

Where Experts Disagree, Models Fail: Detecting Implicit Legal Citations in French Court Decisions

Avrile Floro¹

avrile.floro@ip-paris.fr

Tamara Dhorasoo²

dhorasoo.tamara@uphf.fr

Soline Pellet²

soline.pellet@uphf.fr

Nils Holzenberger¹

nils.holzenberger@telecom-paris.fr

¹Télécom Paris, Institut Polytechnique de Paris ²Université Polytechnique Hauts-de-France

Abstract

Applying computational methods to law at scale requires separating genuine legal reasoning from surface similarity. We study this through a concrete task: detecting *implicit* citations of the French Civil Code, where a court applies a statutory rule without naming it (a *post-hoc* question about the reasoning a court actually used). We release a benchmark of 1,015 passage–article pairs annotated by three legal experts. Our central finding is that their disagreement is itself informative: the third of cases the experts dispute are where models fail. Our best ensemble reaches an F1 score of 0.70 overall. Yet, two-thirds of its false positives fall on those disputed cases, a concentration that holds across all ten models we evaluate. Expert disagreement thus signals uncertainty in the task and in its gold label, which aggregate metrics hide. This should not block useful tools, however: reframed as top-*k* ranking with multi-model consensus, the same signals reach 76% precision for the top-200 candidates of the benchmark without supervision.

1 Introduction

A lawyer researching case law faces an asymmetry: explicit citations are trivial to find through keyword search. However, implicit applications, where a court applies a legal rule without naming it, are hidden. Consider a practitioner seeking examples of how article 2274 of the French Civil Code (the presumption of good faith) is applied in practice. Searching for “article 2274” retrieves decisions that explicitly cite this provision. Yet many decisions apply the same legal reasoning without numerical reference, using formulations such as “the mere observation of the increase in rental debt is not sufficient to establish bad faith.” This blind spot is also relevant to quantitative legal scholarship. Take a researcher studying whether French courts have expanded the scope of the good-faith presumption over the past decade. If the analysis captures only

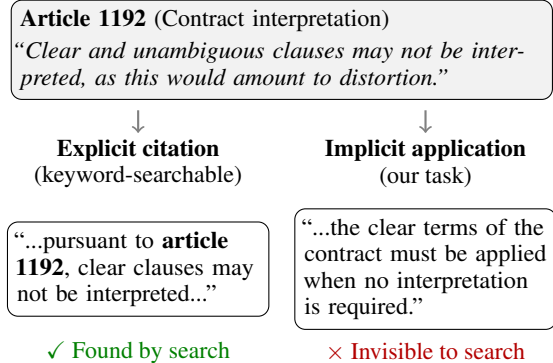


Figure 1: Explicit and implicit statutory citations. While both excerpts apply article 1192, only the left one can be found by keyword search.

decisions that explicitly cite article 2274, it misses cases where the same provision is applied implicitly, and could potentially skew conclusions about jurisprudential trends. Our work begins to address this problem by evaluating the reliability of automatic detection. Specifically, this paper tackles the task of detecting *implicit statutory citations* (Figure 1): given a passage from a court decision and a candidate Civil Code article, determine whether the passage applies that article’s legal rule without explicitly mentioning it.

This task is both practically important and methodologically challenging. Indeed, it requires distinguishing genuine legal reasoning from semantic similarity. But how difficult is this task, and where do current methods fail? We make four contributions¹ that characterize both the limits and the practical potential of computational approaches to this problem. **First**, we introduce an adversarial benchmark for implicit citation detection in French civil law (§3). We train a bi-encoder on explicit citations, use it to retrieve semantically similar candidates, then perform adversarial filtering using o3 with a conservative prompt. The final dataset com-

¹Data and code: <https://github.com/avrilemay/implicit-legal-citations> (code);
doi:10.5281/zenodo.21206799 (model & data)

prises 1,015 pairs. **Second**, we conduct an annotation study with three legal experts (§4). It reveals the difficulty of the task: experts frequently disagree on whether a passage applies legal reasoning or merely states facts. **Third**, we show that expert disagreement predicts model failure (§5–6). Our supervised ensemble achieves an overall F1 of 0.70. This aggregate masks failure on disputed cases: most of its false positives fall on the cases where annotators disagreed, a pattern observed across all ten models tested. **Fourth**, despite these limits, we show a path toward practical assistance tools (§5.3). Reframing classification as top- k ranking and exploiting LLM consensus yields 76% precision when k is set to 200 in an unsupervised setting.

2 Related Work

Legal NLP and Citation Detection Explicit citation extraction identifies numerical statute references, and can be solved with straightforward pattern matching. Several tasks in legal NLP aim at connecting legal statutes and cases, where no explicit link exists between the two. Statutory reasoning is the task of predicting whether a given statute applies to a given case (Holzenberger et al., 2020). Legal Statute Identification is the more general task of deciding which statutory provisions are relevant to a given case from a closed set (Paul et al., 2024, 2022). An open-set version of the task exists, as statute retrieval in the COLIEE competition (Goebel et al., 2023), with related prior-case-retrieval benchmarks (Joshi et al., 2023; Paul et al., 2025). Our task differs from all three: instead of retrieving relevant articles for a factual query (an *ex ante* prediction), we determine whether a court applies a specific article’s legal rule within its reasoning (a *post-hoc* detection). These two are independent: a decision may omit a statute that could apply and may, instead, invoke one that is of debatable relevance. The latter is precisely a case of interest to legal scholarship. Our task also differs from natural language inference because our negative categories (factual descriptions, party claims, special regimes) are often logically compatible with the article. Hence, an NLI model may predict entailment even though they do not involve any judicial application. Havaldar et al. (2025) show that NLI models struggle to recognize entailments that are implied rather than explicit.

Legal-domain Language Models Language models fine-tuned on legal-domain documents

have generally improved performance on legal tasks (Dominguez-Olmedo et al., 2025): LegalBERT (Chalkidis et al., 2020) showed gains from domain pretraining; JuriBERT (Douka et al., 2021), which we use as our base encoder, achieves good results on French legal benchmarks. We further use LLMs fine-tuned through continued learning on legal documents, such as SaulLM (Colombo et al., 2024b) and LawMA (Dominguez-Olmedo et al., 2025). Legal LLMs must however reach a certain size to compete with larger, proprietary models (Colombo et al., 2024a). Our experiments show that specialized models capture legal semantics better than general encoders. They are nonetheless vulnerable to failure on ambiguous cases. The model used in §3, o3, ranks 12th on LegalBench (Guha et al., 2023), making it a reasonable baseline.²

Annotation Disagreement in NLP It is the norm in NLP to consider annotator disagreement as noise and the majority vote is used to suppress it. However, recent work challenges this view. Cabitza et al. (2023) and Uma et al. (2021) argue for treating annotator variation as informative signal rather than noise. Kim et al. (2025) distinguish “consensual” from “non-consensual” instances, showing that annotator agreement is predictive of model behavior. Our work provides a legal-domain example of this insight. While cases need to be decided upon, which justifies the use of a gold label, we have quantified the concentration of errors on disputed cases and demonstrated that this pattern is consistent across architectures. The boundary between fact and law has long been recognized as difficult to draw. Šavelka and Ashley (2018) find low agreement on distinguishing factual from legal sentences in U.S. court opinions. The concept of legal “open texture” (Hart, 1961) offers an analogy: when a general rule is applied to particular facts, legal concepts have a core of settled meaning but a penumbra of uncertainty. In a clear case, the court simply states that the rule applies. In a borderline case, the court must first examine the facts in detail to decide whether it does. The sentence that carries this decision then reads like a finding of fact. This is why fact and law become hard to tell apart in such passages.

²See https://www.vals.ai/benchmarks/legal_bench, consulted on February 19, 2026.

3 Dataset Construction

We introduce a benchmark for detecting *implicit* citations of the French Civil Code in court decisions, when a legal rule is applied with no numerical reference to the article.

3.1 Source Data

We used Judilibre, the open API of the French Ministry of Justice, as the main source for our data. Personal names are pseudonymized at the source. We restrict our scope to decisions from *tribunaux judiciaires* (first-instance civil courts). At the time of collection in July 2025, only 86 of the 164 courts had published decisions and no data was available prior to December 2023. We collected 182,155 decisions with usable motivation sections (i.e., Judilibre’s zoning delimits a non-empty, unredacted *motivations* zone), spanning December 2023 to July 2025. We divided each motivation into chunks with a sentence-aware algorithm (Appendix A). We focus on first-instance courts because they publish more decisions than appellate courts, broadening the coverage. Furthermore, in French civil law, only first-instance and appellate courts (not the Cour de cassation) can rule and interpret contracts. They are governed by the obligations articles that dominate the benchmark (Book III, 77%, Table 1). Finally, limiting the corpus to a single jurisdiction level ensures homogeneity.

3.2 Bi-Encoder Training on Explicit Citations

We first train a retrieval model on *explicit* citations, i.e. cases where chunks contain verbatim article numbers, extracted with regex patterns and TF-IDF filtering (details in Appendix A). Before training, we mask all legal references in the text (article numbers, code names, statute references) with special tokens. We force the model to learn semantic associations rather than pattern matching. We fine-tune a JuriBERT-based bi-encoder with Multiple Negatives Ranking Loss (Henderson et al., 2017), encode all Civil Code articles and use a FAISS index (Johnson et al., 2021) for nearest-neighbor retrieval.

3.3 Implicit Candidate Generation

Our objective is to uncover *implicit* citations. We apply the bi-encoder to the held-out split of its data (Appendix A). First, we eliminate chunks containing explicit keywords such as ‘article’, ‘loi’, ‘code’. Second, for each chunk, we retrieve the top-5 near-

est articles. Finally, a candidate is discarded if its article is cited elsewhere in the decision. We also account for statutory renumbering (Appendix B). This yields 40,566 candidate pairs.

3.4 Adversarial Filtering with o3

We applied adversarial filtering (Zellers et al., 2018) to select challenging examples. We keep candidates labeled as positive by a strong model. The 565 negatives identified after human annotation are cases where the model failed. This creates difficult negative cases, where models sharing similar reasoning patterns with the one being used as a filter will likely fail. We used OpenAI’s o3 model to evaluate these pairs. The prompt is engineered to be strict since the model is asked to reply ‘NO’ in case of doubt (full prompt in Appendix C). From this pool, o3 returned 4,206 positive predictions (10.4%), involving 497 articles. In accordance with the adversarial filtering objective, we only selected the pairs where o3 predicted the implicit use of an article. Finally, to avoid circularity, o3 was used only to create the dataset and was not evaluated afterwards on the final task.

3.5 Selection for Annotation

From the 4,206 pairs predicted positive by o3, one of the legally-trained authors selected pairs for annotation, prioritizing coverage across distinct Civil Code articles. The final dataset contains 1,015 pairs of (chunk, article) over 418 Civil Code articles (per-article frequency in Appendix D). Table 1 shows the distribution across the books of the Civil Code. Book III covers contracts, successions, and matrimonial regimes. It accounts for 77% of the dataset, which is expected because cases related to obligations dominate first-instance civil litigation. Length statistics are reported in Appendix D.

Book	Art.	Pairs
I — Persons	52	86 (8.5%)
II — Property	11	21 (2.1%)
III — Obligations	310	785 (77.3%)
IV — Security	43	121 (11.9%)
V — Mayotte	2	2 (0.2%)

Table 1: Dataset distribution across Civil Code books.

3.6 Bounding Selection Bias

Every retained pair was predicted positive by o3. Because the benchmark is a filtered sample, it is possible that o3 wrongly rejected genuine implicit

citations. To bound this risk, we drew a control sample of 100 (chunk, article) pairs retrieved by the bi-encoder but *rejected* by o3, stratified over 100 distinct Civil Code articles to avoid article-frequency confounds. They were labeled by the same three legal experts under the protocol of §4 (A_1 and A_2 independently, A_3 adjudicating disagreements). A_1 judged all 100 non-implicit and A_2 judged 98 non-implicit. A_3 resolved the two disputed pairs as non-implicit (facts only). No pair was confirmed as an implicit application. With zero positives in 100 trials, the rate of missed implicit citations among o3-rejected candidates is bounded below roughly 3% (95% CI). This is an upper bound on the selection bias, not an evidence of its absence.

4 Annotation Study

Protocol Annotation was carried out independently by three of the authors, all legally trained in French civil law. Each entry pair was reviewed and assigned a label: YES, when the article was deemed to be applied implicitly, or NO otherwise. Negative cases could be further qualified (e.g., the passage states facts, or relies on a different legal regime). Appendix E details the annotation interface built for the campaign. Annotators A_1 and A_2 labeled the whole dataset. A_3 was the adjudicator for the 339 cases where A_1 and A_2 disagreed.

Agreement Statistics Agreement between A_1 and A_2 reached 66.6% (Cohen’s $\kappa = 0.33$), consistent with prior work on implicit reasoning tasks (Šavelka and Ashley, 2018; Troiano et al., 2019). A_3 resolved the 339 disputed cases (33.4%), for a total of 450 YES and 565 NO. Table 2 shows that A_2 labeled 67% of cases as YES versus 50% for A_1 . This asymmetry explains the moderate κ . Across the disputed cases, A_3 sided with A_1 in 73.5% of cases, versus 26.5% with A_2 . This suggests that A_1 adopted a more restrictive reading.

	A_2 =Yes	A_2 =No
A_1 =Yes	425	83
A_1 =No	256	251

Table 2: Confusion matrix between primary annotators.

Structure of Disagreement The decisions of Annotator A_3 on the 339 disputed cases reveal two patterns (Table 9 in Appendix F). 92.6% of disagreements resolve to NO: borderline cases are

more often judged as non-implicit in the end. Moreover, the fact versus law boundary dominates in the disputed cases. 43.4% of disagreements involve chunks describing facts or party claims without judicial reasoning, according to A_3 ’s categorization. This pattern is consistent with the distinction drawn by Hart (1961) between the “core” of legal concepts (clear cases) and their “penumbra” (borderline cases where application is genuinely uncertain).

Implications for Evaluation Recent works state that annotator disagreement should not be interpreted as mere noise, but as a signal reflecting genuine ambiguity (Aroyo and Welty, 2015; Pavlick and Kwiatkowski, 2019; Uma et al., 2021). We embrace this perspective: the 339 disputed cases may constitute instances where trained experts reached different conclusions. This prompts our question: do models fail indiscriminately, or do their errors concentrate upon these disputed cases?

5 Experiments

We evaluated several approaches on the task of implicit citation detection. Although we frame annotations as binary classification, the ultimate goal is retrieval: given a passage from a Court decision, identify which Civil Code articles are implicitly applied, if any. We therefore progress from supervised classification (§5.1) to zero-shot LLM classification (§5.2), and finally to unsupervised ranking that combines multiple signals (§5.3).

5.1 Supervised Classifiers

Models We evaluate several pre-trained encoders on the classification task. To avoid overfitting considering the limited size of our dataset (1,015 entries), we freeze the encoders’ parameters, extract sentence-level representations, and then train simple classifiers on top. We test eight encoders across three categories: general French models CamemBERT (Martin et al., 2020) and CamemBERTav2 (Antoun et al., 2024); legal-domain models JuriBERT (Douka et al., 2021), SaulLM-7B (Colombo et al., 2024b), and Lawma (Dominguez-Olmedo et al., 2025); and general-purpose models LLaMA-3.1-8B (Grattafiori et al., 2024), MPNet (Song et al., 2020), and MiniLM (Wang et al., 2020). All experiments rely on 5-fold cross-validation, with folds grouped by decision ID. Each fold contains 203 examples, for a total of 1,015. Performance is primarily assessed using Matthews Correlation

Coefficient (MCC), which is more informative than raw accuracy under class imbalance. For each encoder, we vary input formatting, pooling methods, layer selection, and the classifier head. Additional details are reported in Appendix G. As a baseline, we use TF-IDF with logistic regression.

Model	Head	F1	Acc	MCC
SAUL	LR	.69	.74	.47
LLaMA	MLP1	.69	.74	.46
LawMA	MLP1	.64	.73	.46
ST-MPNet	MLP2	.63	.72	.43
CamemBERT	MLP2	.67	.72	.43
CamemBERTav2	MLP1	.65	.71	.42
TF-IDF	LR	.63	.71	.41
JuriBERT	LR	.62	.69	.37
ST-MiniLM	MLP2	.61	.69	.37

Table 3: Supervised classification results (5-fold CV, F1 on positive class). Configuration details in Appendix G.

Table 3 shows that among individual models, the strongest result is obtained by SAUL-7B with an MCC of 0.47. The TF-IDF baseline is competitive (MCC = 0.41) and surpasses JuriBERT (MCC = 0.37), despite the latter being pre-trained on French legal corpora. It suggests that lexical overlap between articles and judicial passages captures part of the signal for this task.

Ensemble To improve performance, we move beyond single models and construct ensembles with nested cross-validation. We explore weighted averaging, stacking, and rank-based fusion. The highest-scoring configurations all achieve the same MCC of 0.53 (Appendix H). We retain the stacking ensemble with a logistic-regression meta-learner combining CamemBERTav2, JuriBERT, LLaMA-3.1-8B, and SAUL-7B. It reaches an F1 of 70% on the positive class and an accuracy of 77%. Both legal-domain models (JuriBERT for French, SAUL-7B for English) appear in every top-performing ensemble, suggesting their representations complement those of generalist encoders. Appendix H houses the full results.

Purpose-Built Entailment and Cross-Encoder Baselines To test whether the difficulty is specific to the frozen-encoder models, we add two models that are architecturally distinct and fine-tuned: XLM-RoBERTa-large fine-tuned on XNLI (an encoder purpose-built for entailment) and a fine-tuned BGE-reranker-v2-m3 cross-encoder (same 5-fold CV, Appendices I and J). Neither matches the supervised ensemble. More importantly, both

reproduce the disagreement-failure pattern of §6: their false positives concentrate on disputed true negatives (OR = 2.26 for XLM and 2.22 for BGE, both $p = 0.0001$), within the range of Figure 3. The concentration is therefore not an artifact of the models we use.

5.2 Zero-Shot LLM Approaches

Our dataset was constructed using o3 as an adversarial filter (§3). It raises the question of whether other LLMs share similar reasoning patterns. We evaluate ten instruction-tuned LLMs with zero-shot prompting (Appendix K). We include models from the Qwen family (Qwen et al., 2025; Yang et al., 2025) (2.5-7B, 2.5-32B, and 3-32B with thinking mode), LLaMA-3.1 (Grattafiori et al., 2024) (8B and 70B), as well as Mistral-Nemo-12B (Mistral AI, 2024), Gemma-2-27B (Gemma Team et al., 2024), Aya-Expanse-32B (Dang et al., 2024), Command-R-35B (Cohere, 2024), and the legal-domain SaulLM-7B (Colombo et al., 2024b). Model sizes range from 7B to 70B parameters.

Results Table 4 reports performance (full results per annotator in Appendix L). Most models over-predict positive cases, with yes-rates ranging from 22% to 96%, far from the 44% base rate. This positive bias validates our adversarial construction: most off-the-shelf models have similar accuracy as o3. Curiously, despite SAUL-7B achieving the best supervised performance (MCC = 0.47), it shows extreme positive bias in a zero-shot setting (95% yes-rate, MCC = 0.03). It suggests that legal pre-training improves representation quality but does not calibrate the model’s zero-shot judgment. This agrees with findings showing that the benefits of domain-specific pre-training depend on the task (Gururangan et al., 2020). Retrieval-augmented few-shot prompting on the three zero-shot models with the best performance (Qwen-2.5-32B, Qwen-2.5-7B, and LLaMA-3.1-8B) does not approach the ensemble either (Appendix M). The best gain is +0.02 MCC.

The positive bias that is shared across models makes zero-shot LLMs unreliable binary classifiers. However, the wide range of models’ yes-rate can be leveraged into a new strategy. Instead of using the predictions of a single model, we can treat the agreement among differently calibrated models as a ranking signal.

Model	Yes%	F1	Acc	MCC
Qwen-2.5-32B	66	.66	.63	.31
Qwen-2.5-7B	54	.63	.63	.28
LLaMA-3.1-8B	22	.47	.65	.28
LLaMA-3.1-70B [†]	84	.65	.55	.25
Mistral-Nemo-12B	75	.63	.56	.21
Gemma-2-27B	94	.63	.49	.17
Qwen3-32B [‡]	85	.63	.52	.16
Aya-Expansive-32B	94	.62	.47	.08
Command-R-35B	96	.61	.46	.04
SAUL-7B	95	.61	.46	.03

Table 4: Zero-shot LLM results ($n = 1,015$, F1 on positive class). [†]4-bit quantization. [‡]Reasoning model with thinking mode.

5.3 Toward Unsupervised Retrieval

We use four LLMs with the lowest positive prediction rates: LLaMA-3.1-8B, Qwen-2.5-7B, Qwen-2.5-32B, and Mistral-Nemo-12B (Table 4). Their yes-rates, though still significant, are lower than those of the remaining models. We hypothesize that overly permissive models flood the ranking with weak candidates, which is undesirable in this setting. We use various ranking strategies. **Lexical baselines** uses TF-IDF and BM25 similarity between chunk and article. **Cross-encoder** refers to BGE-reranker-v2-m3 (Chen et al., 2024) used out-of-the-box. It is a multilingual encoder that is applied directly to the French chunk-article pairs. **LLMs intersection** ranks candidates by agreement among LLMs: UNION (at least one positive), INTER2 (2 positives), INTER3, and INTER4 (full agreement). Ties are broken using the cross-encoder score. **Unsupervised Ensemble** combines the four selected LLM predictions with equal weights and gives a bonus for agreement. We also include the normalized scores from TF-IDF, BM25, and the cross-encoder. The weights were set heuristically without tuning on labels to remain truly unsupervised. The ranking is stable across weight choices. The formula and sensitivity study are provided in Appendix N.

Results Table 5 reports Average Precision (AP). The zero-shot ensemble reaches an AP of 0.67, above the random ranking by +0.23. The INTER4 intersection method also performs well, reaching AP = 0.66. The consensus acts as a strong signal, even in an unsupervised environment. Table 6 shows precision and recall at various cutoffs. At $k = 200$, precision is 76% with 34% recall. Figure 2 confirms the unsupervised ensemble outperforms baselines (full results per annotator in Ap-

pendix O). We discuss practical implications in §6.

Method	AP	Δ vs Random
Random	.44	–
TF-IDF	.57	+ .13
BM25	.57	+ .13
Cross-Encoder	.61	+ .17
LLM_Union	.62	+ .18
LLM_Inter2	.63	+ .19
LLM_Inter3	.65	+ .20
LLM_Inter4	.66	+ .21
Unsup. Ensemble	.67	+ .23

Table 5: Average Precision for unsupervised ranking methods on gold labels ($n = 1,015$).

k	TP	FP	P@ k	R@ k	FP red. (%)
50	38	12	.76	.08	57
100	77	23	.77	.17	59
200	151	49	.76	.34	56
300	211	89	.70	.47	47

Table 6: Unsupervised ensemble ranking at various cut-offs. FP red. = reduction vs. random.

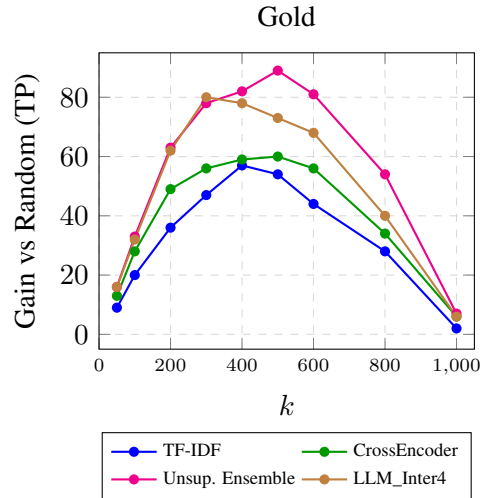


Figure 2: Gain in true positives retrieved compared to random ranking on gold labels.

Limits The same difficulty observed with supervised classifiers persists with unsupervised ranking. At $k = 100$, 65% of false positives (15/23) come from the 33% of disputed cases (Table 24, Appendix O). This concentration calls for a systematic analysis.

6 Discussion

False Positives Concentrate on Disagreement Cases The previous sections found that legal experts disagree on implicit citation detection, and

that both supervised and unsupervised methods reach reasonable performance when aggregated. We now examine the relationship between these two findings: do model errors distribute uniformly? Do they concentrate on the cases that were difficult for human annotators? Our supervised ensemble predicts 66 false positives. Table 7 shows that these errors concentrate on cases of annotator disagreement. The false positive rate is also higher when annotators disagree, by a factor of 1.7 (14.3% vs. 8.4%). To ensure this pattern is robust, we examine false positive rates for the nine individual classifiers used in the supervised experiments. We focus on true negatives (gold = NO, $n = 565$). We compare error rates on consensual negatives versus disagreed ones. After FDR correction, seven of the nine models show significantly higher false positive rates on disagreement cases, as does the supervised ensemble. Figure 3 shows that odds ratios (OR)³ range from 1.28 to 2.91 across models (per-model rates in Appendix P).

Subset	Cases	TN+FP	FP	FPR
Agree	676	251	21	8.4%
Disagree	339	314	45	14.3%
Total	1015	565	66	11.7%

Table 7: False positives by annotator agreement. False positive rate $FPR = FP / (TN+FP)$.

Contrary to what one might expect, models do not become *less* confident on difficult cases. They remain confident while being wrong. Confidence tracks accuracy quite well on agreed cases but not on disagreed ones (Table 27, Appendix Q). Calibration Error is approximately twice as high on disagreement cases (0.30 versus 0.15 for the supervised ensemble). This pattern holds across all models. The concentration persists after controlling for chunk and article length and lexical overlap (logistic regression, Appendix R). Yet, we cannot definitively establish whether the residual difficulty reflects genuine legal ambiguity or other unmeasured factors. Nonetheless, the fact that 43% of disagreements involve the fact versus law boundary (Table 9) is consistent with the hypothesis that some examples represent ambiguous applications of legal rules. A symmetric analysis of the false

³Restricted to true negatives ($n = 565$). $OR = \text{odds}(FP | \text{disagree}) / \text{odds}(FP | \text{agree})$, with $\text{odds}(p) = p / (1 - p)$. $OR > 1$: more FPs on disputed cases.

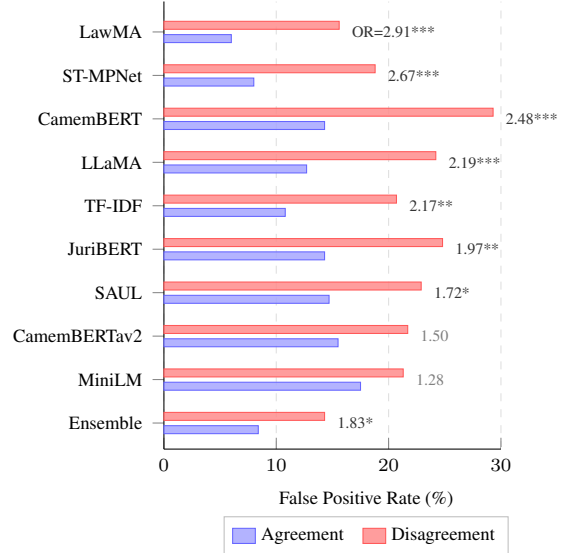


Figure 3: False positive rates by annotator agreement status (agree versus disagree). Odds ratios right of bars. Significance after FDR correction: *** $p < .001$, ** $p < .01$, * $p < .05$; gray = n.s.

negatives shows the same direction but does not reach significance (Appendix S), which is consistent with the very limited number of items tested (25).

Implications for Legal Practitioners Table 6 illustrates the implications of this ranking strategy from a practitioner’s perspective. A legal professional researching how a specific Civil Code article is applied in practice would query the system and receive a ranked, top- k list of court excerpts likely to apply it implicitly. With a cutoff set at a few hundred candidates, the unsupervised ensemble keeps precision high while roughly halving false positives compared to random sampling. Professionals could review a manageable shortlist rather than the full set. Moreover, the remaining false positives tend to be ambiguous cases. This pattern may be due to our adversarial filtering, which added difficult examples to the dataset. Even if it is the case, it suggests that professionals reviewing the flagged candidates would find borderline cases rather than obvious errors. Triage is necessary because the unfiltered test split alone contains 553,082 chunks. Manual review is unfeasible, whereas a ranked shortlist of a few hundred candidates is tractable for an expert.

Evaluation Aggregate metrics can be deceptive when models behave differently depending on the level of difficulty. Splitting data based on annotator agreement makes it possible to assess whether

improvements are limited to easy cases, where annotators agree, or also extend to more ambiguous cases.

7 Qualitative Analysis

We present three examples from the supervised ensemble illustrating when and why the model succeeds and fails (more examples in Appendix T). Appendix U complements them with a quantitative error analysis of the supervised ensemble's 66 false positives. In 56% of the cases, statutory language is present in the chunk but is not applied by the court. In 42%, the model retrieves the right legal domain but an incorrect article.

True positive, agreement: Art. 1192 (contract interpretation) $A_1=A_2=Yes$. Gold=Yes. Conf. 0.81 ✓

Article “*On ne peut interpréter les clauses claires et précises à peine de dénaturation.*” [Clear and unambiguous contractual clauses may not be interpreted, as this would amount to distortion.]

Excerpt “*À l'inverse, sera écartée une contestation qui serait à l'évidence superficielle ou artificielle et le juge est tenu d'appliquer les clauses claires du contrat qui lui est soumis, si aucune interprétation n'en est nécessaire.*” [Conversely, a challenge that is clearly superficial or artificial will be dismissed, and the judge must apply the clear terms of the contract when no interpretation is required.]

Analysis The court applies the principle of article 1192 without citing the number but both texts contain the same words (“clauses claires”). The logic according to which clear clauses require no interpretation is used and it is a clear implicit citation. Both annotators agreed and the model gave it a high confidence score.

False positive, disagreement: Art. 1361 (proof by writing) $A_1=No$, $A_2=Yes$, $A_3=No$ (special regime). Gold=No. Conf. 0.87 ×

Article “*Il peut être suppléé à l'écrit par l'aveu judiciaire, le serment décisoire ou un commencement de preuve par écrit corroboré par un autre moyen de preuve.*” [Written proof may be substituted by judicial admission, decisive oath, or a beginning of proof in writing corroborated by another means of proof.]

Excerpt “*L'offre de crédit produite n'est pas l'original du contrat, et la copie de cet acte juridique ne constitue ici qu'un commencement de preuve par écrit.*” [The credit offer produced is not

the original contract, and the copy of this legal document constitutes here only a beginning of proof in writing.]

Analysis A_2 noticed that the court invokes article 1361's concept of “commencement de preuve par écrit” to declare the evidentiary threshold not met. A_1 disagreed because consumer credit falls under the Code de la consommation, which has its own proof rules. A_3 sided with A_1 . In this case, the judge uses the same wording as the Civil Code but the applicable legal framework is different. The difficulty here is to determine where one regime ends and another begins.

True positive, disagreement: Art. 1219 (exception of non-performance) $A_1=No$, $A_2=Yes$, $A_3=Yes$. Gold=Yes. Conf. 0.71 ✓

Article “*Une partie peut refuser d'exécuter son obligation, alors même que celle-ci est exigible, si l'autre n'exécute pas la sienne et si cette inexécution est suffisamment grave.*” [A party may refuse to perform its obligation, even if it is due, if the other party does not perform its own and if this non-performance is sufficiently serious.]

Excerpt “*Cependant, il n'est pas établi que la SARL NAJI AUTO est privée en permanence d'électricité depuis cette date et l'exception totale d'inexécution n'est donc pas justifiée.*” [However, it is not established that SARL NAJI AUTO has been permanently deprived of electricity since that date, and the total exception of non-performance is therefore not justified.]

Analysis A_1 's negative choice shows it is difficult to identify the proportionality test when the reasoning is mixed with factual finding. A_2 recognized that the judge applies the requirement that non-performance must be “sufficiently serious” to justify the suspension of one's obligations. Because the electricity deprivation is not established, the court concludes the exception is not justified. A_3 confirmed. The model detected the implicit citation despite the formulation.

8 Conclusion

Using a combination of computational methods and human annotation, we build a corpus for the detection of implicit citations in court decisions. Extensive experiments with state-of-the-art NLP methods highlight that the cases where legal experts disagree are also those where computational methods tend to fail. On the practical side, our rank-

ing approach demonstrates that useful assistance may be possible even when perfect classification is not, by surfacing candidates for human review rather than rendering binary judgments. One possibility consistent with our data but not proven by it, is that annotator disagreement reflects genuine ambiguity in the underlying task rather than noise. If so, the concentration of model errors on disputed cases may indicate a shared difficulty: both humans and machines struggle with the same borderline instances. This interpretation aligns with Hart’s notion of legal “open texture” (Hart, 1961). Whether this theoretical framework fully explains the observed pattern is left for future work.

Limitations

Dataset size. 1,015 examples is small by NLP standards but it is quite typical for expert-annotated legal datasets, and our statistical findings are solid. Annotation required advanced legal expertise in French law that cannot be crowd-sourced in order to adjudicate fine-grained distinctions between judicial reasoning and factual description.

Adversarial selection bias. The only candidates retained were predicted positive by o3 because we used it as a filter. Reciprocally, our negatives are cases where o3 was wrong. Models that share reasoning patterns with it might be disadvantaged. Furthermore, models with different reasoning might perform better on our benchmark than on a randomly selected dataset. An annotation of 100 pairs that were rejected by o3 (§3.6) found no missed implicit applications, bounding this risk below 3%. Moreover, the disagreement-failure pattern reproduces on architecturally distinct models (Appendices I and J), indicating it is not an artifact of o3 reasoning.

Zero-shot ensemble weights. The weights in our unsupervised ensemble were set based on intuition rather than optimization. This preserves the unsupervised approach but performance could be improved with cross-validation tuning. A sensitivity analysis (Appendix N) shows that the ranking is robust to weight changes.

References

Wissam Antoun, Francis Kulumba, Rian Touchent, Éric de la Clergerie, Benoît Sagot, and Djamé Seddah. 2024. [Camembert 2.0: A smarter french](#)

[language model aged to perfection](#). *Preprint*, arXiv:2411.08868.

Lora Aroyo and Chris Welty. 2015. [Truth is a lie: Crowd truth and the seven myths of human annotation](#). *AI Magazine*, 36(1):15–24.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. [Toward a perspectivist turn in ground truthing for predictive computing](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. [LEGAL-BERT: The muppets straight out of law school](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904. Association for Computational Linguistics.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Cohere. 2024. Command R: Retrieval-augmented generation at production scale. <https://cohere.com/blog/command-r>.

Pierre Colombo, Telmo Pires, Malik Boudiaf, Rui Melo, Dominic Culver, Etienne Malaboeuf, Gabriel Hauteux, Johanne Charpentier, and Michael Desa. 2024a. [SaulLM-54B & SaulLM-141B: Scaling up domain adaptation for the legal domain](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 129672–129695. Curran Associates, Inc.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre F. T. Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and Michael Desa. 2024b. [SaulLM-7b: A pioneering large language model for law](#). *Preprint*, arXiv:2403.03883.

John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. [Aya expand: Combining research breakthroughs for a new multilingual frontier](#). *Preprint*, arXiv:2412.04261.

Ricardo Dominguez-Olmedo, Vedant Nanda, Rediet Abebe, Stefan Bechtold, Christoph Engel, Jens Frankenreiter, Krishna Gummadi, Moritz Hardt, and Michael Livermore. 2025. [Lawma: The power of specialization for legal annotation](#). In *International Conference on Learning Representations*, volume 2025, pages 35059–35103.

- Stella Douka, Hadi Abdine, Michalis Vazirgiannis, Rajaa El Hamdani, and David Restrepo Amariles. 2021. [JuriBERT: A masked-language model adaptation for French legal text](#). In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 95–101, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. [Gemma 2: Improving open language models at a practical size](#). *Preprint*, arXiv:2408.00118.
- Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Julian Rabelo, Ken Satoh, and Masaharu Yoshioka. 2023. [Summary of the competition on legal information, extraction/entailment \(colice\) 2023](#). In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ICAIL '23*, pages 472–480, New York, NY, USA. Association for Computing Machinery.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Aditya K, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. [Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 44123–44279. Curran Associates, Inc.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. [Don’t stop pretraining: Adapt language models to domains and tasks](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.
- Herbert Lionel Adolphus Hart. 1961. *The Concept of Law*. Clarendon Press, Oxford.
- Shreya Havaladar, Hamidreza Alvari, John Palowitch, Mohammad Javad Hosseini, Senaka Buthpitiya, and Alex Fabrikant. 2025. [Entailed between the lines: Incorporating implication into NLI](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32274–32290, Vienna, Austria. Association for Computational Linguistics.
- Matthew Henderson, Rami Al-Rfou, Brian Strope, Yunhsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. [Efficient natural language response suggestion for smart reply](#). *Preprint*, arXiv:1705.00652.
- Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. [A dataset for statutory reasoning in tax law entailment and question answering](#). In *Proceedings of the Natural Legal Language Processing Workshop 2020 co-located with the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2020), Virtual Workshop, August 24, 2020*, volume 2645 of *CEUR Workshop Proceedings*, pages 31–38. CEUR-WS.org.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. [Billion-scale similarity search with gpus](#). *IEEE Transactions on Big Data*, 7(3):535–547.
- Abhinav Joshi, Akshat Sharma, Sai Kiran Tanikella, and Ashutosh Modi. 2023. [U-CREAT: Unsupervised case retrieval using events extrAcTion](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13899–13915, Toronto, Canada. Association for Computational Linguistics.
- Do-Kyung Kim, Hyeseon Ahn, Youngwook Kim, and Yo-Sub Han. 2025. [Analyzing offensive language dataset insights from training dynamics and human agreement level](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9780–9792, Abu Dhabi, UAE. Association for Computational Linguistics.
- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. [CamemBERT: a tasty French language model](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.
- Mistral AI. 2024. Mistral NeMo. <https://mistral.ai/news/mistral-nemo>.
- Shounak Paul, Rajas Bhatt, Pawan Goyal, and Saptarshi Ghosh. 2024. [Legal statute identification: A case study using state-of-the-art datasets and methods](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24*, pages 2231–2240, New York, NY, USA. Association for Computing Machinery.
- Shounak Paul, Dhananjay Ghumare, Pawan Goyal, Saptarshi Ghosh, and Ashutosh Modi. 2025. [IL-PCSR: Legal corpus for prior case and statute retrieval](#). In *Proceedings of the 2025 Conference on Empirical*

Methods in Natural Language Processing, pages 14588–14611, Suzhou, China. Association for Computational Linguistics.

Shounak Paul, Pawan Goyal, and Saptarshi Ghosh. 2022. [LeSICiN: A heterogeneous graph-based approach for automatic legal statute identification from indian legal documents](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11139–11146.

Ellie Pavlick and Tom Kwiatkowski. 2019. [Inherent disagreements in human textual inferences](#). *Transactions of the Association for Computational Linguistics*, 7:677–694.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.

Jaromír Šavelka and Kevin D. Ashley. 2018. [Segmenting U.S. court decisions into functional and issue specific parts](#). In *Legal Knowledge and Information Systems – JURIX 2018: The Thirty-first Annual Conference*, pages 111–120. IOS Press.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. [Mpnnet: Masked and permuted pre-training for language understanding](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2019. [Crowdsourcing and validating event-focused emotion corpora for German and English](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, Florence, Italy. Association for Computational Linguistics.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. [Learning from disagreement: A survey](#). *Journal of Artificial Intelligence Research*, 72:1385–1470.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. [Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. [SWAG: A large-scale adversarial dataset](#)

[for grounded commonsense inference](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

A Bi-Encoder Training Details

This appendix expands the retrieval pipeline of §3.2. It details the sentence-aware chunking, explicit-pair extraction, data splits and FAISS indexing.

Sentence-aware chunking Motivation sections are segmented into chunks of up to 100 tokens using a custom algorithm. We split on sentence boundaries (., !, ?) while preserving legal abbreviations (e.g., *C. civ.*, *art.*, *al.*). Each chunk is made of up to two sentences if they fit within the token limit.

Explicit pair extraction To identify chunks that explicitly cite Civil Code articles, we use complementary methods. We apply a set of regular expressions designed to capture common variations in French legal citations, such as article ranges (e.g., *articles 1352 à 1352-9*), coordinated enumerations (*et*), and abbreviated references to French law Codes (*C. civ.*, *C. com.*). We compute TF-IDF cosine similarity between each chunk and the corresponding article text. We discard candidate matches when their similarity falls below 0.15. It helps eliminate accidental lexical overlaps. For each retained positive pair, we construct a negative counterpart. We sample randomly an article that has a TF-IDF similarity with the chunk below 0.05. We keep these negatives separate from the training data. They are used exclusively for evaluation purposes.

Data splits We split the 182,155 decisions into train (70%), validation (15%) and test (15%) by `decision_id`. We get 2,563,287 chunks for train, 555,658 for validation, and 553,082 for test. The bi-encoder is trained on 89,817 explicit positive pairs extracted from the training split (from 36,385 decisions). The validation set produces 19,642 positive pairs that were used for the selection of the threshold. The test split (553,082 chunks) is used exclusively for implicit candidate generation (§3).

Model architecture We use special tokens ([ARTICLE], [DECISION], [LOI], [MONTANT]) to fine-tune JuriBERT.

FAISS indexing We encode all 2,836 Civil Code articles (articles 1–2534). Embeddings are L2-normalized and indexed with IndexFlatL2 for exact search. On validation pairs, 95% of true matches fall within a squared L2 distance of 0.574. This is the retrieval threshold we have chosen.

B Article Renumbering

As mentioned in §3.3, the French Civil Code underwent a major reform in 2016, renumbering many articles in the contract law section (*Ordonnance* n°2016-131 of February 10, 2016). For example, the former article 1134 (binding force of contracts) became article 1103. Court decisions from our corpus (2023–2025) may cite and use either the old or new numbering depending on when the underlying contract was formed. To avoid false implicit candidates, we constructed an equivalence table mapping pre-2016 article numbers to their post-2016 counterparts. When a chunk’s candidate article is explicitly cited in the decision under either its old or new number, we exclude that candidate from the implicit pool. Our table covers approximately 150 article pairs and was compiled from official legislative sources. However, it is not exhaustive, as some articles were substantially modified rather than simply renumbered.

C Adversarial Filtering Prompt

The adversarial filtering step of §3.4 relied on the following prompt (in French, with English translation):

Tu es un juriste assistant spécialisé en droit français. Ta tâche consiste à déterminer si un extrait de décision judiciaire met en œuvre, applique ou reprend de manière implicite la règle de droit d’un article de loi donné, c’est-à-dire sans que le numéro ou la référence de l’article ne soit explicitement cité dans l’extrait, mais en reprenant son contenu, sa règle ou son principe. Réponds par ‘OUI’ ou ‘NON’ suivi d’une brève justification. IMPORTANT: Si tu n’es pas sûr ou qu’il y a un doute, réponds NON.

[You are a legal assistant with expertise in French law. Your task is to determine whether a court excerpt implicitly applies the rule contained in a given statutory article. The article must not be explicitly cited in the excerpt, but its substance, reasoning, or underlying principle may nonetheless be relied upon. Answer YES or NO, and provide a short justification. IMPORTANT: if there is any doubt, answer NO.]

D Dataset Statistics

The benchmark assembled in §3.5 is characterized here. The 1,015 pairs come from 829 decisions.

An article appears on average 2.43 times (median 2), and 92% appear five times or fewer.

Table 8 reports the length statistics for chunks and articles. Chunks average 50 words in the benchmark (median 50, max 88). Articles have 60 words on average in the benchmark (median 46). The distribution of the benchmark is very similar to the one of the full Civil Code (mean 62, median 46).

	Mean	Med.	Min	Max
Chunks (benchmark)	49.9	50	14	88
Articles (benchmark)	60.2	46	7	564
Articles (full Civil Code)	61.5	46	3	883

Table 8: Length statistics in words

E Annotation Interface

The annotation study of §4 used the interface described here, designed to ease the task while providing access to supplementary context when needed. For each entry, annotators were shown the candidate article together with its position in the Civil Code (book, title, chapter, and section), as well as the text chunk to be evaluated. Annotators could access the full court decision, in which the target chunk was visually highlighted.

Annotator instructions. The following instructions were given to the annotators: *L’article est-il appliqué implicitement ?* [Is the article implicitly applied?]. The response options were: *Oui* [Yes]; *Non* [No]; *Non, faits ou prétentions des parties uniquement* [No, facts or claims of the parties only]; *Non, application d’un régime spécial* [No, application of a special regime]; *Je ne sais pas* [I don’t know]; *À revoir* [To review].

F Structure of Disagreement

The disagreement structure summarized in §4 is detailed in Table 9, which reports how A_3 resolved each disputed case.

Resolution category (A_3)	N	%
NO – facts/party claims	147	43.4%
NO – residual category	106	31.3%
NO – special regime	61	18.0%
YES – implicit application	25	7.4%

Table 9: Structure of disagreement cases resolved by Annotator A_3 .

G Grid Search Details

The supervised encoders of §5.1 were tuned over the grid presented here:

Input configurations

- cfg1: [article_text] [SEP] [chunk]
- cfg2: [ARTICLE] [article_text] [SEP] [CHUNK] [chunk]
- cfg3: [ARTICLE] Article {num}: [article_text] [SEP] [CHUNK] [chunk]

Embedding extraction For encoders, we extract hidden states using four configurations: (i) mean pooling over last layer, (ii) mean pooling over penultimate layer (-2), (iii) mean pooling over average of last 4 layers (avg4), (iv) CLS token from last layer. For LLMs, we use configurations (i)–(iii) only.

Classification heads Logistic regression (LR), and MLP with one or two hidden layers (MLP1, MLP2). MLP1 has one hidden layer of size 256. MLP2 has two hidden layers of sizes 256 and 64.

TF-IDF baseline Unigrams and bigrams, with a minimum document frequency (min_df) of 2, capped at max_features of 50k, and sublinear TF scaling. IDF is computed on the training folds only. Logistic regression with SAGA solver, L2 regularization, and inverse regularization strength C set to 1.0.

Best configurations Table 10 reports the best configuration.

Encoder	Cfg	Pool	Layer	Head
SAUL-7B	cfg3	mean	avg4	LR
LLaMA-3.1-8B	cfg2	mean	-2	MLP1
LawMA-8B	cfg1	mean	last	MLP1
CamemBERT	cfg3	mean	avg4	MLP2
ST-MPNet	cfg1	mean	avg4	MLP2
CamemBERTav2	cfg1	mean	avg4	MLP1
TF-IDF	cfg2	–	–	LR
JuriBERT	cfg1	cls	last	LR
ST-MiniLM	cfg3	mean	-2	MLP2

Table 10: Best configuration per encoder (5-fold CV).

Search space We have a grid of 36 configurations per transformer and 27 per LLM. We use MCC calculated on out-of-fold predictions to select the best configuration.

Model	Thr	P	R	F1	MCC
SAUL-7B	.79	.73	.66	.69	.47
LLaMA-3.1-8B	.61	.73	.65	.69	.46
LawMA-8B	.84	.79	.54	.64	.46
ST-MPNet	.63	.76	.54	.63	.43
CamemBERT	.48	.70	.65	.67	.43
CamemBERTav2	.59	.71	.59	.65	.42
TF-IDF	.47	.73	.56	.63	.41
JuriBERT	.74	.69	.56	.62	.37
ST-MiniLM	.54	.69	.55	.61	.37
Ensemble	.61	.81	.62	.70	.53

Table 11: Supervised metrics (5-fold CV, positive class). Thr = threshold, P = precision, R = recall.

Full metrics Table 11 presents the complete metrics for all supervised models.

Confusion matrices Table 12 reports confusion matrices.

Model	TP	TN	FP	FN
Ensemble	278	499	66	172
SAUL-7B	296	456	109	154
LLaMA-3.1-8B	291	457	108	159
LawMA-8B	241	501	64	209
CamemBERT	293	437	128	157
CamemBERTav2	267	458	107	183
TF-IDF	251	473	92	199
ST-MPNet	244	486	79	206
JuriBERT	252	451	114	198
ST-MiniLM	248	454	111	202

Table 12: Confusion matrices for supervised models (5-fold CV, optimized threshold).

H Ensemble Search Details

We expand the ensemble search of §5.1. We use nested cross-validations to avoid data leakage while evaluating ensemble methods. Folds are grouped by decision_id. Different strategies are explored: logistic regression, weighted average and rank fusion.

Top 5 configurations Table 13 reports the best ensembles.

Ensemble metrics Our retained ensemble stacks CamemBERTav2 + JuriBERT + LLaMA + SAUL with a logistic regression meta-learner. Table 14 reports detailed metrics at the optimized threshold.

I Fine-tuned Cross-Encoder (BGE-reranker) Details

The fine-tuned BGE-reranker cross-encoder introduced in §5.1 is detailed here. We fine-tuned BGE-

Method	Models	MCC
Stacking-LR	Cam2+Juri+LLa+SAUL	.53
Stacking-LR	Cam+Juri+SAUL+Mini+MP+Law	.53
Stacking-LR	Cam+Juri+SAUL+Mini+MP+TF+Law	.53
Weighted Avg	Cam2+Juri+SAUL+Law	.53
Weighted Avg	Cam+Juri+SAUL+Mini+MP+Law	.53

Table 13: Top 5 ensembles. Cam=CamemBERT, Cam2=CamemBERTav2, Juri=JuriBERT, LLa=LLaMA, Mini=ST-MiniLM, MP=ST-MPNet, TF=TF-IDF, Law=LawMA.

Metric	Value
Models	Cam2+Juri+LLa+SAUL
Meta-learner	Logistic Regression
Threshold	.61
Accuracy	.77
Balanced Accuracy	.75
MCC	.53
Precision (yes/no)	.81 / .74
Recall (yes/no)	.62 / .88
F1 (yes/no)	.70 / .81
TP / TN / FP / FN	278 / 499 / 66 / 172

Table 14: Supervised ensemble metrics (5-fold nested CV).

reranker-v2-m3 as a binary classifier on the chunk-article pairs under a 5-fold cross-validation for 3 epochs. We use AdamW and a learning rate of 2×10^{-5} for the encoder body and 1×10^{-4} for the classification head.

It is the strongest of the purpose-built baselines, with an accuracy of 0.72, an F1 of 0.69 and an MCC of 0.43. Its YES-rate (at 47.4%) is close to the 44% base rate. See Table 15.

Metric	Value
Accuracy	0.72
F1 (positive)	0.69
MCC	0.43
YES-rate	47.4%
OR (disputed versus consensual TN)	2.22
95% CI	[1.51, 3.28]
p-value	0.0001

Table 15: Fine-tuned BGE-reranker-v2-m3

J Entailment (NLI) Baseline Details

The entailment (NLI) baseline introduced in §5.1 is described here. We use XLM-RoBERTa-large fine-tuned on XNLI (joeddav/xlm-roberta-large-xnli) and we fine-tune it on our benchmark with the same

grouped 5-fold cross-validation as the supervised classifiers in §5.1. Each (chunk, article) pair is given to the model as a premise and a hypothesis. The premise is the court excerpt (chunk) and the hypothesis is the sentence “*Ce passage applique la règle suivante : [article text]*” (“This passage applies the following rule: ...”). We tried three hypothesis templates and kept this one as the best performing, though the choice had little effect: all variants fell within 0.02 MCC of one another.

Fine-tuning uses AdamW at a learning rate of 1×10^{-5} , a batch size of 16, a maximum length of 512 tokens, for 4 epochs. Without fine-tuning, the model stays near chance (Table 16). On the other hand, the fine-tuning improves the MCC from 0.07 (zero-shot) to 0.35.

Method	F1	Acc	MCC	OR
XNLI zero-shot	.46	.55	.07	—
XNLI fine-tuned	.63	.68	.35	2.26
TF-IDF	.63	.71	.41	2.17
Sup. ensemble	.70	.77	.53	1.83

Table 16: NLI baseline versus paper models (positive class). TF-IDF and ensemble figures from Tables 3 and 14.

K Zero-Shot Prompts

The zero-shot evaluation of §5.2 used the two prompts reproduced here:

K.1 Standard Prompt

This prompt was used for non-reasoning models:

Tu es un expert en droit civil français. Ta tâche est de déterminer si un article du Code civil est implicitement appliqué dans un extrait de décision de justice. IMPORTANT:

- Réponds UNIQUEMENT par “oui” ou “non”
- “oui” = l’article est implicitement appliqué (le raisonnement juridique utilise cet article sans le citer)
- “non” = l’article n’est pas appliqué (simple mention des faits, ou autre régime juridique)

Ne donne aucune explication, juste “oui” ou “non”.

[English: You are an expert in French civil law. Your task is to determine whether a Civil Code article is implicitly applied in a court decision excerpt. IMPORTANT: Answer ONLY “yes” or “no”. “yes” = the article is implicitly applied (the legal reasoning uses this article without citing it). “no” = the article is not applied (mere mention of facts, or different legal regime). Provide no explanation, just “yes” or “no”.]

K.2 Reasoning Model Prompt

For Qwen3-32B with thinking mode, we use a different prompt:

Tu es un expert en droit civil français. Ta tâche est de déterminer si un article du Code civil est implicitement appliqué dans un extrait de décision de justice.

Analyse l'extrait en considérant:

1. Le raisonnement juridique utilisé par le juge
2. Les concepts juridiques mobilisés (même sans citation explicite)
3. La cohérence entre l'article proposé et le raisonnement de la décision

À la fin de ton analyse, conclus OBLIGATOIREMENT par une ligne contenant uniquement:

RÉPONSE: oui

ou

RÉPONSE: non

- “oui” = l'article est implicitement appliqué (le raisonnement juridique utilise cet article sans le citer)
- “non” = l'article n'est pas appliqué (simple mention des faits, ou autre régime juridique par exemple)

[English: You are an expert in French civil law. Your task is to figure out if a Civil Code article is implicitly applied in a court decision excerpt. Analyze the excerpt by looking at: (1) The legal reasoning the judge uses, (2) The legal concepts involved (even if not explicitly cited), (3) Whether the proposed article actually fits with what the decision is saying. At the end of your analysis, you MUST conclude with a line containing only: ANSWER: yes, or, ANSWER: no. “yes” = the article is implicitly applied (the legal reasoning relies on this article without citing it). “no” = the article is not applied (just factual mention, or different legal framework for instance).]

L Zero-Shot Full Results

These tables give the complete zero-shot results behind §5.2: Table 17 presents the zero-shot metrics against the gold standard while Tables 18 and 19 showcase the individual annotators.

M Retrieval-Augmented Few-Shot Prompting

Few-shot prompting, mentioned in §5.2, is evaluated on the three strongest zero-shot models: Qwen-2.5-32B, Qwen-2.5-7B and LLaMA-3.1-8B. Table 20 shows that few-shot prompting never approaches the supervised ensemble (MCC = 0.53).

For each query we retrieve $k/2$ positive and $k/2$ negative in-context examples by BM25 from the training fold (5-fold CV). We exclude the query's article. Few-shot brings at most +0.02 MCC for

Model	Yes	P	R	F1	Acc	MCC
Qwen2.5-32B	66	55	82	66	63	31
Qwen2.5-7B	54	57	70	63	63	28
LLaMA-3.1-8B	22	70	35	47	65	28
LLaMA-3.1-70B [†]	84	50	94	65	55	25
Mistral-Nemo-12B	75	50	85	63	56	21
Gemma-2-27B	94	46	98	63	49	17
Qwen3-32B [‡]	85	48	91	63	52	16
Aya-Expansive-32B	94	45	96	62	47	8
Command-R-35B	96	45	97	61	46	4
SAUL-7B	95	45	95	61	46	3

Table 17: Zero-shot results vs. gold ($n = 1,015$), positive class, in %. [†]4-bit quantization. [‡]Reasoning model.

Model	Yes	P	R	F1	Acc	MCC
Qwen2.5-32B	66	61	80	69	64	30
Qwen2.5-7B	54	62	67	64	63	26
LLaMA-3.1-70B [†]	84	55	93	69	59	24
LLaMA-3.1-8B	22	71	31	44	59	23
Mistral-Nemo-12B	75	55	83	66	58	18
Qwen3-32B [‡]	85	54	91	67	56	17
Gemma-2-27B	94	52	97	68	54	15
Aya-Expansive-32B	94	51	96	66	52	7
Command-R-35B	96	50	97	66	51	4
SAUL-7B	95	51	95	66	51	4

Table 18: Zero-shot results vs. A₁, positive class, in %. [†]4-bit quantization. [‡]Reasoning model.

Qwen-2.5-7B. It degrades LLaMA-3.1-8B by pushing its YES-rate from 22% toward 50% without improving discrimination.

N Unsupervised Ensemble: Weights and Sensitivity

We present the heuristic of §5.3 and report the sensitivity study. The ranking score is:

$$S = 0.25 \sum_{i=1}^4 v_i + \sum_{j=2}^4 \beta_j \mathbf{1} \left[\sum_i v_i \geq j \right] + 0.2 \text{TF-IDF} + 0.2 \text{BM25} + 0.4 \text{CE}.$$

where $v_i \in \{0, 1\}$ is LLM i 's vote. $\beta_2 = 0.3$, $\beta_3 = 0.5$ and $\beta_4 = 1.0$ are cumulative so that the unanimous bonus is $0.3 + 0.5 + 1.0 = 1.8$. TF-IDF, BM25 and cross-encoder (CE) scores are normalized to $[0, 1]$. Ties are broken by $\varepsilon \cdot \text{CE}$ with $\varepsilon = 10^{-6}$.

We test whether the ranking score is sensitive to the exact choice of weights. Overall, it is not, as the following results show:

Model	Yes	P	R	F1	Acc	MCC
Qwen2.5-32B	66	76	74	75	67	26
Qwen2.5-7B	54	75	60	67	60	18
LLaMA-3.1-70B [†]	84	71	88	79	68	17
Gemma-2-27B	94	69	97	81	69	17
Mistral-Nemo-12B	75	71	80	76	65	16
LLaMA-3.1-8B	22	81	27	40	47	16
Qwen3-32B [‡]	85	70	88	78	67	14
SAUL-7B	95	68	96	79	67	6
Command-R-35B	96	67	97	79	67	4
Aya-Expansive-32B	94	67	94	78	65	1

Table 19: Zero-shot results vs. A₂, positive class, in %.
[†]4-bit quantization. [‡]Reasoning model.

Model	$k=0$	$k=2$	$k=4$	$k=8$
Qwen-2.5-32B	.309	.306	.324	.324
Qwen-2.5-7B	.275	.294	.267	.248
LLaMA-3.1-8B	.284	.213	.210	.212

Table 20: Few-shot MCC on gold ($n = 1,015$; Qwen-2.5-32B at $k=8$; $n = 1,008$). k in-context examples retrieved by BM25.

- Scaling the LLM block by a factor λ keeps AP almost unchanged across a wide range of values. AP is 0.668 ± 0.002 for $\lambda \in [0.25, 10]$ ($P@200 = 0.755$). However, entirely removing the LLM block does affect performance, since it then drops ($\lambda = 0$, $AP = 0.604$). $\lambda \rightarrow \infty$ is the symmetric ablation. When we keep the LLM block only (we set the TF-IDF, BM25 and cross-encoder weights to zero), we notice a slight increase ($\lambda \rightarrow \infty$, $AP = 0.675$).
- Perturbing all weights randomly by factors in $[0.5, 1.5]$ over 200 configurations shows an AP standard deviation of only 0.001.
- Removing any single component changes AP by at most 0.005.

O Unsupervised Ranking Details

Complementing the unsupervised ranking of §5.3, this appendix gives the tie-breaking rule and the average-precision scores for each annotator.

Tie-breaking There were ties because of the binary outputs (0 or 1) of the LLMs. To break them we used a trace from the cross-encoder. The final score is $\text{Score}_{\text{final}} = \text{Score}_{\text{LLMs}} + \varepsilon \times \text{Score}_{\text{CE}}$, with $\varepsilon = 10^{-6}$.

Full results by annotator Figure 2 (main text) reports the gain in true positives on gold labels. Figures 4 and 5 show the same curves for A₁ and A₂ individually.

Table 21 reports AP for each annotator separately. A₂ says yes more often (67% positive) and thus shows higher AP values. The relative ranking of methods is nonetheless consistent across annotators.

Method	A ₁	A ₂	Gold
Random	.50	.67	.44
TF-IDF	.61	.74	.57
BM25	.61	.73	.57
Cross-Encoder	.63	.76	.61
LLM_Inter4	.67	.78	.66
Unsupervised Ensemble	.69	.79	.67

Table 21: Average Precision by annotator for all ranking methods.

Coverage analysis Tables 22 and 23 display recall at diverse cutoffs. At $k = 300$ (29.6% of the corpus), the unsupervised ensemble retrieves 47% of all true positives. Table 24 details the FP/disagreement breakdown.

k	Unsup. Ensemble	Inter4	CrossEnc.	Random
100	17%	17%	16%	10%
200	34%	33%	30%	20%
300	47%	47%	42%	30%
500	69%	65%	62%	49%

Table 22: Recall at k (%) for gold labels.

k	Unsup. Ensemble	Inter4	CrossEnc.	Random
100	77	76	72	44
200	151	150	137	89
300	211	213	189	133
500	310	294	281	222

Table 23: True positives retrieved at k (out of 450 total).

Results by annotator Figures 4 and 5 show the gain in true positives for A₁ and A₂. The unsupervised ensemble is the best method for both annotators. The gain in absolute value changes with the annotator.

P Per-Model False Positive Rates

These are the per-model false-positive rates behind the odds ratios reported in §6. Table 25 reports false positive rates on agreement versus disagreement for

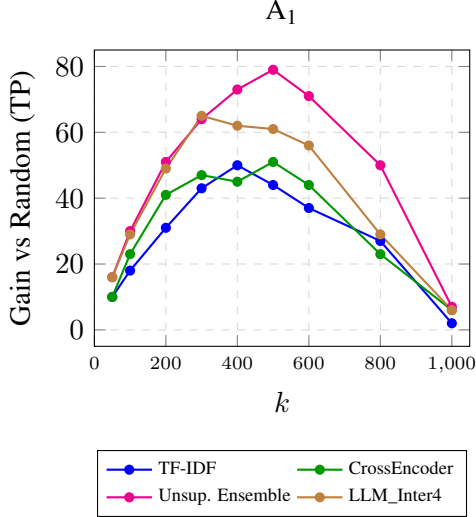


Figure 4: Gain in true positives vs. random ranking (A_1).

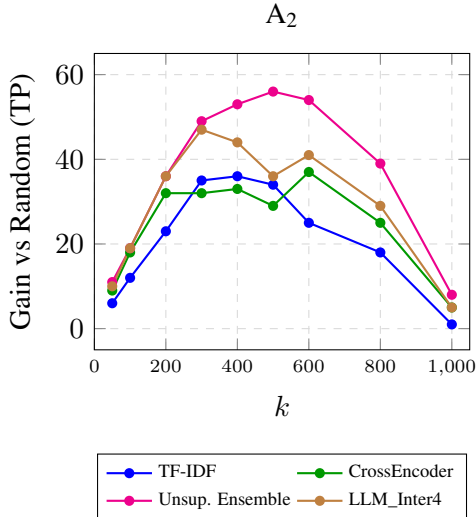


Figure 5: Gain in true positives vs. random ranking (A_2).

all individual models and the supervised ensemble. Figure 3 (main text) represents the same data.

Q Calibration Details

The calibration gap discussed in §6 is detailed in Table 26, which reports Expected Calibration Error (ECE) for all models. It is computed separately on agreement and disagreement cases. All models show higher ECE on disagreement cases. Table 27 provides a breakdown by confidence bin for the retained supervised ensemble.

R Controlling for Surface Confounds

To support the claim in §6 that error concentration reflects a genuine effect rather than a surface arti-

k	FP (Agree)	FP (Disagree)	% Disagree
50	2	10	83%
100	8	15	65%
200	20	29	59%

Table 24: False positive origin by annotator agreement. Disagreement cases (33% of data) account for 59–83% of FPs until $k = 200$.

Model	FPR _{ag}	FPR _{dis}	OR	p
LawMA-8B	6.0%	15.6%	2.91	<.001
ST-MPNet	8.0%	18.8%	2.67	<.001
CamemBERT	14.3%	29.3%	2.48	<.001
LLaMA-3.1-8B	12.7%	24.2%	2.19	<.001
TF-IDF	10.8%	20.7%	2.17	.001
JuriBERT	14.3%	24.8%	1.97	.002
SAUL-7B	14.7%	22.9%	1.72	.013
CamemBERTav2	15.5%	21.7%	1.50	.058
ST-MiniLM	17.5%	21.3%	1.28	.253
Ensemble	8.4%	14.3%	1.83	.030

Table 25: False positive rates on agreement (FPR_{ag}, $n = 251$) vs. disagreement (FPR_{dis}, $n = 314$) subsets. Seven models and the ensemble significant after FDR correction ($p < 0.05$).

fact, we test whether disagreement-related errors can instead be explained by surface features. We restrict our analysis to true negatives (gold = NO, $n = 565$). We regress a false-positive indicator on a disagreement indicator. We test three nested specifications: M0 (disagreement only), M1 (+ log chunk and article length), and M2 (+ TF-IDF overlap). The groups have very similar surface statistics. The disagreement odds ratio barely moves once the controls are added (Table 28): at most 0.08 across models, and $1.83 \rightarrow 1.85$ for the ensemble. Surface properties do not explain the concentration of errors on disputed cases.

S False Negatives by Agreement

The symmetric false-negative analysis noted in §6 is reported in Table 29. We repeat the agreement/disagreement split on false negatives (gold = YES). The supervised ensemble has FNR 37.9% on agreement versus 44.0% on disagreement (OR = 1.29, $p = 0.53$). The direction matches the false-positive analysis. However, there are only 25 positives within the disagreement subset. This is a structural consequence of the 92.6% of disputed cases resolving to NO. As a consequence, statistical power is limited. Six of nine individual classifiers show OR > 1. None is individually significant.

Model	ECE _{ag}	ECE _{dis}	Δ
TF-IDF	.17	.34	+ .17
ST-MiniLM	.16	.31	+ .15
Ensemble	.15	.30	+ .15
CamemBERTav2	.15	.30	+ .14
ST-MPNet	.16	.30	+ .14
CamemBERT	.16	.29	+ .13
LLaMA-3.1-8B	.17	.28	+ .11
JuriBERT	.21	.32	+ .11
Lawma-8B	.18	.28	+ .10
SAUL-7B	.20	.29	+ .09

Table 26: Expected Calibration Error by agreement status.

Bin	Agreement			Disagreement		
	Acc.	Gap	n	Acc.	Gap	n
0.1–0.2	.26	+ .12	175	.03	– .12	117
0.2–0.3	.39	+ .14	66	.06	– .19	54
0.3–0.4	.54	+ .19	65	.00	– .35	33
0.4–0.5	.70	+ .25	43	.07	– .38	43
0.5–0.6	.55	+ .01	38	.07	– .48	29
0.6–0.7	.82	+ .16	73	.13	– .51	15
0.7–0.8	.94	+ .18	93	.12	– .64	24
0.8–0.9	.98	+ .13	122	.38	– .47	24

Table 27: Full calibration for supervised ensemble.

T Additional Qualitative Examples

Beyond the three cases analyzed in §7, this appendix presents six annotated examples. All predictions were made using the supervised ensemble (Stacking-LR: CamemBERTav2 + JuriBERT + LLaMA + SAUL).

T.1 True positive, agreement: Art. 2274 (good faith presumption)

A₁=A₂=Yes. Gold=Yes. Conf. 0.75 ✓

Article “*La bonne foi est toujours présumée, et c’est à celui qui allègue la mauvaise foi à la prouver.*” [Good faith is always presumed, and it is for the party alleging bad faith to prove it.]

Excerpt “*Néanmoins, ce seul constat de l’accroissement de la dette locative ne suffit pas à caractériser par lui seul la mauvaise foi de la débitrice.*” [Nevertheless, the mere observation of the increase in rental debt is not sufficient in itself to establish the bad faith of the debtor.]

Analysis The presumption of the article 2274 is applied by the court that requires affirmative proof of bad faith. Circumstantial evidence (merely observing increased debt) does not qualify. Both annotators recognized the application and the model identified it properly with satisfying confidence.

Model	M0	M1	M2
LawMA-8B	2.91	2.91	2.92
ST-MPNet	2.67	2.70	2.75
CamemBERT	2.47	2.52	2.53
LLaMA-3.1-8B	2.19	2.19	2.21
TF-IDF	2.17	2.16	2.17
JuriBERT	1.97	1.97	1.97
SAUL-7B	1.72	1.72	1.72
CamemBERTav2	1.50	1.50	1.51
ST-MiniLM	1.28	1.27	1.27
Ensemble	1.83	1.83	1.85

Table 28: Odds ratio of false positive for disagreement versus agreement on true negatives ($n = 565$)

Subset	Cases	TP+FN	FN	FNR
Agree	676	425	161	37.9%
Disagree	339	25	11	44.0%

Table 29: False negatives by annotator agreement (gold = YES). FNR = FN / (TP+FN).

T.2 True positive, agreement: Art. 1310 (solidarity not presumed)

A₁=A₂=Yes. Gold=Yes. Conf. 0.71 ✓

Article “*La solidarité est légale ou conventionnelle; elle ne se présume pas.*” [Solidarity is statutory or contractual; it may not be presumed.]

Excerpt “*La solidarité ne se présume pas et qu’aucun texte ne l’institue pour le paiement des charges de copropriété du seul fait de la nature de cette dette.*” [Solidarity is not presumed, and no text establishes it for the payment of co-ownership charges solely by virtue of the nature of this debt.]

Analysis The same words as the article 1310 (“ne se présume pas”) are used by the court. It rejects the existence of solidarity between co-owners and also outlines that no statutory provision establishes solidarity. Doing so, it reaffirms that solidarity requires an explicit legal or contractual basis. Both annotators agreed on an implicit citation. The model was right as well.

T.3 True negative, agreement: Art. 1241 (tort liability)

A₁=A₂=No. Gold=No. Conf. 0.27 ✓

Article “*Chacun est responsable du dommage qu’il a causé non seulement par son fait, mais encore par sa négligence ou par son imprudence.*” [Every person is liable for the damage they have caused not only by their own act, but also by their negligence or imprudence.]

Excerpt “*Le principe du droit à indemnisation in-*

tégrale du demandeur est en revanche contesté par l'assureur qui reproche à M. [F] [S] d'avoir commis des fautes à l'origine de son dommage." [The principle of the plaintiff's right to full compensation is, however, contested by the insurer, who accuses Mr. [F] [S] of having committed faults that caused his damage.]

Analysis There are some semantic similarity ("fautes," "dommage"), but this extract reports the insurer's position and not a judicial reasoning. The court only describes what a party claims. Both annotators identified this as a factual chunk and the model rightfully assigned low confidence.

T.4 False positive, disagreement: Art. 2251 (renunciation of prescription)

A₁=Yes, A₂=No, A₃=No. Gold=No. Conf. 0.69 ×

Article "*La renonciation à la prescription est expresse ou tacite. La renonciation tacite résulte de circonstances établissant sans équivoque la volonté de ne pas se prévaloir de la prescription.*" [Prescription may be renounced expressly or tacitly. A tacit renunciation is inferred from circumstances that unequivocally show an intent to forgo the defense.]

Excerpt "*Elle conteste toute renonciation à son droit de se prévaloir de la forclusion de l'action des époux [R] qui ne peut se présumer et doit résulter, lorsqu'elle est tacite, d'une volonté de renoncer non équivoque.*" [She denies having waived her right to raise the time-bar against the [R] spouses' claim. This waiver cannot be presumed; if tacit, it must stem from an unequivocal intent to give up that defense.]

Analysis A₁ noticed the legal standard being invoked but A₂ and A₃ noted that "Elle conteste..." is a party's argument. The court reports what one party argues, not applying the rule itself. The model incorrectly predicted this case. Cases like this illustrate that it can be difficult to distinguish the description of facts and party claims from the actual application of law to those facts.

T.5 False negative, disagreement: Art. 1118 (acceptance of offer)

A₁=Yes, A₂=No, A₃=Yes. Gold=Yes. Conf. 0.47 ×

Article "*L'acceptation est la manifestation de volonté de son auteur d'être lié dans les termes de l'offre...*" [Acceptance is the expression of its author's intention to be bound by the terms of the offer...]

Excerpt "*Cette affirmation est exacte, mais la bonne réception de la demande de Monsieur [E] [L] n'en signifie pas pour autant l'acceptation.*" [This assertion is correct, but the proper receipt of Mr. [E] [L]'s request does not thereby signify acceptance.]

Analysis Article 1118 core principle is the distinction between receipt and acceptance. It is a key concept of contract formation. Receiving a communication is different from acceptance without a manifestation of intent to be bound. A₂ may have focused on the absence of the article characteristic vocabulary. A₁ and A₃ recognized the underlying legal distinction being applied. The model was wrong but exhibits moderate confidence. Detecting implicit citations is more difficult when the statutory language is not echoed in the decision.

T.6 False positive, agreement: Art. 606 (major repairs)

A₁=A₂=No. Gold=No. Conf. 0.80 ×

Article "*Les grosses réparations sont celles des gros murs et des voûtes, le rétablissement des poutres et des couvertures entières...*" [Major repairs comprise those to main walls and vaults, the restoration of beams and entire roofs...]

Excerpt "*Le bail commercial précise au titre de l'entretien et des réparations que « le bailleur aura à sa charge les réparations afférentes aux gros murs et voûtes, le rétablissement des poutres et des couvertures entières...*" [The commercial lease specifies, under the heading of maintenance and repairs, that the lessor shall bear the cost of repairs relating to main walls and vaults, the restoration of beams and entire roofs...]

Analysis Both annotators saw that the court is citing the *contract* rather than applying the *statute*. The lease mirrors the language of article 606, which is a standard practice in commercial leases. However, the court is interpreting a contractual clause, not applying the Civil Code article. This is a type of model failure when there is an almost perfect lexical match that is deceptive (detailed in §U).

U Quantitative Error Analysis

To complement the qualitative analysis of §7, we manually evaluated the 66 false positives produced by the supervised ensemble. Table 30 shows the distribution. We focus on false positives because their failure modes leave observable traces in the chunk. False negatives, on the other hand, would

require reasoning about absent evidence.

Failure mode	N	%	Disag.	Conf.
Statutory language present, not applied by the court	37	56	28	0.75/0.76
Right legal domain, wrong rule	28	42	17	0.76/0.76
Other	1	2	0	0.71/0.71
Total	66	100	45	0.75/0.76

Table 30: Failure modes of the supervised ensemble on its 66 FP. Confidence shown as mean/median.

The first category corresponds to cases where statutory language is used in the chunk but not applied by the court. Often, the article’s vocabulary is echoed in a party’s claim or a contract clause. The model mistakes the presence of legal language for its application (cf. the art. 2251 and art. 606 examples in Appendix T). The second category of errors involves cases where the legal domain is the correct one but the rule retrieved is not the appropriate one. It can be a sibling article of the same domain or a special regime that displaces the general article (cf. the art. 1361 example in §7). One case fits neither pattern. On 45 of the 66 pairs, the annotators disagreed, which is consistent with §6.