

Outside the Four Corners: Charting the Missing Evidence in Korean AI Training-Data License Screening

Jiwoo Lee^{1♡*} Sungwon Chae^{2♡}

¹Yonsei University

²Seoul National University

kit777777t@yonsei.ac.kr, csw0815@snu.ac.kr

Abstract

As training corpora grow, screening licenses for AI training data at scale is likely to fall to large language models, but a screening system must distinguish when a document settles a proposed use from when the answer depends on facts outside it. We formalize this distinction under Korean copyright law. Supreme Court of Korea authority supports five types of extrinsic fact that may be required to decide whether a grant reaches a use or binds the rightsholder: consideration, usage of trade, the rightsholder's conduct, market substitution, and upstream title. A practicing Korean intellectual property lawyer applies this framework to Korean licenses, terms of service, and standard-form contracts, yielding a benchmark of whether each document settles five AI-related uses and, if not, which facts are missing. We evaluate frontier proprietary, open-weight, and Korean-specialized models under increasingly detailed legal procedures and broader retrieved context. Neither the lawyer's screening procedure, additional context, nor even the governing clause reliably improves models' ability to distinguish determinate from indeterminate cases; the governing clause instead increases abstention on many determinate items. These interventions mainly shift decision thresholds rather than discriminability, a difference obscured by aggregate accuracy. Automated license screening has two limits: some questions are indeterminate from the instrument because resolving them requires facts outside it, and models over-abstain on questions the document does settle.

1 Introduction

Large language models became general-purpose assistants through instruction tuning (Ouyang et al., 2022), which changed what they are used for, not

where their knowledge comes from. Loss falls as a power law in model size and in training tokens (Kaplan et al., 2020), and the compute-optimal ratio puts more of the budget on tokens than early practice assumed (Hoffmann et al., 2022). However far post-training has come, the pretraining corpus still decides what a model knows, so the appetite for text grows with every generation and is often met by copying documents whose owners are now in court over exactly that.¹ Whether a corpus may lawfully be used is therefore settled before training, and at this volume by a program. License metadata will not settle it: an audit of dataset provenance found license omission above 70% and error rates above 50% on major hosting platforms (Longpre et al., 2024a), and lifecycle tracing shows that datasets whose surface license appears to permit commercial use often do not (Kim et al., 2025). If the tags cannot be trusted the text must be read, and at pre-training scale nobody reads it: “does this license permit this use?” is answered by a model inside an automated pipeline.

Reading the text does not settle the question either. Whether a license authorizes a use raises two questions the text often leaves open. One is what the grant covers, which turns on the consideration paid, trade practice, how the rightsholder has treated the use before, and market substitution. The other is whether the grant binds the rightsholder at all, which turns on whether the licensor held the right it purports to grant. None of these facts is in the contract. Among the documents a screener can reach we find none that supplies them. The gap lies not in the document but in the legal question, which stays open. The two failure modes call for different remedies. Better retrieval finds a

♡ denotes equal contribution.

* Member of the Korean Bar, Attorney at Law, Shin & Kim LLC.

¹*The New York Times Co. v. Microsoft Corp.*, No. 1:23-cv-11195 (S.D.N.Y. filed Dec. 27, 2023), now consolidated in *In re OpenAI, Inc., Copyright Infringement Litigation*, No. 1:25-md-03143 (S.D.N.Y.); *Dow Jones & Co. v. Perplexity AI, Inc.*, No. 1:24-cv-07984 (S.D.N.Y. filed Oct. 21, 2024).

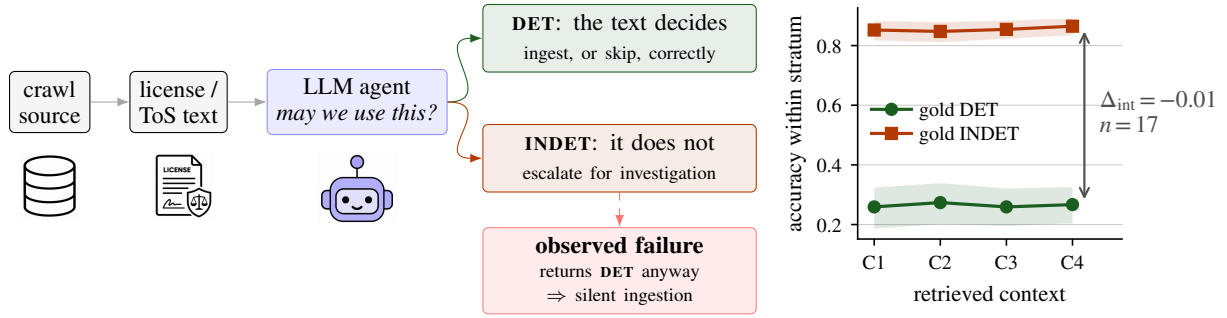


Figure 1: Screening at Scale and What It Misses. *Left*: the pipeline; the costly failure is a confident DET on an item the document does not settle. *Right*: T1 (whether the text settles the use) accuracy per gold stratum as context grows from C1 to C4 (§4.2), 95% cluster-bootstrap bands, n non-degenerate systems.

clause the pipeline missed, but nothing retrieves a fact that was never written down. A system that answers anyway may admit data the license does not reach, with no record that a judgment was made. The safe behavior is to abstain and escalate for further legal and factual investigation, and nothing we measure does that. Identifier matching commits to a verdict on most items the document does not settle. The models fail in both directions, one frontier system committing on nearly half of the unsettled items and another on one in twenty under the same prompt. Aggregate accuracy does not tell those failures apart: the keyword rule reaches 0.599 balanced accuracy, which only 5 of 57 model-and-level cells exceed.

We separate the two cases. Incidental underspecification is what existing work addresses: the absent value belongs to the form of the problem, and supplying it resolves the question. Constitutive indeterminacy is different. The absent fact is in no document at all. Appendix C sets it against the related distinctions, which all assume the missing information exists and could be supplied.

Contract and policy NLP treats absence as a label to predict rather than a state to resolve (Kooreeda and Manning, 2021; Hendrycks et al., 2021; Ravichander et al., 2019), and provenance work audits license tags rather than reading the text to a decision (Longpre et al., 2024a; Kim et al., 2025; Chae et al., 2026). Closest is Scharrenberg and Sun (2025), who argue that compliance turns on qualitative standards while model development is governed by quantitative ones. They make that argument for EU law, and we measure it where the case law names what the qualitative standard looks to. Korean legal NLP has largely addressed public law (Hwang et al., 2022; Kim et al., 2024; Lee et al., 2025; Jung et al., 2026); private-law work reads adjudicated disputes (Cho et al., 2026) or

expert-answered consultations (Lee et al., 2026), not the instruments themselves.

Two things make the constitutive case measurable here. Korean case law on the scope of a license names the kinds of extrinsic evidence that resolve a grant of uncertain scope. That lets us fix the coding set before annotation and makes the gap enumerable. If the gap is constitutive, retrieved context should move where a screener sets its threshold far more than it improves how well the screener separates the two classes. We test that prediction with a discriminability statistic rather than with stratum accuracies, because discriminability and threshold are algebraically independent.

Throughout, the reference is not another system’s output or a crowd majority but one practicing Korean intellectual property lawyer’s judgment on the same instruments, recorded through the screening protocol that lawyer wrote for the purpose and that Appendix A reproduces in full. What we report is the distance between that judgment and what an automated screener produces, and where it comes from.

Our key questions and contributions are:

- The first question is what a license settles. With no text-and-data-mining exception in Korea, whether a corpus may be used for training turns on the license itself and not on a statutory safe harbor, so we formulate screening as a question about the document rather than the law: whether its text settles a stated use and, if not, which of five types of extrinsic fact, each drawn from Supreme Court of Korea authority, must be established outside it.
- The second question is what a screener can answer when the deciding facts are not in the document. Measuring 20 systems on over a hundred Korean licenses, terms of service

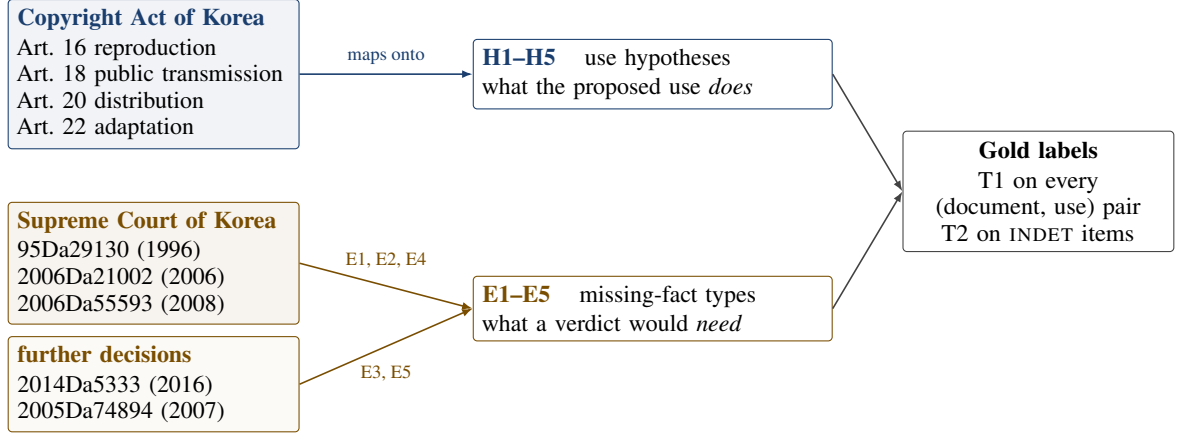


Figure 2: Legal Provenance of the Two Label Axes. Statutory rights fix what each use (H1–H5) implicates; Supreme Court decisions supply the missing-fact types (E1–E5), E1, E2, E4 as express factors and E3, E5 from further holdings (Appendix B; both sets in Table 1).

and standard-form contracts, across a protocol axis and a context axis, we find that neither axis improves how well systems separate settled from unsettled instruments, that none beats a supervised metadata lookup that ignores the text, and that the systems also abstain on questions the document does settle. The omitted facts still have to be established outside the document.

The single-jurisdiction scope is a precondition rather than a limitation: scope-of-grant doctrine is jurisdiction-specific, so labels are defensible only where the annotator is qualified in the governing law. We do not adjudicate infringement, decide fair use, or audit any dataset or provider, and the only prescription we offer is what to measure (§6).

2 Legal Background

Korean copyright law contains no provision permitting reproduction for text and data mining (TDM) (Ministry of Culture, Sports and Tourism and Korea Copyright Commission, 2026). Absent a license, the principal route to excusing training is the general fair-use provision, Article 35-5 of the Copyright Act, applied case by case, and, as of September 2026, no Korean court has ruled on whether AI training infringes (Ministry of Culture, Sports and Tourism and Korea Copyright Commission, 2026). Courts elsewhere have begun to rule on it, and other jurisdictions have legislated the TDM exceptions Korean law does not supply. Neither route is open here, which is why the question reaches Korean practitioners as one of license construction rather than of statutory exception. Where a statute supplies a TDM safe harbor a screening system can

Use hypotheses: what is proposed to be done		
H1	Training	reproduction; adaptation where the use transforms
H2	RAG indexing	reproduction; public transmission
H3	Output commerce	adaptation; whether a new form of exploitation
H4	Weight release	reproduction; public transmission (online) or distribution (tangible)
H5	Retention	reproduction for a limited term; retention set by the grant
Missing-fact types: what a verdict would need		
E1	Consideration	what was paid, and the circumstances of the bargain
E2	Usage of trade	whether the industry treats this use as separately licensed
E3	Rightsholder conduct	prior tolerance, technical measures, objections
E4	Market substitution	whether the use displaces the market the grant was made for, or opens a separate one
E5	Upstream title	whether the licensor held the right and could sublicense

Table 1: The Two Coding Sets. Closed for annotation, not the law: a further fact is marked CONTESTED.

often answer from the statute. In Korea the answer turns on the license, and where the license is silent, on facts about the transaction it does not record (Appendix B).

Korean case law names the evidence that resolves a grant whose scope is disputed. Where a license granted before a medium existed is silent on that medium, the Supreme Court held that the question turns on, among other things, whether the medium was unknown when the contract was made,

the parties’ knowledge, experience and economic position, usage of trade, whether the consideration is grossly disproportionate to a grant of that breadth, and whether the new medium displaces the existing market or creates a separate one.² Market substitution here is evidence of what the parties meant to grant, not the fourth fair-use factor of Article 35-5(2), which asks whether a use outside any grant should be excused. E4 codes only the first. Infringement and statutory limitation are outside the task. The five missing-fact types come from this body of authority. The set is closed as operationalized for annotation, not as a claim about the law: the decisions list their factors as illustrative, and foreseeability would on the law be a sixth (Appendix B). Figure 2 gives the provenance of each label axis. Which decision supplies which, and why foreseeability at contracting is carried as the revision-era stratum rather than as a sixth code, are in Appendix B. None of these facts is in the document. They are facts about a transaction and an industry, and a standard-form instrument, drafted before any particular counterparty exists, is not the kind of document that records them.

2.1 Choice of instruments

We work with six document types: public-sector open licenses (*Gonggongnuri*/KOGL, whose 2026 revision added an AI-specific type restricted to training use; [Ministry of Culture, Sports and Tourism and Korea Culture Information Service Agency, 2026](#)), national data-platform terms, the ministry’s standard-form contracts (*pyojun gyeyak-seo*), news-publisher terms, user-generated content platform terms, and data-marketplace terms, chosen so that the same legal question arrives in very different drafting cultures. Appendix B describes those cultures and how far each settles its items, and Appendix C places the task against the NLP literature.

3 Task and Corpus

3.1 Task definition

Given a document d and a use hypothesis h , T1 asks whether d alone settles whether h is permitted: DET or INDET. For INDET items, T2 asks which types of extrinsic fact are required, as a subset of a fixed set of five. Where a document does settle

the use, a secondary label records the direction it settles it in, permitted or prohibited. Table 1 gives both coding sets with the exclusive right each use implicates; E1, E2 and E4 come directly from the Supreme Court’s factor list for a grant of uncertain scope, and E3 and E5 rest on separate strands of the same authority. The five codes do not answer one question. E1–E4 bear on the meaning of the grant: what the parties agreed the license covers. E5 bears on its effectiveness: whether a grant that clearly covers the use binds the rightsholder, which it does not where the licensor lacked the right or the authority to sublicense. A document can settle the first and leave the second open. T1 is INDET in either case, because in either case a screener cannot clear the use from the text. Because T1 asks about information sufficiency rather than legality, it is closer to a question of fact, which is what the formulation is for. A document that explicitly prohibits the use is DET, because it has settled the question.

Each hypothesis is judged assuming the preceding uses are permitted, so that H4 assumes H1 is allowed. Without this rule the five items of a document become fully dependent and the effective sample collapses to the document count. Under it the items are not redundant: 73% of documents (76 of 104) do not carry the same T1 verdict across all five hypotheses, so the document-hypothesis pair is the unit on which a screening decision is taken.

A grant permitting commercial use of the work does not determine whether the outputs of a model trained on it may be commercialized, nor does a redistribution clause determine whether model weights may be distributed. Whether distributing weights amounts to redistributing the work is unsettled in Korean law (rule R4, Appendix A). Our annotation therefore treats H3 and H4 as determined only where the document addresses machine-learning use explicitly. The labels show what that costs a reader. Documents settle the two uses that act on the data about half the time (H1 50%, H2 47%), and settle the downstream uses less often and less evenly: commercializing outputs 38%, retention after training 22%, and distributing weights 19%, the least often settled of the five.

3.2 Corpus

104 Korean documents $\times$ 5 hypotheses give 520 items across the six instrument types of §2, all labeled and all scored. A second axis is revision era: last revised in 2021 or earlier (PRE) versus

²Supreme Court of Korea, 95Da29130, 30 July 1996; the facts of the case are in Appendix B. We use it for its test, not its outcome.

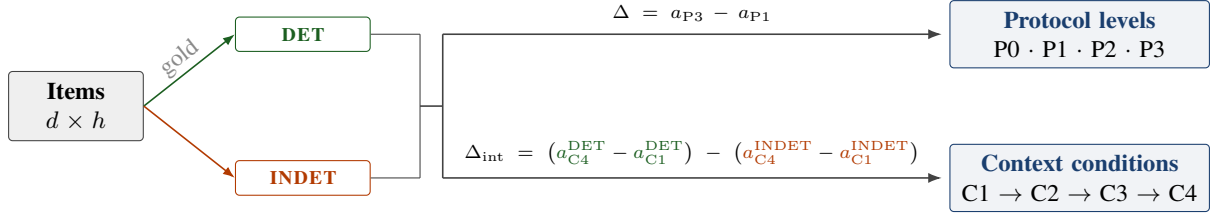


Figure 3: The Measurement Design. The protocol axis (P0–P3) varies only the written procedure, the context axis (C1–C4) only the text the model sees (§4); a is T1 accuracy, and only Δ_{int} is read per gold stratum.

2023 or later (POST), with 2022 left out as the year the question became salient and therefore the one year a date does not place a document on either side. We discard nothing to balance the strata: documents dated 2022 stay in the corpus and drop out only of PRE/POST comparisons. Era operationalizes the foreseeability factor: a document written before generative AI became salient can address these uses only through general language, which is the condition under which the question becomes indeterminate.

Era is not the same as whether a document says artificial intelligence: the terms in Table 2 have permitted use for artificial-intelligence research and development since 2018, seven years before the 2025 version quoted there, so a keyword filter would have flagged that 2018 text as addressing AI, and Table 7 sets out five archived versions spanning the transition. The base rate follows from coding rules fixed before any rate was computed: 35.2% of items are DET, 183 of 520. The corpus is assembled purposively rather than sampled, to span the drafting cultures of §2, so every rate we report is a rate within it and not an estimate of prevalence in Korean licensing at large.

Labels were assigned by one of the authors, a lawyer admitted to practice in the Republic of Korea whose practice is in intellectual property and includes copyright licensing and rights clearance, following the codebook reproduced in Appendix A. Every item was marked twice, the second pass two weeks later and blind to the first; the two agree on 92.3% of items and the rest were adjudicated a week after that (Appendix A.10). The codebook is not the annotator’s own taxonomy. Three of the five types are express factors in the three scope-of-grant decisions discussed in §2, and the other two rest on identified holdings within the same line of authority (Appendix B). A mark therefore records which factor the document leaves open under a stated rule rather than an unstructured impression. A second annotator working from the same authorities would be answering the same question; whether

they would reach the same answers is untested, and the line between determinacy and indeterminacy retains a normative element that one annotator cannot calibrate. Each item carries a three-level confidence rating and a CONTESTED flag, which record the annotator’s own difficulty signal; all 520 items are scored everywhere, and no item is removed on either ground.

4 Experiments

4.1 Protocol and context conditions

Figure 3 sets out the design. The protocol levels vary only the written procedure given to the model, holding model and document fixed, so the change in T1 accuracy from P1 to P3 separates what expert collaboration contributes from what prompting effort contributes. The context conditions vary only how much of the document and its surroundings the model sees, and we report that effect as the interaction Δ_{int} defined in §4.2: a gain confined to the DET stratum is the signature of retrieval finding a clause it had missed, a gain confined to the INDET stratum is the signature of a shifted threshold, and one flat curve is equally consistent with a model that could not use long inputs at all (Appendix D).

The four levels are the four realistic options a practitioner has. P0 is the incumbent regular-expression filter over permissive and restrictive phrases, with no model in the loop; P1 is a generic prompt asking whether the use is permitted, with only the output format specified, which is what an engineer writes without legal input; P2 adds the task definition, the DET/INDET distinction, the fixed inventory of missing-fact types and the antecedent-use assumption, but no procedure for applying them; and P3 is the gold-label lawyer’s own screening protocol serialized into the system prompt (Appendix A). Because P3’s author also produced the gold labels, the P1→P3 contrast is an upper bound on what writing the procedure down buys, not an estimate of it.

Document: a user-generated-content platform’s terms of service, in force July 2025

Grant, for AI R&D	“The content you provide may also be used by [the platform] for research and development in fields such as artificial-intelligence technology, in order to improve [its] services and to provide new ones, or used by [its] affiliates for the same purposes under delegation.”
Opt-out, and its exception	“Content that you have withheld from public access . . . will not be used for research and development in fields such as artificial-intelligence technology after the time of that action, except where it had already been processed, before that action, in a way that makes it technically difficult to distinguish.”

Annotation: five criteria, one mark each; T1 is derived, not annotated: any X $\Rightarrow$ INDET

Use hypothesis	E1	E2	E3	E4	E5	T1
H1 Use as training data	O+	.	O+	.	.	DET
H2 Index for retrieval-augmented generation	O+	.	O+	.	.	DET
H3 Commercialize the model’s outputs	O-	X	X	X	.	INDET
H4 Distribute the trained weights	X	X	X	X	.	INDET
H5 Retain the original after training	X	X	X	.	.	INDET

O+ judgeable, permissive O- judgeable, restrictive 0 judgeable, neutral X missing: needs a fact outside the document . not engaged by this use (N/A)

Table 2: A Worked Annotation on One Document. Training and indexing are settled permissively; outputs, weights and retention stay open (§3); upstream title is not engaged because the instrument allocates the risk of title within itself (Appendix A.3). Originals in Table 8.

4.2 Setup

Context is grown in four nested conditions: a BM25 retrieval over the document’s passages with the use hypothesis as the query (C1), the clause passages recovered for it (C2), C2 plus the governing statutes (C3), and C3 plus other documents from the same provider (C4). C1 never consults the gold label: retrieval may find the governing clause or miss it, and that is what it measures. Half (52) of the 104 documents have no other document from the same provider in the corpus, so on their items C4 coincides with C3. The conditions nest, so any improvement is attributable to the material added; we do not inject case law, which would supply the doctrinal derivation the labels themselves encode. The two axes are varied one at a time and cross at a single point. The protocol sequence of §4.1 holds context at C2, and the context grid holds the prompt fixed at the P2 wording throughout. The grid’s C2 column and the protocol table’s P2 column are therefore the same configuration issued twice. They agree on 93.4% of 9,858 paired verdicts, so a smaller difference should not be read as an effect. Away from that column the two tables vary different things, and their cells should not be read against each other. The estimand is the interaction between context condition and gold label, $\Delta_{\text{int}} = (a_{C4}^{\text{DET}} - a_{C1}^{\text{DET}}) - (a_{C4}^{\text{INDET}} - a_{C1}^{\text{INDET}})$, where a_c^g is T1 accuracy (whether the text settles the use) on gold stratum g under condition c , reported as an interval rather than as two separately tested curves. T2, on items the text leaves open, names which extrinsic facts are missing. Four controls each target

one alternative explanation: an oracle clause, no document at all, an abstention-encouraged prompt, and a nonsense-label prompt that replaces DET and INDET with arbitrary tokens while keeping their definitions (Ling et al., 2026).

We evaluate 20 systems: the keyword filter, open-weight models from 7B to 70B parameters, and frontier proprietary and open-weight systems reached through a common API. T1 uses balanced accuracy and MCC against an always-INDET baseline, together with the signal-detection pair d' and c (§5.2 gives both), which separate how well a system discriminates the two classes from where it sets its threshold. T2 uses mean Jaccard against two references, emitting all five codes and same-document lookup. A third reference, for T1, reads no document text and predicts T1 from the use hypothesis and revision era alone. A system that does not clear it is recovering metadata regularity rather than reading. Every interval is a paired bootstrap re-sampled over the underlying instrument rather than the document record, which leaves 97 clusters over 104 documents. We do not fine-tune: at this corpus size a supervised model would measure how well one annotator’s coding convention can be recovered, not how well a system screens. Appendix D gives the controls, the references, the clustering and that argument in full. Our primary error measure is the over-commitment rate, the share of gold-INDET items on which a system returns a DET verdict in either direction: each such verdict settles, with no record that a judgment was made, a question the document leaves open.

Model	Protocol	Bal. acc	MCC	O.c.	T2 Jac.
Keyword filter	P0 keyword	0.599	0.207	0.632	0.170
GPT-5.6 Sol	P1 naive	0.524	0.087	0.056	0.264
	P2 structured	0.603	0.227	0.181	0.367
	P3 expert	0.544	0.099	0.202	0.351
Claude Opus 5	P1 naive	0.530	0.058	0.481	0.156
	P2 structured	0.616	0.245	0.205	0.376
	P3 expert	0.571	0.143	0.285	0.341
Kimi K3	P1 naive	0.566	0.149	0.190	0.258
	P2 structured	0.584	0.198	0.148	0.367
	P3 expert	0.570	0.153	0.205	0.368
Solar-Pro4	P1 naive	0.568	0.206	0.065	0.258
	P2 structured	0.545	0.164	0.042	0.433
	P3 expert	0.559	0.219	0.030	0.476

Table 3: The Screening Protocol. Four systems at C2; all 19 in Table 11. Bal. acc is balanced accuracy, MCC is the Matthews correlation coefficient, O.c. is over-commitment (§4.2), and Jac. is the Jaccard index.

5 Results

5.1 Effect of the screening protocol

The keyword filter, with no model in it, performs above chance, and Table 3 does not order the protocol levels above it. Of the 57 model-and-level cells the protocol produces, 5 beat its balanced accuracy of 0.599 and 7 its MCC of 0.207. It reaches those figures by answering DET most of the time and committing on 0.632 of the unsettled items. The pipeline failure of Figure 1 scores best on the metric a benchmark would report. Enlarging its context only makes identifier matching commit more often: over-commitment reaches 1.000 at C3 and C4, because the statutory text added at C3 supplies a restrictive term the filter matches whatever the use in question.

Across the systems the protocol sequence covers, it barely touches the determinacy judgment. Mean balanced accuracy runs 0.547, 0.551, 0.560 from P1 to P3, a spread of 1.30 points, and the ordering inside that band is not stable: P3 is ahead of P2 on 11 systems and behind on 8. MCC and d' move as little. Over the roster the P1-to-P3 change in d' is +0.083 [−0.025, +0.203] while the criterion moves −0.035 [−0.114, +0.038]. Neither interval excludes zero, so at the roster level neither quantity moves by a measurable amount.

A protocol written by a model to the same specification as the lawyer’s changes balanced accuracy by −0.004 across 15 systems. What it changes is where the systems stop. The criterion moves +0.523, up on 13 of 15 ($p = 0.007$), while discriminability moves −0.094, down on 11 ($p = 0.118$). The threshold difference is several times larger than

the difference in discrimination, and balanced accuracy nets that to nothing. A length-matched placebo with no doctrinal content and the protocol with its steps shuffled both land inside the same range, moving balanced accuracy by −0.021 and +0.014 against P3 on the seven systems they were run on, so on this corpus nothing we can remove from the protocol is worth a measurable amount.

What the protocol does move is the second task. Mean T2 Jaccard rises monotonically, 0.244 → 0.363 → 0.383: the intervention that moves balanced accuracy by 1.30 points moves missing-fact agreement by 13.9. That rise happens entirely underneath a baseline with no model in it. Emitting all five types unconditionally scores 0.487, and one cell shown a document exceeds it, a control condition rather than any cell of Table 11 or Table 12. The other cells that clear the floor are ones we ran with the document withheld, where systems name all five types on almost every item. Pooling the other hypotheses’ gold sets for the same document scores 0.623, which no cell beats. Holding the second task to the standard of the first, we find that the ladder moves how many types a system names rather than which. 36% of indeterminate responses name all five. The code-frequency profile is a third as variable as the gold profile (0.108 against 0.311) while tracking its shape ($r = 0.96$). The naming reproduces the taxonomy’s marginals more than it discriminates between items. The sharpest case is upstream title, whose gold pattern is categorical rather than graded: the code is marked on none of the indeterminate items in the two adhesion-contract types or the public licenses, because those instruments allocate the risk of title within themselves. It is marked on +0.453 more of them in the three types where an intermediate holder sublicenses material contributed by others. Predictions recover +0.052 of that gap on average, and only 11 of 19 systems get the sign right, because most emit the code almost everywhere and a few almost nowhere. The structure is doctrinal and categorical, which makes its absence from the predictions hard to read as difficulty. Naming which extrinsic facts a verdict requires is a taxonomic judgment the coding set carries. Deciding whether the document settles the question is not.

The clearest effect of the protocol is neither accuracy nor bias but the narrowing of the spread between systems. On identical inputs, P1 over-commitment ranges from 0.056 to 0.680 across the 19 systems. At P2 that range is 0.347 wide, a nar-

rowing by a factor of 1.8, and at P3 it widens again to 0.466. The criterion does not track that range: it sits at -0.684 at P1 and -0.719 at P3 but reaches -1.021 at P2, where the range is narrowest. It stays on the opposite side of neutral from the keyword filter throughout. What the protocol secures is not correctness but a predictable operating point, and most of that comes from stating the task (P2) rather than from the full procedure (P3).

5.2 Context saturation

Table 12 reports every system at each context condition, and the number to compare against is not a constant answer but a lookup table. Predicting T1 from the use hypothesis and revision era alone (PRE, POST, or undated), reading none of the text, scores 0.651 balanced accuracy and 0.305 MCC under document-level leave-one-out. Of the 323 scored cells in this paper, 0 exceed it on balanced accuracy and 0 on MCC. Fitted on the other documents’ gold labels, it is a supervised bound on how much of the label survives without the text, not a like-for-like contest. Nothing we ran that reads the document beats it, and within the grid the best cell is 0.621 against 0.500 for answering INDET everywhere. The systems differ in which way they err, their over-commitment running 0.000 to 0.344. Decomposing the C1-to-C4 change by gold label separates the two things over-commitment merges. Over 17 systems the indeterminate stratum gains $+0.013$ on average, with 3 intervals excluding zero upward and 0 downward; the determinate stratum averages $+0.008$, with 3 up and 3 down. Only the indeterminate stratum gains directionally, which is the second signature of §4.1 and not the first: the determinate intervals split evenly, so its mean is not a gain the roster agrees on. The interaction averages -0.005 over a span of -0.177 to $+0.116$, excludes zero downward on 4 systems and upward on 4, and its sign is split 9–8 across the roster.

We exclude two systems, Llama-3.1-8B and Gemma-3-12B, answering INDET on at least 99% of items at C1 or C4: with the determinate stratum near-empty at that end, $\Delta d'$ rests on one or two rows.

None of that settles whether the systems read the enlarged context better. Δ_{int} equals $-(\Delta H + \Delta F)$ for hit rate H and false-alarm rate F , so it is orthogonal to $d' = z(H) - z(F)$. A threshold move and a sensitivity gain produce the same value. Measured directly over the roster, $\Delta d'$ is $+0.098$ $[-0.017, +0.228]$ and Δc is -0.044

$[-0.097, +0.012]$. Against criterion levels of -1.883 to -0.229 , even the upper end of the d' interval leaves every system too far from neutral for its verdicts to be relied on. A non-parametric $h - f$, which the probit cannot distort where the hit rate is near one, gives $+0.020$ $[-0.009, +0.053]$. That point sits inside the $\delta = 0.05$ margin, though the interval does not, which we read as the probit amplifying a threshold move. Per system the sign of $\Delta d'$ is positive on 14 of 17 ($p = 0.013$), but the roster interval still includes zero, so the direction is not a usable effect. The keyword baseline moves the opposite way on both indices, -0.594 and $+2.649$, so neither curve is a property of the corpus.

Response rates over the same 17 systems run 0.612–0.927 at C1 and 0.667–0.960 at C4, almost every cell above the base rate of 0.648, so enlargement is not correcting an over-confident prior. The criterion excludes zero downward on 5 systems and upward on 4, pushing some toward abstention and others toward commitment.

Scored per item, the failure is one-sided and falls on the items the document settles. Across every cell shown a document, gold-determinate items average 0.286 accuracy against 0.823 for indeterminate ones, and 25% of the determinate items are decided correctly by a tenth or fewer of those cells, spread over 25 of the 80 documents that carry one, against 0 indeterminate items.

What the systems write on those items is uniform enough to name. Taking the determinate items on which at least 85% of the roster abstains under the expert protocol at C2, 65% of the abstaining answers give the same ground: the instrument does not *explicitly* address machine-learning use. The lawyer’s marks treat this as limitation, not the silence of rule R1 (Appendix A.7): the grant is confined to enumerated purposes, so a license confined to operating a service settles training against permission. The systems ask the document to say the word and abstain when it does not. The pattern is not confined to the two hypotheses whose rule turns on explicit mention; it appears across all five and across every instrument type.

Nor can a pipeline separate right answers from wrong ones on these items by asking the systems how sure they are. The protocol asks for a confidence level, and 8 systems supply one on at least nine tenths of their answers. Over those 4,156 answers, accepting only the high-confidence ones costs a third of coverage and buys $+0.008$ accuracy

(0.651 against 0.643). The ordering does not even run the right way: the 1.6% of answers marked low confidence (68 answers) are the most accurate of the three levels, at 0.765, though too few to establish an inversion. A screening pipeline cannot read escalation off the model’s self-assessment, which is the practical form the first task’s difficulty takes.

Supplying the clause the lawyer relied on does not repair the failure on determinate items; it deepens it. The oracle arm replaces the document with that clause, or an explicit statement that none exists, so a failure to locate the provision should disappear there. Instead, determinate accuracy moves -0.040 and indeterminate accuracy $+0.052$, with determinate accuracy falling on 15 of 19 systems (two-sided sign test over the 17 that move, $p = 0.002$). Handed the provision that settles the question, the systems abstain more. That forecloses the two repairs a practitioner would try first. The clause was neither missed nor buried.

6 Discussion and Conclusion

Two failures sit beneath the aggregate numbers: systems commit on questions the document leaves open, and abstain on questions it settles. The first cannot be trained away, because the facts it turns on are in no document, so abstention has to be a first-class output. Systems whose only answers are permitted and prohibited resolve the INDET case of Figure 1 silently, in whichever direction their prior favors. A screening component should be evaluated on how often it escalates correctly. Naming the missing fact makes that escalation actionable. Because the codes come from doctrine, each tells the reviewer what to investigate, whether evidence from the parties, market practice, or the chain of title.

Expert collaboration changes what the screener writes, not what it knows. Neither stating the task nor writing the lawyer’s procedure down closes a measurable part of the distance to the reference lawyer’s judgment, and the gain in missing-fact agreement is in how many types get named, not which: almost no configuration reaches the all-five baseline, and none the same-document lookup. What stating the task buys is a predictable operating point, which a component specification can rely on. A fixed, doctrine-derived set of codes is the right output format, not yet a working one. Nothing we vary moves the distance to the expert, not scale, not context, not the protocol, and not even

the governing clause, which lowers determinate accuracy on 15 of 19 systems. Two ceilings sit one on the other. On indeterminate items no screener can settle the question, because the facts it turns on have to be established by further legal and factual investigation rather than read off the instrument. A lawyer directs that investigation but cannot resolve it from the document either. A web-search tool does not recover them either: usage of trade and a rightsholder’s course of conduct, including prior tolerance, are facts about a particular transaction rather than pages a query returns. Below that, on items the document does settle, the systems still fail, and that failure is theirs, not the instrument’s. The upper ceiling is a boundary a screener must recognize and route across. The lower one is a deficiency to close first.

Limitations

The corpus is 104 documents with correlated items, so intervals come from a document-level cluster bootstrap that can detect a sizeable interaction but not establish a small one. The metadata lookup, which no cell beats, bounds how regular the labels are; it is not a screener. On 3 systems the same justification text recurs on more than ten percent of unrelated items; we report these systems rather than exclude them. No second lawyer annotated the corpus, so we report no inter-expert agreement. Because the annotator also wrote the P3 protocol, the P1-to-P3 comparison is an upper bound on what a written protocol transfers (§4.1). The two blind passes agree on 92.3% of items, which measures one annotator’s stability over time rather than agreement across readers. The corpus consists of publicly posted instruments and excludes negotiated agreements that are never made public; such agreements may record the consideration and bargaining history that standard-form instruments do not, so the rates reported here do not carry over to them. Each document was retrieved once, and its labels describe that version, not the instrument as later amended. We read every instrument under Korean law; which law governs a license, and whether Korean copyright law reaches a training act performed abroad, are questions of private international law outside the task. A screener deployed in another jurisdiction would need the same exercise under the law that governs there, and that transfer is untested. Results also depend on model versions and prompts at evaluation time.

Ethical considerations

Data provenance. The documents are licenses, terms of service and standard-form contracts publicly posted by their issuers, retrieved once each with source, issuer and access date recorded. Beyond the short passages quoted for analysis in Table 8, with issuers replaced by their role, we redistribute no document text, name no private issuer and release no data artifact. Quoted passages may nonetheless be traceable to their issuer through public search. The labels characterize the text as written, not any issuer’s practice or intent. The corpus contains no personal data.

No legal advice. The labels are interpretations made for research of documents as written, not legal opinions on any transaction, and no attorney-client relationship arises from this work. A determinate label records that the instrument settles the question on its face, not that a court would so hold; an indeterminate label records that the text leaves the question open, not that the use is unauthorized. Nothing here should be relied on to decide whether any actual use is lawful.

Acknowledgments

The views expressed herein are those of the authors alone and do not represent the official position of their affiliated organizations or law firm. GPT-5.6 Sol, Claude Opus 5, and Gemini 3.7 Flash were used to improve the writing of this paper.

References

- Gustaf Ahndritz, Tian Qin, Nikhil Vyas, Boaz Barak, and Benjamin L. Edelman. 2024. [Distinguishing the knowable from the unknowable with language models](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 503–549. PMLR.
- Emily Allaway and Kathleen McKeown. 2025. [Evaluating defeasible reasoning in LLMs with DEFREASING](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10540–10558, Albuquerque, New Mexico. Association for Computational Linguistics.
- Anthropic. 2026. [System card: Claude Opus 5](#).
- Misha Benjamin, Paul Gagnon, Negar Rostamzadeh, Chris Pal, Yoshua Bengio, and Alex Shee. 2019. [Towards standardization of data licenses: The Montreal data license](#). *Preprint*, arXiv:1903.12262.
- Michael J. Bommarito II, Jillian Bommarito, and Daniel Martin Katz. 2025. [The KL3M data project: Copyright-clean training resources for large language models](#). *Preprint*, arXiv:2504.07854.
- Daniel Braun and Florian Matthes. 2024. [AGB-DE: A corpus for the automated legal assessment of clauses in German consumer contracts](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10389–10405, Bangkok, Thailand. Association for Computational Linguistics.
- Sungwon Chae, Keonwoo Kim, Hoki Kim, Jaeyeon Ju, and Sangchul Park. 2026. [Position: Current model cards are insufficient for downstream governance of open-weight foundation models](#). In *Forty-third International Conference on Machine Learning Position Paper Track*.
- Minhan Cho, Soyoung Park, S. Shyam Sundar, Daemin Choi, and Jinyoung Han. 2026. [KCC: Korean civil case dataset for legal information retrieval](#). In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’26, page 3117–3123, New York, NY, USA. Association for Computing Machinery.
- Danish Contractor, Daniel McDuff, Julia Katherine Haines, Jenny Lee, Christopher Hines, Brent Hecht, Nicholas Vincent, and Hanlin Li. 2022. [Behavioral use licensing for responsible AI](#). In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22, page 778–788, New York, NY, USA. Association for Computing Machinery.
- Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. [The power of noise: Redefining retrieval for RAG systems](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’24, page 719–729, New York, NY, USA. Association for Computing Machinery.
- DeepSeek-AI. 2026. [DeepSeek-V4: Towards highly efficient million-token context intelligence](#). *Preprint*, arXiv:2606.19348.
- Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Babu Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A Huerta, and Hao Peng. 2025. [Context length alone hurts LLM performance despite perfect retrieval](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23281–23298, Suzhou, China. Association for Computational Linguistics.
- Ran El-Yaniv and Yair Wiener. 2010. [On the foundations of noise-free selective classification](#). *Journal of Machine Learning Research*, 11(53):1605–1641.

- Yonatan Geifman and Ran El-Yaniv. 2017. [Selective classification for deep neural networks](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Gemma Team. 2025. [Gemma 3 technical report](#). Preprint, arXiv:2503.19786.
- Google DeepMind. 2026. [Gemini 3.7 Flash model card](#).
- Luke Guerdan, Solon Barocas, Ken Holstein, Hanna Wallach, Steven Wu, and Alexandra Chouldechova. 2025. [Validating LLM-as-a-judge systems under rating indeterminacy](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Lena Held and Ivan Habernal. 2025. [Contemporary LLMs struggle with extracting formal legal arguments](#). In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 292–303, Suzhou, China. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. [CUAD: An expert-annotated NLP dataset for legal contract review](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. [Training compute-optimal large language models](#). Preprint, arXiv:2203.15556.
- Elias Horner, Cristinel Mateis, Guido Governatori, and Agata Ciabattoni. 2025. [Toward robust legal text formalization into defeasible deontic logic using LLMs](#). Preprint, arXiv:2506.08899.
- Abe Hou, William Jurayj, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2024a. [Gaps or hallucinations? scrutinizing machine-generated legal analysis for fine-grained text evaluations](#). In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 280–302, Miami, FL, USA. Association for Computational Linguistics.
- Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2024b. [Decomposing uncertainty for large language models through input clarification ensembling](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 19023–19042. PMLR.
- Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekish, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. [RULER: What’s the real context size of your long-context language models?](#) In *Conference on Language Modeling (COLM)*.
- Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. [A multi-task benchmark for Korean legal language understanding and judgement prediction](#). In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Joseph Marvin Imperial, Daniel Slate, and Noam Kolt. 2026. [ScrapeBench: Evaluating legal compliance of AI agents in website scraping](#). In *ICML 2026 Workshop on AI for Law*.
- Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. 2025. [Sufficient context: A new lens on retrieval augmented generation systems](#). In *The Thirteenth International Conference on Learning Representations*.
- JiHyeok Jung, TaeYoung Yoon, and HyunSouk Cho. 2026. [CALRK-Bench: Evaluating context-aware legal reasoning in Korean law](#). Preprint, arXiv:2603.26332.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. [Scaling laws for neural language models](#). Preprint, arXiv:2001.08361.
- Jaekyeom Kim, Sungryull Sohn, Gerrard Jeongwon Jo, Jihoon Choi, Kyunghoon Bae, Hwayoung Lee, Yongmin Park, and Honglak Lee. 2025. [Do not trust licenses you see: Dataset compliance requires massive-scale AI-powered lifecycle tracing](#). Preprint, arXiv:2503.02784.
- Najoung Kim, Phu Mon Htut, Samuel R. Bowman, and Jackson Petty. 2023. [\(QA\)²: Question answering with questionable assumptions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8466–8487, Toronto, Canada. Association for Computational Linguistics.
- Yeeun Kim, Youngrok Choi, Eunkyung Choi, JinHwan Choi, Hai Jin Park, and Wonseok Hwang. 2024. [Developing a pragmatic benchmark for assessing Korean legal language understanding in large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5573–5595, Miami, Florida, USA. Association for Computational Linguistics.
- Kimi Team. 2026. [Kimi k3: Open frontier intelligence](#). Preprint, arXiv:2607.24653.
- Michael Kirchhof, Gjergji Kasneci, and Enkelejda Kasneci. 2025. [Position: Uncertainty quantification needs reassessment for large language model agents](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 81665–81677. PMLR.

- Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. 2025. [AbstentionBench: Reasoning LLMs fail on unanswerable questions](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Yuta Koreeda and Christopher Manning. 2021. [ContractNLI: A dataset for document-level natural language inference for contracts](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jihyung Lee, Daehui Kim, Seonjeong Hwang, Hyounghun Kim, and Gary Lee. 2025. [KoBLEX: Open legal question answering with multi-hop reasoning](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4019–4053, Suzhou, China. Association for Computational Linguistics.
- Yongtae Lee, Surin Lee, Sumin Kim, S M Wahidur Rahman, and Heung-No Lee. 2026. [KoLegalQA: A Korean legal QA dataset for trustworthy and explanation-grounded legal AI](#). In *Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026)*, pages 240–255, San Diego, California. Association for Computational Linguistics.
- Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. [Same task, more tokens: the impact of input length on the reasoning performance of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Bangkok, Thailand. Association for Computational Linguistics.
- Belinda Z. Li, Been Kim, and Zi Wang. 2025. [QuestBench: Can LLMs ask the right question to acquire information in reasoning tasks?](#) In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Zipeng Ling, Shuliang Liu, Yuehao Tang, Junqi Yang, Shenghong Fu, Seonil Son, Chen Huang, Kejia Huang, Yao Wan, Zhichao Hou, and Xuming Hu. 2026. [LLM abstention can be a prompt artifact, in addition to genuine uncertainty](#). *Preprint*, arXiv:2507.16199.
- Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019. [CLAUDETTE: An automated detector of potentially unfair clauses in online terms of service](#). *Artificial Intelligence and Law*, 27(2):117–139.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Shuang Liu, Zelong Li, Ruoyun Ma, Haiyan Zhao, and Mengnan Du. 2025. [ContractEval: Benchmarking LLMs for clause-level legal risk identification in commercial contracts](#). In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 291–291, Suzhou, China. Association for Computational Linguistics.
- Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. 2024a. [A large-scale audit of dataset licensing and attribution in AI](#). *Nature Machine Intelligence*, 6(8):975–987.
- Shayne Longpre, Robert Mahari, Ariel N. Lee, Campbell S. Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole J Hunter, Kevin Klyman, Christopher Klamm, Hailey Schoelkopf, Nikhil Singh, Manuel Cherep, Ahmad Mustafa Anis, An Dinh, Caroline Shamiso Chitongo, Da Yin, and 30 others. 2024b. [Consent in crisis: The rapid decline of the AI data commons](#). In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Subhabrata Majumdar. 2026. [Limits of reliability and scaling in language models](#). *Preprint*, arXiv:2607.14112.
- Meta AI. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Meta AI. 2025. [Llama 4 model card](#).
- Ministry of Culture, Sports and Tourism and Korea Copyright Commission. 2026. [Guide on applicability of the fair use doctrine to training of generative AI models](#).
- Ministry of Culture, Sports and Tourism and Korea Culture Information Service Agency. 2026. [KOGL AI type: Conditions of use for artificial-intelligence training data](#).
- Mistral AI. 2026. [Remote agents in Vibe. Powered by Mistral Medium 3.5](#).
- NAVER Cloud. 2025. [HyperCLOVA X SEED think 14B](#).
- OpenAI. 2026. [GPT-5.6 system card](#).
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.
- Barbara Plank. 2022. [The “problem” of human label variation: On ground truth in data, modeling and evaluation](#). In *Proceedings of the 2022 Conference*

- on *Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Qwen Team. 2025a. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Qwen Team. 2025b. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. [Question answering for privacy policies: Combining computational and legal perspectives](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4947–4958, Hong Kong, China. Association for Computational Linguistics.
- Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. [Thinking like a skeptic: Defeasible inference in natural language](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online. Association for Computational Linguistics.
- Noah Scharrenberg and Chang Sun. 2025. [Copyright infringement by large language models in the EU: Misalignment, safeguards, and the path forward](#). In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 125–134, Suzhou, China. Association for Computational Linguistics.
- Aakash Sen Sharma, Debdeep Sanyal, Priyansh Srivastava, Sundar Atreya H., Shirish Karande, Mohan Kankanhalli, and Murari Mandal. 2025. [Nine ways to break copyright law and why our llm won’t: A fair use aligned generation framework](#). *Preprint*, arXiv:2505.23788.
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. [Large language models can be easily distracted by irrelevant context](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.
- Jingwen Tan, Gopi Krishnan Rajbahadur, Zi Li, Xiangfu Song, Jianshan Lin, Dan Li, Zibin Zheng, and Ahmed E. Hassan. 2025. [LicenseGPT: A fine-tuned foundation model for publicly available dataset license compliance](#). In *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering*, FSE Companion ’25, page 39–50, New York, NY, USA. Association for Computing Machinery.
- Upstage. 2026. [Solar Pro 4: The agentic model that finishes the job](#).
- Samuel J. Vincent, Daniel D. Calloway, Fangyi Yu, Andrew M. Bean, and Nabeel Seedat. 2026. [InsufficiencyBench: Evaluating LLM legal advice on underspecified user queries](#). In *ICML 2026 Workshop on AI for Law*.
- Yusuke Watanabe, Yohei Kobashi, Takeshi Kojima, Yusuke Iwasawa, Yasushi Okuno, and Yutaka Matsuo. 2026. [ClinDet-bench: Beyond abstention, evaluating judgment determinability of LLMs in clinical decision-making](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 681–703, San Diego, California, USA. Association for Computational Linguistics.
- Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. [Know your limits: A survey of abstention in large language models](#). *Transactions of the Association for Computational Linguistics*, 13:529–556.
- xAI. 2026. [Model card: Grok 4.6](#).
- Eric Xia, Karthik Srikumar, Keshav Karthik, Advait Renjith, and Ashwinee Panda. 2025. [Beyond the haystack: Sensitivity to context in legal reference recall](#). In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 48–53, Suzhou, China. Association for Computational Linguistics.
- Robin Young. 2026. [Epistemic uncertainty is not the reducible kind](#). *Preprint*, arXiv:2606.12646.
- Zhipu AI & Tsinghua University GLM-5-Team. 2026. [GLM-5: from vibe coding to agentic engineering](#). *Preprint*, arXiv:2602.15763.

A Screening Protocol (P3)

This appendix reproduces the P3 protocol in full. The same document exists in two forms: an annotator-facing form, which governed the production of the gold labels, and a form serialized into a system prompt and evaluated as level P3 of §4.1. They differ in five respects, none doctrinal. Where the text below describes only one form it says so, and §A.9 lists every difference. The doctrinal derivation of the criteria is in Appendix B and is not repeated here. We first set out the other levels and the keyword filter so that the comparison in §5.1 can be read from this appendix alone.

The other levels. P0 stands in for what industry calls license scanning: it matches text against fixed term lists and presupposes that a match settles the question, and it is the cheapest option available. Two properties of the gold labels limit it. The items a document does settle are settled against the use nearly twice as often as for it (118 prohibitions to 64 permissions among the 182 of 183 determinate items that carry a direction; the remaining one records none because all its judgeable marks are neutral), so permissive language marks the minority case. The same document also routinely settles one use while leaving another open, which a rule that ignores the use hypothesis can express only through the passages retrieved for each hypothesis. P1’s system prompt is “you are an assistant with legal expertise; decide whether this use is permitted” (translated from the Korean). P2 adds the task definition, the DET/INDET distinction, the five missing-fact types and the antecedent-use assumption of R3, with no procedure for applying them. P3 serializes a use-to-rights mapping, an ordered procedure, and five decision rules, chiefly that silence is not permission and that a general commercial grant settles neither commercialization of model outputs nor distribution of weights.

The keyword filter in full. P0 is three term lists and a four-branch rule. Each term is given with a gloss; the filter matches the Korean, not the gloss. A term matches wherever it occurs, so an entry cut short matches any ending, and an ellipsis marks intervening text of at most four characters in the permissive entry and twenty in the restrictive one. The machine-learning list is 인공지능 (artificial intelligence), 기계학습 (machine learning), 텍스트·데이터 마이닝 (text and data mining) and 자동화된 수집 (automated collection); the

permissive list is 영리·비영리 (commercial or non-commercial), 영리... 목적을 불문 (whatever the commercial purpose), 상업적 용도 (commercial use), 상업적으로 이용할 수 있 (may be used commercially), 자유롭게 이용 (use freely) and 제한 없이 이용 (use without restriction); the restrictive list is 비영리 목적으로만 (for non-commercial purposes only), 영리를 목적으로... 동의 (consent, where the purpose is commercial), 할 수 없다 (may not), 금지 (prohibited), 파기하여야 (shall destroy) and 사전 서면 동의 (prior written consent). If any machine-learning term matches, the filter returns DET, prohibited where a restrictive term also matches and permitted otherwise. Failing that, a permissive match with no restrictive match returns DET-permitted, any restrictive match returns DET-prohibited, and no match at all returns INDET with all five codes. The lists were fixed before evaluation and were not tuned against the gold labels. Two restrictive entries, 할 수 없다 (may not) and 금지 (prohibited), are general enough to match almost any prohibition in a Korean instrument, including prohibitions directed at something other than the use in question; that generality is part of why the filter commits so readily.

P3 as an upper bound. The P1→P3 contrast measures how much of the gap to a qualified human closes when the procedure is written down. Because P3’s author also produced the gold labels it is an upper bound rather than an estimate, and a more generous one than shared authorship alone would make it: R1 through R5 are the rules that generate the labels, so serializing them discloses part of the labeling function itself, which a protocol written by a different lawyer would not contain. That even this ceiling moves balanced accuracy by 1.30 points (§5.1) makes the null result stronger, not weaker.

A.1 What is being decided

Given a document d and a use hypothesis h , decide whether the text of d settles whether h is permitted.

- DET: the text settles it, in either direction.
- INDET: the text does not settle it; facts outside the document are required.

This is a question about the document, not about the law. A document that expressly prohibits the use is DET: it has settled the question, and the

answer is no. Do not ask whether the use would ultimately be lawful.

In the annotator-facing form, T1 is never recorded directly; it is derived from the five criterion marks (§A.5). The serialized P3 prompt asks the system for T1 and the marks in one response, and scoring uses the reported T1 (§A.9).

A.2 Use hypotheses

Each hypothesis names a distinct act and implicates a distinct exclusive right, though the two ingestion uses often fall together in practice: H1 and H2 receive the same T1 verdict on 95 of 104 documents, whereas H1 and H5 agree on 45. Table 4 lists them.

	Use	Exclusive right implicated
H1	Use the work as training data	reproduction; adaptation where the use transforms
H2	Index the work for retrieval-augmented generation	reproduction; public transmission
H3	Commercialize the outputs of the trained model	adaptation; whether a new form of exploitation
H4	Distribute the trained weights	reproduction; public transmission where released online, distribution only for tangible copies
H5	Retain the original after training concludes	reproduction made for a limited term or purpose; whether it may be kept is set by the grant

Table 4: Use Hypotheses. The five hypotheses and the exclusive right each implicates, as the protocol states them.

A.3 Missing-fact types

For each criterion, ask one question: must I look outside this document to resolve it? Table 5 names the five types and what a verdict on each would need.

As §2 notes, the set is closed for annotation, not exhaustive of the law. An annotator who believes a sixth kind of fact is needed marks the item `CONTES` rather than inventing a code.

E5 concerns the effectiveness of the grant rather than its scope: whether the licensor could grant what the document says it grants. It is not engaged where the instrument allocates the risk of title within itself: adhesion terms that require the user who posts content to hold the rights it grants, and public open licenses whose issuer holds the rights it licenses. It is engaged where an interme-

diat holder sublicenses material contributed by others without warranting title. The test is a coding convention about what a screener can act on at the level of the instrument, not a claim that allocating the risk of title cures a defect in it: such a clause binds the parties, not a true rightsholder. The resulting pattern is categorical rather than graded: the code is marked on none of the indeterminate items in the two adhesion-contract types (platform and news-publisher terms) or in the public open licenses, against 69.2% in data-marketplace terms, 41.3% in national data-platform terms and 35.3% in the ministry’s standard-form contracts.

Foreseeability at contracting is a factor in the governing case law but is deliberately not a code, because it is the one factor a reader can approximate from the document itself, through its revision date. It enters as the revision-era stratum (§3.2).

	Type	What a verdict would need
E1	Consideration and bargaining structure	what was paid, and the circumstances of the bargain
E2	Usage of trade	whether the industry treats this use as separately licensed
E3	The rightsholder’s course of conduct	prior tolerance, technical measures, objections
E4	Market substitution or new-market creation	whether the use displaces an existing market or opens one
E5	Upstream title	whether the licensor held the right and could sublicense

Table 5: Missing-Fact Types. The coding set, as used for annotation (§A.3).

A.4 Notation

Each item carries five marks, one per criterion in order, on one line; each mark takes one of the five values in Table 6. For example, `0+ . X X .` reads: E1 permissive, E2 not engaged, E3 and E4 missing, E5 not engaged.

Mark	Meaning
0+	judgeable from the document; weighs toward permission
0-	judgeable from the document; weighs against permission
0	judgeable from the document; neutral
X	missing; needs a fact outside the document
.	not engaged by this use

Table 6: Criterion Marks. Five marks record an item; T1 follows from them (§A.5).

A.5 Derivation of T1

Any X yields INDET. *Otherwise* DET. A determinate item's direction is prohibited if any mark is 0-, permitted if any mark is 0+ and none is 0-, and unrecorded if all judgeable marks are 0.

At least one criterion must carry a judgeable mark (0, 0+ or 0-) or X. An item on which all five are marked not-engaged is a recording error, not a determinate item; none occurs in the corpus.

A.6 Procedure

1. Read the whole document first. Fix three things before judging any hypothesis: the last revision date; what the granting clause actually licenses (reproduction, transmission, distribution, or an undifferentiated "use"); and any purpose limitation attached to the grant.
2. Take one hypothesis. Eliminate criteria not engaged by it (.).
3. For each remaining criterion, ask whether resolving it requires a fact outside the document. If so, mark X.
4. For criteria that are judgeable, mark direction (0+, 0- or 0).
5. Record the governing provision and a one-clause justification. Record confidence, and flag **CONTESTED** where a qualified reader could reasonably differ.
6. Repeat for the next hypothesis under R3.

A.7 Decision rules

R1. Silence is not permission. A document that does not address a use has not permitted it. Absence of prohibition is not a grant. Where the document is silent, the criteria that would resolve the question are X, not 0.

R2. Extrinsic knowledge may never produce a judgeable mark. If resolving a criterion requires what the reader knows about the industry, the parties, or the market, that criterion is X by that very fact. The label records what the document settles, not what the reader knows.

R3. Each use is judged assuming the antecedent uses are permitted. H4 is judged on the assumption that H1 is allowed. Without this rule the five items of a document become fully dependent and the effective sample collapses to the document count.

R4. A general commercial grant settles neither H3 nor H4. Permission to use the work commercially does not by itself determine whether outputs of a model trained on the work may be commercialized; a redistribution clause about the work does not determine whether weights may be distributed, because whether distributing weights redistributes the work is unsettled. H3 and H4 are DET only where the document addresses machine-learning use explicitly.

R5. Explicit prohibition is DET. The question is whether the text decides, not whether the answer is favorable.

The rules do not state how a purpose limitation bears on uses outside it. The annotator treated a grant confined to enumerated purposes as settling other uses against permission (§5.2); a protocol that stated this might move the determinate stratum, which we did not test.

A.8 Confidence and **CONTESTED**

Each item carries a three-level confidence rating (H/M/L) and a **CONTESTED** flag. An item on which a qualified reader could reasonably reach a different mark is flagged **CONTESTED**. Flagged items are not removed: all 520 items are scored in every condition. In the annotator-facing form only, an item unresolved after three minutes is marked L and flagged rather than deliberated further: the difficulty is itself a datum, and forcing it into a confident mark would corrupt the label.

A.9 What changed in serialization

The annotator-facing protocol and P3 differ in five respects, none doctrinal: the annotator-facing text addresses the reader as "you", while P3 is phrased as instructions to the system; P3 specifies the output as JSON with named fields; P3 drops the three-minute rule (§A.8), which has no analog for a model; P3 numbers the procedure of §A.6, which the annotator-facing form presents as a single reading pass over steps 2–4; and the annotator derives T1 from the five marks (§A.5), while P3 asks the system for T1 and the marks in one response. Scoring uses the reported T1. Estimating transfer directly, rather than bounding it as above, would need a second qualified annotator.

A.10 Annotation provenance

All items were annotated by one of the authors, a lawyer admitted to practice in the Republic of Korea and working in intellectual property, whose

practice includes copyright licensing and rights-clearance work, following this protocol.

Annotation ran in three stages. The first pass marked every item. A second pass two weeks later re-marked every item from this protocol alone, with the first pass hidden, so that a mark reproduced only if the protocol produced it twice. The two passes agree on 92.3% of items. Those that disagreed were held for a third sitting a week later, at which the annotator saw both marks and settled the final label. Those adjudicated marks are the gold we score against. The design measures stability over time rather than agreement across readers. It does not substitute for a second qualified reader. It shows where the protocol underdetermines a mark for this annotator, not whether another Korean copyright lawyer would mark the same item the same way.

Documents that are not licensing instruments (statutes, administrative rules, official guidance) retrieved by the same search were marked out of scope by the annotator rather than deleted, and none of them is among the 104 annotated documents.

B Legal Background in Detail

§2 states the law we rely on; this appendix gives the derivation. Figure 2, in §2, sets out where each label axis comes from. U.S. readers will recognize cognates in UCC §1-303: usage of trade appears in the Supreme Court’s list, and course of performance and course of dealing correspond to the post-contract conduct that later decisions weigh (E3, below). The market-substitution factor resembles the fourth fair-use factor in wording only (§2).

No Korean TDM exception as of September 2026. Korean bills creating a text-and-data-mining exception have been introduced repeatedly since 2021 and none has been enacted; the Ministry of Culture, Sports and Tourism’s 2026 guidance, which names the EU and Japan among the jurisdictions that have adopted such an exception, analyzes training in Korea under the general fair-use provision instead ([Ministry of Culture, Sports and Tourism and Korea Copyright Commission, 2026](#)). Nor does the framework act on artificial intelligence change this. That act, in force since January 2026, contains no provision on copyright, reproduction or licensing. Its provisions touching training data are industrial-policy measures and, for operators of high-impact systems only, a duty to establish measures for explaining, as far as tech-

nically feasible, an overview of the training data used.³

The same question in litigation. A dispute over a syndication agreement concluded before generative models existed would reach a court in exactly this form: whether the agreement reaches model training would be resolved by evidence about how it was made, which is the dependence on extrinsic fact our task asks a model to recognize. Such agreements are rarely published, so a screener could not read the instrument, let alone the facts behind it.

Foreseeability as a stratum. One factor of the Supreme Court’s list is deliberately not among our codes: all three decisions name foreseeability at contracting as a consideration (95Da29130 asks whether the medium was unknown when the contract was made, 2006Da21002 and 2006Da55593 whether the use was foreseeable when the license was granted), so on the law it would be a sixth code. It is the only factor of the set a reader can approximate from the document itself, through its revision date, so coding it would mark as missing something the document’s date already approximates. We carry it as the revision-era stratum (§3.2) instead, which predicts that documents last revised before generative AI became salient leave more items open than later ones. The labels are consistent with that: PRE documents settle 24% of their items against 39% for POST ones (Table 9), though the annotator saw each revision date before marking (§A.6).

Two drafting cultures. The six instrument types of §2 are chosen so that the same legal question arrives in very different drafting cultures. In a ministry standard-form contract the parties tick which of the author’s economic rights are licensed, and the form typically has no field for the manner or medium of use, while a platform’s terms grant broadly and in prose. Both often leave AI training undetermined, for different reasons. The labels separate the cultures quantitatively: public open licenses settle 47% of their items, the most of any type, and ministry standard-form contracts 20%, the least, because a ticked reproduction right does not say whether it reaches a use the parties did

³Framework Act on the Development of Artificial Intelligence and the Creation of a Foundation for Trust, Act No. 20676 of 21 January 2025 (in force 22 January 2026), as amended by Act No. 21311 of 20 January 2026 (in force 21 July 2026).

not contemplate. The four remaining types fall between, from 35% to 37%.

One default rule does exist, and it is narrower than it looks. Where a contract does not express whether copyright was assigned or merely licensed, 95Da29130 presumes a license rather than an assignment, with rights reserved to the author. That presumption resolves which kind of contract the document is; it does not resolve whether a particular new use falls inside a grant whose existence is not in doubt. That second question is the one the factor test of §2 is for, and the one our items ask.⁴

Three of the five missing-fact types are drawn from this line, and the line consists of three decisions rather than one: 95Da29130 on the scope of a license over a new medium, 2006Da21002 on how the scope of a license is construed, and 2006Da55593 on the same question together with the burden of proving a right to use.⁵ E1, E2 and E4 appear as express factors in all three. In 95Da29130 itself a 1984 license of reproduction in “all recordings” was held to reach compact disc, the court weighing among other things that the medium was already commercialized abroad, that the consideration was not grossly disproportionate, and that it substituted for the LP market rather than opening a separate one.

The rightsholder’s course of conduct (E3) rests on a different strand of authority. It is not among the new-medium factors of 95Da29130. That decision admits post-contract conduct only for a different question, reading the authors’ later, unopposed contracting with another company as evidence that the grant was non-exclusive. Two later decisions weigh post-contract conduct directly: 2014Da5333 the rightsholder’s own failure to object, which is what E3 codes, and 2006Da55593 the licensee’s later dealings with other authors, which shows the court looking past the instrument to how the parties behaved.⁶

⁴Breach of the agreed method or conditions of use does not of itself infringe the reproduction right (Supreme Court of Korea, 2015Da1017, 23 November 2017), so whether a use falls outside the scope of the grant, rather than breaching a condition of it, is the question on which infringement exposure first turns.

⁵Supreme Court of Korea, 95Da29130, 30 July 1996; 2006Da21002, 22 December 2006; 2006Da55593, 24 April 2008.

⁶Supreme Court of Korea, 2006Da55593, *supra* (how the licensee acquired rights from other authors after contracting); 2014Da5333, 17 August 2016 (post-delivery conduct weighed; transfer of rights in a derivative work does not of itself transfer rights in the underlying work).

Upstream title (E5) likewise has case-law footing. A phonogram producer’s neighboring right is independent of the author’s economic right, yet reproducing the phonogram necessarily reproduces the works recorded on it. The producer therefore cannot reproduce or distribute without the author’s license, and the party asserting a right to use bears the burden of proving it.⁷ The same structure appears when a platform licenses to others material its users posted: the platform is then the intermediate holder, and its terms with users, such as those in Table 2, are the first-tier grant beneath it (Appendix A.3).

Infringement and statutory limitation arise only once a use falls outside the grant, and the task does not reach them. We note them to explain why, in Korean practice, the license rather than a defense carries the weight. Nor does Korean fair use make the license question moot. Where a use is defended as transformative, the Supreme Court has held that a court may weigh whether the degree of transformation exceeds the level required to create a derivative work. Construing the quotation exception of Article 28, the Supreme Court has held that a commercial purpose considerably narrows the range of free use.⁸ The transformativeness argument that carries much of the weight in United States litigation is therefore likely to be measured in Korea against the derivative-work threshold rather than treated as sufficient on its own. Nor does the private-use exception. Copying a work for business use inside a company is not reproduction for private use, which makes that exception an unlikely basis for treating internal corpus assembly as permitted.⁹

The framework is neither dated nor confined to new media. A 2007 decision restated the factor test as the general standard for construing the scope of a license. A 2008 decision applied it outside the new-medium setting, to what a party must show to establish a right to use at all. A 2013 judgment still cites the 1996 decision for its distinct assignment-or-license presumption.¹⁰

⁷Supreme Court of Korea, 2006Da21002, *supra*, holding [1]; 2006Da55593, *supra*, holding [1]; 2005Da74894, 22 February 2007.

⁸Supreme Court of Korea, 2021Da272001, 11 July 2024; 2012Do10786, 26 August 2014.

⁹Supreme Court of Korea, 2011Do5835, 15 February 2013, holding that in-house copying for business use is not private-use copying within the meaning of Article 30 of the Copyright Act, the private-use exception.

¹⁰Supreme Court of Korea, 2005Da74894, 22 February 2007; 2006Da55593, *supra*; 2011Do5835, 15 February 2013.

In force	Era	Names AI	Opt-out
2014 (i)	PRE	no	—
2014 (ii)	PRE	no	—
2018	PRE	yes	no
2024	POST	yes	no
2025	POST	yes	yes

Table 7: Revision History of One Instrument. Five archived versions of the terms in Table 2: whether each names AI and offers an opt-out.

One instrument across the transition. A clause permitting use for research and development “in fields such as artificial-intelligence technology” first appears in the 2018 version of the terms annotated in Table 2, before generative pretraining was a practical question. That is why the keyword is not a proxy for the era. What the 2025 revision adds is an opt-out, together with a carve-out excluding content already processed into a form that is “technically difficult to distinguish”. We annotate the 2025 and 2018 versions, which form one cluster in the bootstrap of Appendix D; the other three versions are not annotated.

Korean originals. Table 8 gives every instrument passage quoted in this paper, in the original with our translation below it, so that either reading can be checked against the other.

C Related Work

Contract and policy document NLP treats absence as something to detect rather than resolve. ContractNLI (Koreeda and Manning, 2021) makes non-mention a terminal label, CUAD (Hendrycks et al., 2021) extracts clause types, CLAUDETTE (Lippi et al., 2019) detects unfair clauses, and PrivacyQA (Ravichander et al., 2019) types policy silence as a cause of unanswerability. Absence of evidence is a label to predict there, not a state to resolve. Hou et al. (2024a) separate gaps from errors. The closest structural analog in another language is AGB-DE (Braun and Matthes, 2024), a corpus of German consumer standard-form contract clauses with expert legal assessments; there too the assessment is of the clause as drafted, not of whether the instrument determines a use its drafter did not contemplate. The over-abstaining direction we report is not peculiar to us: clause-level risk identification finds open models over-emitting “no related clause” (Liu et al., 2025). The null has precedent in this setting as well: Held and Habernal (2025) find that neither fine-tuning nor prompting beats a domain-pretrained encoder on legal argument extraction,

and that supplying annotation guidelines did not help.

Data provenance and license compliance work approaches the same datasets from the metadata side. Audits of license metadata (Longpre et al., 2024a), terms-of-service restriction trends (Longpre et al., 2024b) and provenance tracing (Kim et al., 2025) establish that surface license information is unreliable, and Chae et al. (2026) that standard open-source licenses sit poorly with open-weight foundation models and can undermine the enforceability of acceptable-use policies. Nearest to us, Tan et al. (2025) fine-tune a model on dataset license text to predict a commercial-use verdict with an explicit *unclear* class, and Imperial et al. (2026) find that agents comply with scraping restrictions on fewer than five percent of prohibiting sites while over-refusing on permissive ones. LicenseGPT reads the text and can decline, which is what makes it the closest system. Three things separate the task. It asks one question, where a single instrument in our corpus settles one of our five uses and leaves another open. Its *unclear* is a terminal bucket, where ours is decomposed into a fixed, doctrine-derived inventory of what a verdict would require. It is offered as a tool that speeds a lawyer up, where we measure where the instrument cannot decide. The remaining provenance work audits license tags or proposes better ones. Responses to unreliable metadata include new license instruments (Benjamin et al., 2019; Contractor et al., 2022) and datasets built provenance-clean from the outset (Bommarito II et al., 2025). These change what future documents say. They do not help a screener read the documents that already govern existing data.

Abstention and information sufficiency supply the decision-theoretic frame. Abstention benchmarks (Kirichenko et al., 2025) and surveys (Wen et al., 2025) establish that models answer when they should not; selective prediction supplies the decision-theoretic frame (El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017). Closest to us, Joren et al. (2025) separate the context was insufficient from the model failed to use it, but their insufficiency is a retrieval-bundle property and repairable by construction. QuestBench (Li et al., 2025) identifies a missing variable that is always obtainable. Hou et al. (2024b) presuppose that a clarification exists. All three are our incidental case, as are the related distinctions that treat the missing information as supplyable in principle: re-

Passage	Korean original, and our translation below it
Grant, for AI research and development (Table 2)	또한 여러분이 제공한 소중한 콘텐츠는 [플랫폼] 서비스를 개선하고 새로운 [플랫폼] 서비스를 제공하기 위해 인공지능 분야 기술 등의 연구 개발 목적으로 [플랫폼]에서 사용되거나, [플랫폼]의 위탁에 의해 위와 같은 목적으로 [플랫폼] 계열사에서 사용될 수 있습니다. <i>The content you provide may also be used by [the platform] for research and development in fields such as artificial-intelligence technology, in order to improve [its] services and to provide new ones, or used by [its] affiliates for the same purposes under delegation.</i>
Opt-out clause and its exception (Table 2)	여러분이 위의 관리기능 또는 요청을 통하여 삭제, 비공개 등의 조치를 함으로써 일반 공중의 접근 및 열람을 허용하지 않은 콘텐츠는, 해당 조치 전에 이미 기술적으로 구분이 어렵게 처리된 경우 외에는, 그 조치 시점 이후에는 인공지능 분야 기술 등의 연구 개발 목적으로도 사용되지 않습니다. <i>Content that you have withheld from public access, by deletion or by making it private through the controls or requests above, will not be used for research and development in fields such as artificial-intelligence technology after the time of that action, except where it had already been processed, before that action, in a way that makes it technically difficult to distinguish.</i>
Rights-holding requirement (upstream title) (Table 2; Appendix A.3)	게시물 게재로 여러분은 [플랫폼]에게 그러한 권한을 부여하게 되므로, 여러분은 이에 필요한 권리를 보유하고 있어야 합니다. <i>Because posting content grants [the platform] those rights, you must hold the rights necessary to do so.</i>
Scope-of-grant clause, ministry standard-form contract (Appendix B)	이용허락 대상이 되는 권리는 저작권재산권 중 당사자가 합의한 권리로 한다. ☐ 복제권 ☐ 공연권 ☐ 공중송신권 ☐ 전시권 ☐ 배포권 ☐ 대여권 ☐ 2차적저작물작성권 <i>The rights licensed shall be those of the author's economic rights that the parties agree upon.</i> ☐ reproduction ☐ public performance ☐ public transmission ☐ exhibition ☐ distribution ☐ rental ☐ production of derivative works.

Table 8: Korean Originals and Translations. Every instrument passage quoted in the paper, with our translation. Issuer names are replaced by a bracketed role, and the checkboxes in the last row are reproduced from the ministry form.

solvable versus subjective gaps (Majumdar, 2026), underspecification uncertainty as its own category (Kirchhof et al., 2025), and missing information as a prediction target (Vincent et al., 2026; Watanabe et al., 2026). Ours is a property of the governing instrument, and no amount of retrieval supplies it. The distinction has a counterpart in the uncertainty literature, where Young (2026) argues that reducibility is a property of an uncertainty paired with a way of acquiring evidence rather than a kind of uncertainty on its own. What our screener cannot reduce by reading more of the document, a lawyer reduces by investigating something that is not a document. On the generation side, Sharma et al. (2025) argue that refusal-based copyright filters buy compliance at the cost of usefulness. Naming the missing fact is the screening-side answer to that objection, since it returns something a reviewer can act on rather than a bare refusal.

Korean legal NLP has concentrated on public law. LBOX OPEN (Hwang et al., 2022), KBL (Kim et al., 2024) and KoBLEX (Lee et al., 2025) are Korean legal benchmarks, all on public law. Nearest to us, CALRK-Bench (Jung et al., 2026) asks whether retrieved statutory provisions suffice to answer a consultation question. It is contempo-

aneous and independent, a positive control rather than a competitor. What is missing there is a retrievable provision of public law, whereas here it is an extrinsic fact about a transaction. To our knowledge no Korean NLP resource covers private contractual instruments.

Work on the limits of long context bears on our second result. Performance degrades with input length independently of retrieval quality (Liu et al., 2024; Levy et al., 2024; Du et al., 2025), effective context falls short of claimed context (Hsieh et al., 2024), irrelevant retrieved material actively hurts (Shi et al., 2023; Cuconasu et al., 2024), and in law Xia et al. (2025) show that needle-in-a-haystack evaluation overstates reference recall. This is why the saturation result in §5 is not a null. Added context is costly by default, so a flat curve on indeterminate items is a substantive finding rather than an absence of one. The determinate stratum at matched context length excludes the alternative explanation that the model simply drowned in tokens.

Legal indeterminacy and annotator disagreement are modeled separately from ours. Defeasible inference appears in NLI (Rudinger et al., 2020; Allaway and McKeown, 2025) and in deontic formalizations of statutes (Horner et al., 2025). Docu-

ment information sufficiency is distinct from both, and we treat annotation variation as measurement, not noise (Plank, 2022). Ling et al. (2026) report that an “unknown” option changes model behavior as a nonsense token, which is why we run their control rather than assuming our INDET option is inert. Guerdan et al. (2025) and Kim et al. (2023) apply the vocabulary to adjacent objects: indeterminacy of the measurement protocol, and assumptions unverifiable on current evidence. Work on what a model can know draws the boundary elsewhere: Ahdriz et al. (2024) draw it relative to a larger model, which recedes as models improve, and Guerdan et al. (2025) around a rating protocol. Ours is a property of the governing instrument.

D Experimental Setup and Full Results

D.1 Experimental setup details

Open-weight models that we ran ourselves used one machine with two NVIDIA RTX PRO 6000 Blackwell GPUs, 96 GB each, in bfloat16. The rest were called through one commercial API router, including Llama-3.1-70B, Llama-3.1-8B, Gemma-3-27B and Gemma-3-12B, whose weights are openly available. Decoding is greedy: temperature = 0 and top_p = 1. Every model uses the same completion budget, 16,384 tokens for API models and 8,192 for local ones, and a reasoning-effort setting of low where the model exposes one.

Models. The systems we evaluate: GPT-5.6 Sol (OpenAI, 2026), Claude Opus 5 (Anthropic, 2026), Gemini 3.7 Flash (Google DeepMind, 2026), Grok 4.6 (xAI, 2026), Kimi K3 (Kimi Team, 2026), GLM-5.2 (Zhipu AI & Tsinghua University GLM-5-Team, 2026), DeepSeek-V4-Pro (DeepSeek-AI, 2026), Qwen3-Max (Qwen Team, 2025b), Qwen2.5-32B (Qwen Team, 2025a), Qwen2.5-14B (Qwen Team, 2025a), Qwen2.5-7B (Qwen Team, 2025a), Llama-4-Maverick (Meta AI, 2025), Llama-3.1-70B (Meta AI, 2024), Llama-3.1-8B (Meta AI, 2024), Gemma-3-27B (Gemma Team, 2025), Gemma-3-12B (Gemma Team, 2025), Mistral-Medium-3.5 (Mistral AI, 2026), HyperCLOVA X 14B (NAVER Cloud, 2025), Solar-Pro4 (Upstage, 2026) and Keyword filter.

Intervals. Intervals are paired cluster bootstraps over 97 instrument clusters from 104 documents. Benjamini-Hochberg at $q = 0.05$ leaves 0 of 85 saturation cells on $\Delta d'$ and 3 on Δc , so §5.2 rests on the roster-level intervals.

No fine-tuning. At this corpus size a supervised model would recover one annotator’s coding convention, and an unseen instrument admits no training split. The metadata lookup reads no text and bounds what reading adds.

Each control targets one alternative for a flat INDET curve: an oracle clause, no document, an abstention-encouraged prompt (over-commitment moves between -0.021 and $+0.012$ at C4), and a nonsense-label prompt (Ling et al., 2026). We score T2 with mean Jaccard; always emitting all five codes scores Jaccard 0.487 and micro-F1 0.655, and Jaccard is the floor a reported value has to clear.

D.2 Full results

The tables report the full roster; Table 3 shows four systems and points here for the rest. Corpus composition is Table 9; Table 10 is the context grid by gold label; Table 11 is the full protocol table; and Table 12 gives every cell of the context grid with the second task.

	PRE	POST	Unstated	All
Documents	28	54	22	104
Items	140	270	110	520
DET	33	104	46	183
INDET	107	166	64	337
<i>DET share by use hypothesis</i>				
H1				50%
H2				47%
H3				38%
H4				19%
H5				22%
<i>missing-fact type frequency (INDET items)</i>				
E1 consideration and bargaining structure				140
E2 usage of trade				222
E3 the rightsholder’s course of conduct				196
E4 market substitution or new-market creation				186
E5 upstream title				77

Table 9: Corpus Composition. The 104 documents and 520 items by revision era (**Unstated**: no date or 2022), DET share by use over all 520 items, and missing-fact frequency over the 337 indeterminate ones.

Model	Gold	Per-class accuracy					Δ	Interaction	Signal detection		
		C1	C2	C3	C4				C1 $\rightarrow$ C4	Δ [95% CI]	
GPT-5.6 Sol	DET	0.366	0.350	0.372	0.432	+0.07	+0.080 [+0.01, +0.15]		d' 0.59 $\rightarrow$ 0.70	+0.11 [-0.10, +0.33]	
	INDET	0.825	0.807	0.810	0.810	-0.01			c - 0.64 $\rightarrow$ -0.52	+0.11 [+0.01, +0.21]	
Claude Opus 5	DET	0.372	0.432	0.399	0.410	+0.04	+0.041 [-0.04, +0.13]		d' 0.55 $\rightarrow$ 0.64	+0.09 [-0.12, +0.30]	
	INDET	0.810	0.810	0.813	0.807	+0.00			c - 0.60 $\rightarrow$ -0.55	+0.05 [-0.07, +0.18]	
Gemini 3.7 Flash	DET	0.355	0.448	0.475	0.454	+0.10	+0.116 [+0.03, +0.21]		d' 0.31 $\rightarrow$ 0.50	+0.20 [-0.05, +0.48]	
	INDET	0.751	0.706	0.682	0.733	-0.02			c - 0.52 $\rightarrow$ -0.37	+0.15 [+0.03, +0.28]	
Grok 4.6	DET	0.284	0.306	0.317	0.350	+0.07	+0.098 [+0.02, +0.18]		d' 0.41 $\rightarrow$ 0.47	+0.06 [-0.22, +0.35]	
	INDET	0.837	0.825	0.783	0.804	-0.03			c - 0.77 $\rightarrow$ -0.62	+0.15 [+0.03, +0.29]	
Kimi K3	DET	0.344	0.344	0.383	0.383	+0.04	+0.047 [-0.03, +0.12]		d' 0.57 $\rightarrow$ 0.63	+0.07 [-0.18, +0.31]	
	INDET	0.834	0.819	0.804	0.825	-0.01			c - 0.68 $\rightarrow$ -0.61	+0.07 [-0.05, +0.19]	
GLM-5.2	DET	0.486	0.415	0.448	0.399	-0.09	-0.126 [-0.24, -0.02]		d' 0.39 $\rightarrow$ 0.28	-0.11 [-0.39, +0.17]	
	INDET	0.665	0.656	0.694	0.703	+0.04			c - 0.23 $\rightarrow$ -0.39	-0.16 [-0.32, -0.03]	
DeepSeek-V4-Pro	DET	0.295	0.290	0.317	0.322	+0.03	+0.033 [-0.05, +0.13]		d' 0.49 $\rightarrow$ 0.54	+0.05 [-0.23, +0.35]	
	INDET	0.849	0.843	0.804	0.843	-0.01			c - 0.78 $\rightarrow$ -0.73	+0.05 [-0.09, +0.20]	
Qwen3-Max	DET	0.230	0.273	0.246	0.301	+0.07	+0.095 [+0.02, +0.17]		d' 0.49 $\rightarrow$ 0.59	+0.10 [-0.16, +0.39]	
	INDET	0.890	0.864	0.884	0.866	-0.02			c - 0.98 $\rightarrow$ -0.81	+0.17 [+0.03, +0.30]	
Qwen2.5-32B	DET	0.137	0.175	0.158	0.175	+0.04	+0.012 [-0.05, +0.07]		d' 0.18 $\rightarrow$ 0.51	+0.32 [+0.05, +0.62]	
	INDET	0.899	0.911	0.920	0.926	+0.03			c - 1.18 $\rightarrow$ -1.18	+0.00 [-0.17, +0.11]	
Qwen2.5-14B	DET	0.126	0.104	0.093	0.082	-0.04	-0.070 [-0.11, -0.03]		d' 0.55 $\rightarrow$ 0.69	+0.14 [-0.19, +0.52]	
	INDET	0.955	0.961	0.964	0.982	+0.03			c - 1.41 $\rightarrow$ -1.72	-0.31 [-0.52, -0.15]	
Qwen2.5-7B	DET	0.279	0.262	0.224	0.235	-0.04	-0.136 [-0.23, -0.03]		d' 0.12 $\rightarrow$ 0.32	+0.20 [-0.03, +0.50]	
	INDET	0.760	0.792	0.828	0.852	+0.09			c - 0.64 $\rightarrow$ -0.88	-0.24 [-0.38, -0.07]	
Llama-4-Maverick	DET	0.115	0.126	0.093	0.115	+0.00	-0.021 [-0.06, +0.03]		d' 0.44 $\rightarrow$ 0.67	+0.24 [-0.10, +0.64]	
	INDET	0.950	0.941	0.964	0.970	+0.02			c - 1.41 $\rightarrow$ -1.53	-0.12 [-0.29, +0.06]	
Llama-3.1-70B	DET	0.175	0.148	0.120	0.153	-0.02	-0.054 [-0.12, +0.00]		d' 0.31 $\rightarrow$ 0.42	+0.11 [-0.25, +0.36]	
	INDET	0.893	0.908	0.938	0.926	+0.03			c - 1.08 $\rightarrow$ -1.23	-0.14 [-0.30, -0.01]	
Llama-3.1-8B	DET	0.000	0.000	0.000	0.005	+0.01	+0.005 [+0.00, +0.02]		d' 0.19 $\rightarrow$ 0.57 [†]	+0.38 [+0.00, +0.72]	
	INDET	1.000	1.000	1.000	1.000	+0.00			c - 2.88 $\rightarrow$ -2.69	+0.19 [+0.00, +0.36]	
Gemma-3-27B	DET	0.175	0.158	0.104	0.093	-0.08	-0.177 [-0.27, -0.09]		d' 0.18 $\rightarrow$ 0.44	+0.26 [-0.04, +0.63]	
	INDET	0.866	0.911	0.947	0.961	+0.09			c - 1.02 $\rightarrow$ -1.53	-0.51 [-0.72, -0.31]	
Gemma-3-12B	DET	0.082	0.060	0.033	0.000	-0.08	-0.135 [-0.25, -0.05]		d' 0.22 $\rightarrow$ 0.19 [†]	-0.03 [-0.43, +0.77]	
	INDET	0.947	0.958	0.976	1.000	+0.05			c - 1.49 $\rightarrow$ -2.88	-1.39 [-1.72, -0.91]	
Mistral-Medium-3.5	DET	0.328	0.361	0.279	0.306	-0.02	-0.004 [-0.08, +0.08]		d' 0.58 $\rightarrow$ 0.45	-0.13 [-0.37, +0.12]	
	INDET	0.849	0.843	0.834	0.831	-0.02			c - 0.73 $\rightarrow$ -0.73	+0.01 [-0.12, +0.14]	
HyperCLOVA X 14B	DET	0.197	0.268	0.219	0.191	-0.01	-0.026 [-0.11, +0.05]		d' 0.37 $\rightarrow$ 0.47	+0.10 [-0.20, +0.36]	
	INDET	0.890	0.861	0.893	0.911	+0.02			c - 1.03 $\rightarrow$ -1.10	-0.07 [-0.25, +0.10]	
Solar-Pro4	DET	0.142	0.191	0.158	0.137	-0.01	+0.006 [-0.06, +0.07]		d' 0.72 $\rightarrow$ 0.57	-0.15 [-0.62, +0.32]	
	INDET	0.964	0.950	0.958	0.953	-0.01			c - 1.42 $\rightarrow$ -1.37	+0.05 [-0.13, +0.26]	

Table 10: Context Saturation by Gold Label. Per-class T1 accuracy at each condition under the fixed P2 wording; Δ is $C4 - C1$, the interaction is $\Delta_{\text{DET}} - \Delta_{\text{INDET}}$, and the last columns give the $C1 \rightarrow C4$ change in d' and c with 95% cluster-bootstrap intervals. The interaction is orthogonal to d' , so only $\Delta d'$ and Δc separate a gain from a threshold shift (§5.2). [†]: answers INDET on at least 99% of items at C1 or C4, so d' is not read.

Model	Protocol	Bal. acc	MCC	O.c.	Bias	T2 Jac.
GPT-5.6 Sol	P1 naive prompt	0.524	0.087	0.056	+0.279	0.264
	P2 task-structured	0.603	0.227	0.181	+0.098	0.367
	P3 expert protocol	0.544	0.099	0.202	+0.119	0.351
Claude Opus 5	P1 naive prompt	0.530	0.058	0.481	−0.150	0.156
	P2 task-structured	0.616	0.245	0.205	+0.065	0.376
	P3 expert protocol	0.571	0.143	0.285	+0.017	0.341
Gemini 3.7 Flash	P1 naive prompt	0.555	0.135	0.151	+0.162	0.265
	P2 task-structured	0.569	0.141	0.282	+0.021	0.299
	P3 expert protocol	0.610	0.226	0.234	+0.040	0.342
Grok 4.6	P1 naive prompt	0.537	0.076	0.303	+0.023	0.280
	P2 task-structured	0.536	0.085	0.184	+0.142	0.361
	P3 expert protocol	0.550	0.109	0.223	+0.094	0.376
Kimi K3	P1 naive prompt	0.566	0.149	0.190	+0.115	0.258
	P2 task-structured	0.584	0.198	0.148	+0.144	0.367
	P3 expert protocol	0.570	0.153	0.205	+0.098	0.368
GLM-5.2	P1 naive prompt	0.548	0.106	0.211	+0.108	0.262
	P2 task-structured	0.564	0.125	0.347	−0.040	0.302
	P3 expert protocol	0.622	0.256	0.205	+0.062	0.386
DeepSeek-V4-Pro	P1 naive prompt	0.542	0.110	0.125	+0.198	0.255
	P2 task-structured	0.548	0.115	0.166	+0.152	0.363
	P3 expert protocol	0.551	0.116	0.193	+0.123	0.409
Qwen3-Max	P1 naive prompt	0.557	0.165	0.083	+0.229	0.306
	P2 task-structured	0.563	0.159	0.131	+0.177	0.391
	P3 expert protocol	0.562	0.124	0.303	+0.006	0.311
Qwen2.5-32B	P1 naive prompt	0.538	0.115	0.083	+0.242	0.281
	P2 task-structured	0.543	0.127	0.089	+0.233	0.343
	P3 expert protocol	0.552	0.181	0.045	+0.271	0.402
Qwen2.5-14B	P1 naive prompt	0.604	0.217	0.228	+0.050	0.195
	P2 task-structured	0.533	0.130	0.039	+0.290	0.457
	P3 expert protocol	0.526	0.092	0.062	+0.271	0.464
Qwen2.5-7B	P1 naive prompt	0.563	0.122	0.350	−0.042	0.153
	P2 task-structured	0.527	0.062	0.208	+0.125	0.307
	P3 expert protocol	0.525	0.049	0.496	−0.162	0.255
Llama-4-Maverick	P1 naive prompt	0.569	0.153	0.196	+0.108	0.244
	P2 task-structured	0.540	0.138	0.056	+0.267	0.351
	P3 expert protocol	0.589	0.217	0.128	+0.162	0.346
Llama-3.1-70B	P1 naive prompt	0.547	0.091	0.457	−0.138	0.216
	P2 task-structured	0.522	0.062	0.110	+0.227	0.348
	P3 expert protocol	0.572	0.140	0.353	−0.052	0.296
Llama-3.1-8B	P1 naive prompt	0.505	0.013	0.154	+0.194	0.287
	P2 task-structured	0.500	0.000	0.000	+0.352	0.466
	P3 expert protocol	0.500	−0.001	0.116	+0.237	0.433
Gemma-3-27B	P1 naive prompt	0.524	0.049	0.680	−0.344	0.110
	P2 task-structured	0.528	0.082	0.098	+0.235	0.378
	P3 expert protocol	0.568	0.158	0.169	+0.135	0.472
Gemma-3-12B	P1 naive prompt	0.502	0.006	0.089	+0.262	0.337
	P2 task-structured	0.509	0.041	0.042	+0.304	0.376
	P3 expert protocol	0.516	0.038	0.208	+0.133	0.400
Mistral-Medium-3.5	P1 naive prompt	0.567	0.149	0.199	+0.106	0.293
	P2 task-structured	0.581	0.188	0.160	+0.135	0.329
	P3 expert protocol	0.572	0.183	0.119	+0.183	0.440
HyperCLOVA X 14B	P1 naive prompt	0.553	0.133	0.139	+0.175	0.216
	P2 task-structured	0.563	0.159	0.131	+0.177	0.290
	P3 expert protocol	0.575	0.174	0.166	+0.133	0.406
Solar-Pro4	P1 naive prompt	0.568	0.206	0.065	+0.238	0.258
	P2 task-structured	0.545	0.164	0.042	+0.279	0.433
	P3 expert protocol	0.559	0.219	0.030	+0.281	0.476

Table 11: The Full Protocol Table. Every model-and-level cell, all at C2. Bal. acc is balanced accuracy, MCC is the Matthews correlation coefficient, O.c. is over-commitment, and Jac. is the Jaccard index. Bias is the INDET response rate minus the gold base rate (0.648). The P0 keyword filter is in Table 3.

Model	Ctx	Bal. acc	MCC	Over-commitment	T2 Jaccard
GPT-5.6 Sol	C1	0.596	0.213	0.175	0.366
	C2	0.578	0.173	0.193	0.354
	C3	0.591	0.199	0.190	0.347
	C4	0.621	0.259	0.190	0.344
Claude Opus 5	C1	0.591	0.199	0.190	0.399
	C2	0.621	0.259	0.190	0.396
	C3	0.606	0.230	0.187	0.395
	C4	0.608	0.234	0.193	0.396
Gemini 3.7 Flash	C1	0.553	0.112	0.249	0.344
	C2	0.577	0.155	0.294	0.289
	C3	0.579	0.156	0.318	0.302
	C4	0.593	0.189	0.267	0.294
Grok 4.6	C1	0.560	0.143	0.163	0.360
	C2	0.565	0.151	0.175	0.343
	C3	0.550	0.110	0.217	0.342
	C4	0.577	0.170	0.196	0.330
Kimi K3	C1	0.589	0.202	0.166	0.378
	C2	0.582	0.183	0.181	0.356
	C3	0.593	0.203	0.196	0.356
	C4	0.604	0.229	0.175	0.357
GLM-5.2	C1	0.576	0.148	0.335	0.286
	C2	0.536	0.070	0.344	0.294
	C3	0.571	0.142	0.306	0.283
	C4	0.551	0.104	0.297	0.291
DeepSeek-V4-Pro	C1	0.572	0.171	0.151	0.389
	C2	0.566	0.157	0.157	0.382
	C3	0.561	0.136	0.196	0.369
	C4	0.583	0.192	0.157	0.375
Qwen3-Max	C1	0.560	0.159	0.110	0.399
	C2	0.568	0.168	0.136	0.378
	C3	0.565	0.169	0.116	0.382
	C4	0.584	0.202	0.134	0.357
Qwen2.5-32B	C1	0.518	0.054	0.101	0.360
	C2	0.543	0.127	0.089	0.343
	C3	0.539	0.121	0.080	0.300
	C4	0.550	0.154	0.074	0.297
Qwen2.5-14B	C1	0.541	0.149	0.045	0.453
	C2	0.533	0.130	0.039	0.457
	C3	0.529	0.119	0.036	0.469
	C4	0.532	0.156	0.018	0.472
Qwen2.5-7B	C1	0.519	0.042	0.240	0.306
	C2	0.527	0.062	0.208	0.307
	C3	0.526	0.063	0.172	0.312
	C4	0.543	0.108	0.148	0.343
Llama-4-Maverick	C1	0.532	0.118	0.050	0.374
	C2	0.533	0.115	0.059	0.343
	C3	0.529	0.119	0.036	0.329
	C4	0.543	0.172	0.030	0.330
Llama-3.1-70B	C1	0.534	0.096	0.107	0.323
	C2	0.528	0.084	0.092	0.347
	C3	0.529	0.100	0.062	0.359
	C4	0.539	0.124	0.074	0.366
Llama-3.1-8B	C1	0.500	0.000	0.000	0.471
	C2	0.500	0.000	0.000	0.457
	C3	0.500	0.000	0.000	0.464
	C4	0.503	0.060	0.000	0.470
Gemma-3-27B	C1	0.521	0.056	0.134	0.375
	C2	0.535	0.105	0.089	0.384
	C3	0.525	0.094	0.053	0.410
	C4	0.527	0.111	0.039	0.416
Gemma-3-12B	C1	0.514	0.056	0.053	0.403
	C2	0.509	0.041	0.042	0.375
	C3	0.505	0.027	0.024	0.385
	C4	0.500	0.000	0.000	0.376
Mistral-Medium-3.5	C1	0.588	0.206	0.151	0.355
	C2	0.602	0.231	0.157	0.324
	C3	0.556	0.133	0.166	0.339
	C4	0.568	0.158	0.169	0.335
HyperCLOVA X 14B	C1	0.543	0.120	0.110	0.319
	C2	0.564	0.158	0.139	0.294
	C3	0.556	0.151	0.107	0.315
	C4	0.551	0.148	0.089	0.321
Solar-Pro4	C1	0.553	0.195	0.036	0.446
	C2	0.570	0.224	0.050	0.429
	C3	0.558	0.203	0.042	0.428
	C4	0.545	0.158	0.047	0.431
Always INDET	—	0.500	0.000	0.000	—
Always all five codes	—	—	—	—	0.487
Same-document lookup	—	—	—	—	0.623
Hypothesis $\times$ era lookup	—	0.651	0.305	0.228	—

Table 12: Results by Context Condition. Every system at each condition under the fixed P2 wording; the C2 column repeats the P2 configuration of Table 11, and the other columns are not comparable with that table. Below the rule are the references that read no document (§4.2, §5.2).