incorrect or misgrounded legal answers. Those studies measure the base error rate that a verifier faces; LLRB-1 measures the verifier. The two are complements, and the released error of a gated system needs both (section 1). LLRB-1's deferred generated-answer study (section 12) is the part of this programme that would measure base rates.

Attribution and faithfulness evaluation. The attribution literature defines when a statement is attributable to a source [27], measures citation recall and precision in generated text [28], scores the share of atomic facts supported by a knowledge source [29], and measures the faithfulness of retrieval-augmented answers [30]. Several of these use a natural language inference model or a language model as the judge of support. LLRB-1 uses no model as a labeller, which limits it to what construction can decide and keeps the errors of a model judge out of its ground truth.

Behavioural testing, contrast sets and counterfactual data. CheckList organises behavioural tests of NLP systems into minimum functionality, invariance and directional expectation tests [31]. In those terms LLRB-1's typographic variants are invariance tests and its decisive edits are directional tests: the label must flip. Contrast sets are small, expert-written perturbations of test instances that change the gold label and probe a model's local decision boundary [32]; counterfactually augmented data asks humans to revise an instance minimally so that its label changes [33]. LLRB-1 builds contrast sets automatically, with edits typed by their legal function. The trade-off is explicit: automatic construction gives scale and exact labels, and costs naturalness (section 9). Adversarial distractors that leave a reading-comprehension answer unchanged showed how easily lexical cues mislead such systems [34]; LLRB-1's misattributions play the same role for citations.

What a benchmark should be. Bowman and Dahl ask four things of a benchmark: that good performance imply competence at the task, that labels be accurate and unambiguous, that it have the statistical power

to separate systems, and that it not reward harmful bias [35]. LLRB-1 answers the first only for the failure modes it covers: passing it is necessary for a legal verifier and far from sufficient. It answers the second by construction, short of a lawyer's review. It answers the third with thousands of items in the large families, exact tests and a registered analysis, and falls short in its smallest families. The fourth is outside its scope; it contains no data about persons.

Provenance of the items. LLRB-1 takes no item from any earlier dataset or public benchmark. Its items come from the official text of the acts.

12 The deferred generated-answer study, and reproduction

The generated-answer study. LLRB-1 does not say how often a generated legal answer is wrong. The study that would say it is defined and has not been run [PROPOSED]. Its aim is to measure the rates of fabricated citations, misquotation, temporal error and abstention in answers produced by language-model systems on questions derived from the benchmark, and the change in released error when the verification layer gates those answers. Its design fixes the scoring before any answer exists:

- questions derived from support, temporal and citation items, in a closed-book condition and a condition in which the cited provision is supplied;
- structured answers that state their citations, their quotations and a confidence;
- scoring by the deterministic checks of this benchmark: existence of each cited provision in the corpus, fidelity of each quotation by canonical match, applicability against the schedules, and abstention on questions that cannot be answered;
- risk–coverage curves for gated release, per error class, so that abstention is scored apart from error.

The study needs answers generated by model services, which this programme did not use. Until it is run, the base rates that decide the released error of any gate measured here are unknown, and every statement about LLRB-1 is a statement about verifiers.

Reproduction. The numbers in this report come from one recorded run and can be recomputed from the registered data and the code basis. The corpus can be retrieved again from EUR-Lex and compared with the recorded digest of every rendering. Items are generated by a seeded, deterministic program from the corpus and the schedules. The harness uses the Python standard library and the JavaScript runtime that the code basis requires, imports Lawie's components unchanged from the code basis, refuses to run if a data file differs from its registered checksum or an exercised source file differs from the commit, and records the commit, the digest of every exercised file, the dataset checksums and the tool versions in the run record [MEASURED]. The confirmatory run of 26 November 2026 is that record.

The benchmark's harness and data are held with the eval lab and may be made available to qualified reviewers on request. A reviewer who repeats the run on the code basis should obtain the same decisions item for item; a reviewer who runs a later version of Lawie's components measures a different system, and should say so.

References

[1] 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.

[2] 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.

[3] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L 119, 4.5.2016, 2016. URL https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679.

[4] European Parliament and Council of the European Union. Regulation (EU) 2022/2554 of 14 December 2022 on digital operational resilience for the financial sector. Official Journal of the European Union, L 333, 27.12.2022, 2022. URL https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2554.

[5] European Parliament and Council of the European Union. Directive (EU) 2022/2555 of 14 December 2022 on measures for a high common level of cybersecurity across the Union (NIS 2 Directive). Official Journal of the European Union, L 333, 27.12.2022, 2022. URL https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022L2555.

[6] European Parliament and Council of the European Union. Regulation (EU) 2023/1114 of 31 May 2023 on markets in crypto-assets. Official Journal of the European Union, L 150, 9.6.2023, 2023. URL https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32023R1114.

[7] European Parliament and Council of the European Union. Regulation (EU) 2022/2065 of 19 October 2022 on a Single Market For Digital Services (Digital Services Act). Official Journal of the European Union, L 277, 27.10.2022, 2022. URL https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065.

[8] European Parliament and Council of the European Union. Regulation (EU) 2023/2854 of 13 December 2023 on harmonised rules on fair access to and use of data (Data Act). Official Journal of the European Union, L, 2023/2854, 22.12.2023, 2023. URL https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32023R2854.

[9] 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.

[10] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L, 2024/1689, 12.7.2024, 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32024R1689.

[11] Publications Office of the European Union. Regulation (EU) 2024/1689, consolidated text of 27 July 2026. EUR-Lex, document 02024R1689-20260727, 2026. URL https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:02024R1689-20260727.

[12] European Parliament and Council of the European Union. Regulation (EU) 2026/1744 of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI). Official Journal of the European Union, L, 24.7.2026, 2026. URL https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32026R1744.

[13] European Commission. Commission Decision 2011/833/EU of 12 December 2011 on the reuse of Commission documents. Official Journal of the European Union, L 330, 14.12.2011, p. 39, 2011. URL https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32011D0833.

[14] 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.

[15] 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.

[16] Edwin B. Wilson. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

[17] C. J. Clopper and E. S. Pearson. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

[18] Quinn McNemar. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996.

[19] Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6 (2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

[20] Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The Preregistration Revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018. doi: 10.1073/pnas.1708274114.

[21] Sayash Kapoor, Emily M. Cantrell, Kenny Peng, Thanh Hien Pham, Christopher A. Bail, et al. REFORMS: Consensus-based Recommendations for Machine-learning-based Science. *Science Advances*, 10(18): eadk3452, 2024. doi: 10.1126/sciadv.adk3452.

[22] Klaus Krippendorff. Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research*, 30(3):411–433, 2004. doi: 10.1111/j.1468-2958.2004.tb00738.x.

[23] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723.

[24] Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, et al. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 36 (Datasets and Benchmarks Track)*, 2023.

[25] Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset of 53,000+ Legal Holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 159–168, 2021. doi: 10.1145/3462757.3466088.

[26] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 4310–4330, 2022. doi: 10.18653/v1/2022.acl-long.297.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442.

[32] Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, et al. Evaluating Models' Local Decision Boundaries via Contrast Sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, 2020. doi: 10.18653/v1/2020.findings-emnlp.117.

[33] Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the Difference that Makes a Difference with Counterfactually-Augmented Data. In *International Conference on Learning Representations*, 2020.

[34] Robin Jia and Percy Liang. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, 2017. doi: 10.18653/v1/D17-1215.

[35] Samuel R. Bowman and George E. Dahl. What Will it Take to Fix Benchmarking in Natural Language Understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, 2021. doi: 10.18653/v1/2021.naacl-main.385.