Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Reading Royalty Risk at Scale: LLM-Assisted Construction and Reliability Analysis of an Offset-and-Deduction Clause Corpus from SEC-Filed Technology Licences

Yuyong Kim

ykim288@wisc.edu

University of Wisconsin–Madison

Research Article

Keywords: royalty offset and deduction clauses, technology licensing, SEC EDGAR, legal language resource, LLM-assisted annotation, annotation reliability

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10473033/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Reading Royalty Risk at Scale: LLM-Assisted Construction and Reliability Analysis of an Offset-and-Deduction Clause Corpus from SEC-Filed Technology Licences

Yuyong Kim University of Wisconsin–Madison, Madison, WI 53706, USA
ykim288@wisc.edu · ORCID 0009-0006-4842-666X

Abstract

We introduce **RODEO**, a corpus of 5,111 royalty offset and deduction clause spans harvested from public SEC EDGAR EX-10 material-contract exhibits (709 exhibits, 417 companies) and labelled for their economic function under a seven-function taxonomy. Because exhaustive human annotation of such clauses is costly, we build the resource with an LLM-assisted protocol — independent double classification, adjudication of substantive disagreements, and a stratified human spot-check — and release the corpus, the per-pass and adjudicated labels, and the labelling code, with source identifiers and redistribution-safe excerpts. As a quality-control analysis of the protocol we examine how far the model can be trusted. It is highly self-consistent on the surface (82.4% exact-set agreement across passes) yet poorly calibrated: 82% of predictions carry the maximal confidence score, even that top bin flips label across passes 13.8% of the time, and the model's own uncertainty flag recovers only 1% of genuine disagreements. On a deliberately hard subset, a cross-model check with an independent model family agrees with the adjudicated label only 36.6% of the time, whereas across a broad stratified sample the two independent LLM annotators agree substantially (macro Cohen's $\kappa = 0.78$) — hardest-boundary difficulty and broad-sample stability, respectively. Inter-pass disagreement, not self-reported confidence, is therefore the more useful signal for triaging hard cases in silver annotation — a practical protocol for silver annotation where full human labelling is infeasible. Supervised baselines characterise the corpus's learnability (a TF-IDF logistic-regression classifier reaches macro-F1 0.80; the gating condition NDC is hardest). We are explicit that the released labels are LLM-generated with a human spot-check ("silver") and set out the path to a fully human-adjudicated version.

Keywords: royalty offset; royalty deduction; technology licensing; SEC EDGAR; language resource; contract clause annotation; legal NLP; LLM reliability; LLM calibration

1. Introduction

Royalty-risk allocation clauses decide how much a licensee actually pays once third-party royalties, patent costs, and base definitions are taken into account. Their economic effect

— whether the effective rate rises, falls, or is bounded — turns on fine distinctions that are easy to state and hard to apply: a *credit* against computed royalties (a third-party royalty credit) is not the same as an *exclusion* from the royalty base, and a floor on the *effective rate* (a stacking cap) is not the same as a guaranteed minimum *amount*.

This paper operationalises a seven-function taxonomy of these clauses as a classification task and applies it, at corpus scale, to real licence exhibits. Two questions drive the work. First, can the taxonomy be applied by machine over messy filing text, and how learnable is it? Second — where our contribution ultimately lands — how reliably do contemporary LLMs read these clauses, and can we locate where they fail without an expensive full human annotation?

We make four contributions: (1) **a corpus** of 5,111 royalty-risk clause spans harvested from EX-10 material-contract exhibits on SEC EDGAR (709 exhibits, 417 companies), released with source identifiers and redistribution-safe excerpts; (2) **an LLM-assisted annotation protocol** — independent double classification plus disagreement adjudication — that yields per-span labels, rationales, and a self-consistency signal; (3) **a reliability and calibration analysis** showing that the model's self-reported confidence is a poor difficulty signal whereas inter-pass disagreement is a good one, and that the disagreements concentrate on one specific clause boundary; and (4) **reference baselines** establishing that the taxonomy is strongly learnable from surface features.

The released labels are LLM-generated with a human spot-check; we treat them as a silver resource and set out the path to a human-gold benchmark in the Limitations.

2. Related work

Contract and clause NLP. Automated analysis of contract language has produced a series of annotated resources. CUAD (Hendrycks et al., 2021) labels 41 clause types across 510 contracts for clause *presence*; LEDGAR (Tuggener et al., 2020) offers large-scale provision-type classification; LexGLUE (Chalkidis et al., 2022) aggregates legal-language understanding tasks; ContractNLI (Koreeda and Manning, 2021) casts contract review as document-level natural-language inference; and MAUD (Wang et al., 2023) targets merger-agreement reading comprehension. These operate largely at the level of clause *type* or *presence*. RODEO instead classifies the *economic function* of a narrow, high-stakes family — royalty offset and deduction mechanisms in technology-licence exhibits — where the operative distinctions (a credit against computed royalties vs. an exclusion from the royalty base; a floor on the effective rate vs. a guaranteed minimum amount) are subtle, adversarial to keyword matching, and financially consequential.

LLMs for legal reasoning. Large models are increasingly benchmarked on legal tasks: LegalBench (Guha et al., 2023) assembles 162 legal-reasoning tasks, and GPT-4 has been reported to pass the Uniform Bar Exam (Katz et al., 2024). This line measures *capability* on reasoning and question answering; we instead study the *reliability* of a labelling judgement over real contract text.

Hallucination and reliability of legal LLMs. Dahl et al. (2024) profile legal hallucination in major LLMs and report rates of 58–88% on U.S. case-law queries, observing that models frequently cannot tell when they are hallucinating; related NLLP work scrutinises machine-generated legal analysis and the groundedness of legal question answering. Our contribution is complementary and operates at the clause level: on a real licensing corpus we quantify that a model's *self-reported* confidence is a poor guide to where it errs, and that cross-pass disagreement is a better one.

LLMs as data annotators. A growing literature uses LLMs to pre-label or fully label data, sometimes matching or surpassing crowd workers (Gilardi et al., 2023; Ziems et al., 2024) while cautioning about inherited bias, prompt sensitivity, and over-confidence. We contribute a concrete recipe for the legal-clause setting — double-pass classification with disagreement adjudication and a human spot-check — together with an analysis of where it fails.

Calibration and self-consistency. Neural classifiers are known to be mis-calibrated (Guo et al., 2017); LLMs partly "know what they know" (Kadavath et al., 2022), yet their verbalised confidence is often unreliable (Tian et al., 2023). Self-consistency across sampled generations improves reasoning accuracy (Wang et al., 2023). We repurpose self-consistency at the *annotation-quality* level, using inter-pass agreement — rather than the model's own confidence or ambiguity flag — as the signal for locating hard items, and we report a calibration curve that makes the gap explicit.

Royalty analysis and computable contracts. Royalty-rate assessment has traditionally relied on manual review of comparable agreements (Razgaitis, 2003; Parr, 2007). Surden's (2012) notion of computable contracts frames a follow-on formal model that maps clauses to effective royalty rates (§8). RODEO supplies the large-scale, function-labelled clause data that such doctrinal and computational analyses have lacked.

3. Corpus construction

Source and retrieval. Licence agreements are filed as Exhibit-10 material contracts, predominantly by smaller pharmaceutical, biotechnology, and specialty-chemical firms rather than by large filers. We queried SEC EDGAR full-text search with clause-level seed phrases ("credit against royalties", "royalty stacking", "in no event … royalty", "minimum annual royalty", "net sales shall exclude", and others), restricted results to `EX-10.*` document types, downloaded the matching exhibits, and extracted ±350-character spans around each royalty-risk keyword.

Scale and composition. The corpus comprises **5,111 candidate clause spans** from **709 EX-10 exhibits** across **417 companies**. Industry concentration matches expectations for technology licensing (Table 1).

Table 1. Corpus composition by filer SIC (top classes).

SIC	Industry	Spans

SIC	Industry	Spans
2834	Pharmaceutical preparations	1,912
2836	Biological products	514
2844	Specialty chemicals / prep	298
8731	Commercial R&D	191
3841	Medical devices	172

Redaction. Confidential financial terms are redacted as [***]. Because our labels concern clause *function* rather than the numeric rate, redaction does not impede annotation: the structure ("in no event … less than [***]% of Net Sales") remains fully legible. **Redistribution safety.** We do not redistribute exhibit text; the release carries accession number, exhibit number, character offsets, and short excerpts, with a script to reconstruct spans from EDGAR.

4. Task and label set

The seven functional labels are grounded in the practitioner literature on royalty structuring and drafting (Razgaitis, 2003; Parr, 2007) and were refined by inductive review of the corpus. Each captures a distinct mechanism by which a clause changes the *effective* royalty — a credit against computed royalties (TPRC), an exclusion from the royalty base (RBE), a floor on the effective rate (RSC) or a guaranteed minimum amount (MRF), a sublicensing-revenue adjustment (SRO), or a patent-cost credit (PCO). NDC is treated separately, as a *gating condition* rather than a standalone mechanism, because notice and documentation requirements modulate whether an offset applies and therefore co-occur with the substantive labels rather than replacing them. The boundaries were chosen to separate operations that are economically distinct even when textually similar (credit vs. base-exclusion; rate-floor vs. amount-floor), which is also where the taxonomy is hardest to apply (§6).

Each span is assigned one or more of seven functional labels, plus two housekeeping labels:

- **TPRC** — third-party royalty credit/offset;
- **RSC** — royalty stacking cap (a floor on the *effective rate*);
- **RBE** — royalty-base exclusion (from the Net Sales definition);
- **SRO** — sublicensing-revenue offset;
- **MRF** — minimum royalty floor (a guaranteed *amount*);
- **PCO** — patent-cost offset;
- **NDC** — notice/documentation condition (gates an offset; co-occurs with another label);
- **O** — not a royalty-risk-allocation clause (keyword-matched noise);
- **UNCLEAR** — text too truncated/ambiguous to determine.

The task is multi-label. Three confusion pairs — TPRC/RBE, RSC/MRF, TPRC/PCO — are the a-priori loci of difficulty. Table 2 grounds each label in a real clause from the corpus.

Table 2. Representative clauses for each functional label, with the filer and its EX-10 exhibit (numeric terms are redacted as [***] in the filings).

Label	Filer · exhibit	Representative excerpt
TPRC	ATAI Life Sciences · EX-10.8	"*Third Party Royalties: If the Company … makes payment of Third Party Royalties …*" (creditable against royalties due)
RSC	T2 Biosystems · EX-10.13	"*Royalty Stacking. … In the event that Company is required to make [***] payments to one or more third parties …*"
RBE	Vir Biotechnology · EX-10.28	"*a royalty in respect of Net Sales (as defined below) of Licensed Products covered by Valid Claims …*"
SRO	Islet Sciences · EX-10.1	"*… in respect of such … sublicense … after first deducting any amounts incurred … in settlement or defense*"
MRF	Smart for Life · EX-10.37	"'*Guaranteed Minimum Royalty' … non-returnable advances recoupable against Royalties earned …*"
PCO	Liquidia Technologies · EX-10.18	"*[costs of] Licensee's use of … Patent Rights shall be offset against future royalties, fees, or other payments …*"
NDC	Wyndham Hotels & Resorts · EX-10.4	"*… together with any supporting documentation reasonably requested … [for] the calculation … of the Royalty amount …*"

5. LLM-assisted annotation protocol

Model and output format. We label every span with DeepSeek (chat, served as v4-flash) through a structured-JSON interface. A single system prompt encodes the nine label definitions and an explicit decision tree, and instructs the model to return, for each span, (i) the label set, (ii) an effect direction (up / down / cap / floor / none), (iii) a self-reported confidence on a 1–5 scale, (iv) a boolean *ambiguous* flag, (v) the competing labels it considered, and (vi) a short natural-language rationale in English and Korean. Structured decoding (JSON mode) and a bounded token budget keep outputs parseable; requests are issued concurrently and the run is resumable at the span level.

Three passes. *Pass 1* runs at temperature 0 — the model's most-likely reading. *Pass 2* re-runs the identical prompt at temperature 0.7 as an independent draw; this is a self-consistency probe that exposes where a label is stochastically unstable rather than determined by the text. *Pass 3* is invoked only where the two passes disagree on a *substantive* functional label: a senior-adjudicator prompt is shown both label sets and both rationales and must either commit to a final label set, with its own confidence, or mark the span `needs_human` when the excerpt is genuinely undecidable.

Disagreement tiers. Not all inter-pass disagreements matter equally, so we tier them by the symmetric difference of the two label sets: *NDC-only* (the passes differ solely on the presence of a gating condition), *inclusion-boundary* (one pass says O, the other a substantive label), and *substantive swaps* (they differ on a functional label, e.g. RSC vs. MRF). Only substantive swaps, together with the few self-flagged spans, are sent to adjudication (§6); NDC-only and agreeing spans keep the Pass-1 label.

Human spot-check and merge. A human reviewer inspected a stratified spot-check — adjudicator-resolved disagreements plus per-label samples — through a purpose-built web interface that shows the excerpt, the model's suggestion, and its rationale, and lets the reviewer accept the suggestion with a single keystroke or override it. Decisions are auto-saved in the browser and persisted server-side. The final label of each span follows a fixed priority: human decision, then adjudication, then Pass 1.

6. Results

Noise filtering. Keyword retrieval is deliberately high-recall; the model labels **1,486 of 5,111 spans (29%) as O** (non-royalty-risk mentions — tax withholding, generic payment timing, definitional cross-references). The remaining **3,625 spans** are substantive clauses. The distribution over the seven functional labels (multi-label) is MRF 1,790 · TPRC 1,169 · RSC 946 · NDC 634 · RBE 179 · SRO 150 · PCO 103 (Figure 1); O (1,486) and UNCLEAR (8) are housekeeping labels, not part of the taxonomy, and are omitted from the figure.

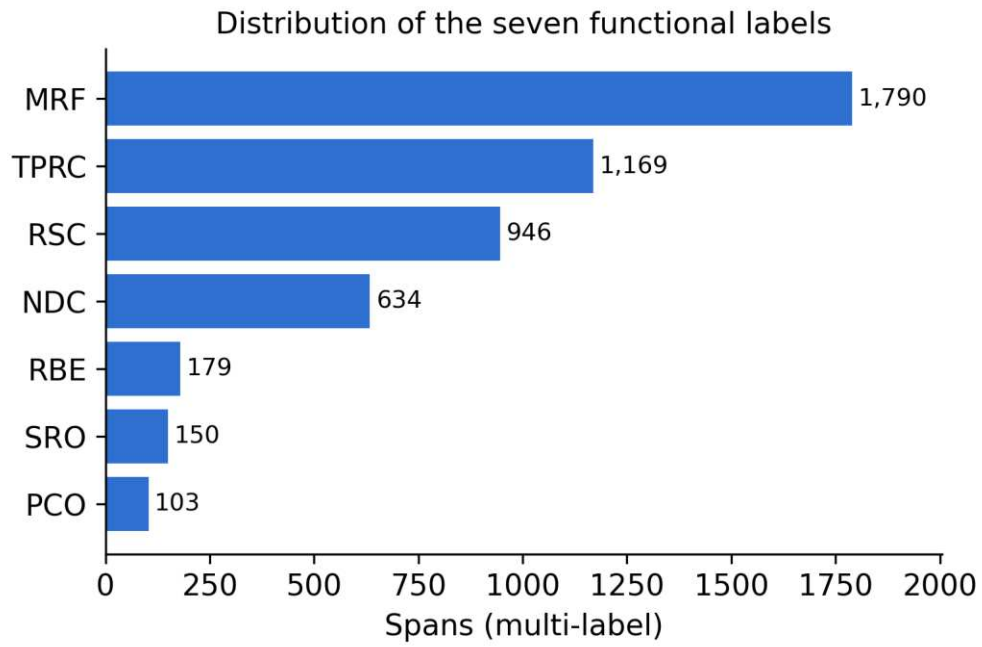


Figure 1. Distribution of the seven functional labels (multi-label). The housekeeping labels O (1,486 non-royalty-risk spans) and UNCLEAR (8) are reported in the text, not shown.

Self-consistency vs. confidence — the central finding. Passes 1 and 2 produce the *exact same label set* on **82.4%** of spans (4,210/5,111); per-label presence agreement is 91–99.7% (Table 3). The model's self-reported confidence is *weakly* informative but badly concentrated and mis-scaled: 82% of Pass-1 predictions carry the maximal confidence of 5, only 17 spans score ≤3, and the model self-flags ambiguity in just 0.2% of cases. Crucially (Table 4), even the maximal-confidence bin is unstable — its predictions flip across passes **13.8%** of the time — while the confidence-4 bin flips **35.0%**. Confidence-5 is thus not the same as reliable, and the model's own ambiguity flag is useless for triage: it recovers only **1%** of the genuine inter-pass disagreements (recall 0.01, precision 0.40). The usable difficulty signal is **inter-pass disagreement**, not the model's self-assessment.

Table 3. Per-label agreement between Pass 1 and Pass 2 (presence).

Label	Agr.	Label	Agr.
PCO	99.7%	RSC	97.8%
SRO	99.3%	MRF	96.5%
RBE	98.5%	O	93.4%
TPRC	98.0%	NDC	91.3%

Table 4. Calibration: self-reported confidence vs. inter-pass disagreement.

Pass-1 confidence	Spans	Disagree with Pass 2
5 (maximal)	4,186	13.8%

Pass-1 confidence	Spans	Disagree with Pass 2
4	908	35.0%
≤3	17	~35%
all	5,111	17.6%

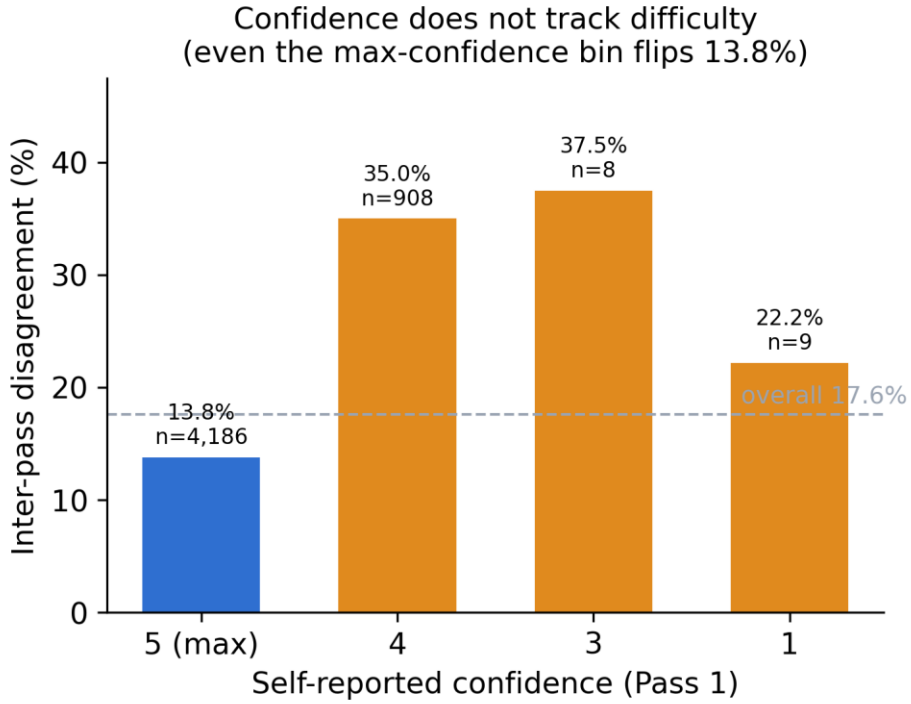


Figure 2. Self-reported confidence does not track difficulty. The maximal-confidence bin (blue), which holds 82% of predictions, still flips its label across independent passes 13.8% of the time; the dashed line is the corpus-wide disagreement rate.

Disagreement structure. Of the 901 inter-pass disagreements (17.6%), most are minor: 353 differ only on NDC and 354 are inclusion-boundary calls (O vs. a substantive label). Only **194** are *substantive label swaps*. These are not spread evenly: **~111 of 194 concern a single question** — whether a third-party royalty credit clause *also* carries a stacking-cap floor (TPRC vs. TPRC+RSC, 74+37 cases). That is, the passes agree there is an offset but disagree on whether an "in no event below X%" floor is present — exactly the RSC/TPRC boundary that is easy to miss. Remaining swaps involve SRO±TPRC, MRF±RSC, PCO±RSC, and RBE↔TPRC (Figure 3). Effect-direction (up/down/cap/floor) agrees across passes on **90.5%** of spans.

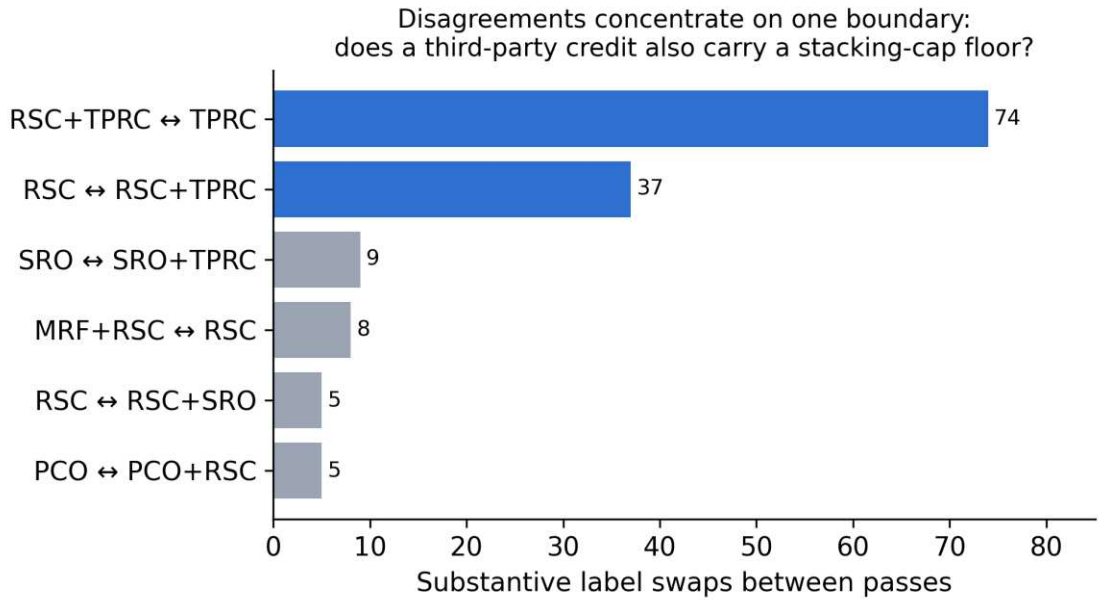


Figure 3. Substantive inter-pass label swaps, by type. Over half concern one boundary (blue): whether a third-party royalty credit clause also carries a stacking-cap floor (TPRC vs. TPRC+RSC).

A worked disagreement. An EX-10.3 exhibit from Aclaris Therapeutics illustrates the modal case with an in-clause numeric example: "… *where the Third-Party royalty obligation is less than [***], Assignee has no right [to] offset any portion of it against the royalty of this Agreement … [amounts above the threshold constitute] Excess Third-Party Royalties … eligible to be offset.*" Pass 1 read the "less than [***]" threshold as an effective-rate floor and labelled {TPRC, RSC}; Pass 2 saw only the third-party credit and labelled {TPRC}. Both readings are defensible: the clause plainly credits third-party royalties (TPRC), but whether its threshold amounts to a stacking-cap floor (RSC) is genuinely arguable from the excerpt — which is precisely why this boundary dominates the disagreements.

Adjudication. The Pass-3 adjudicator, shown both readings, resolved **all 214** hard spans with confidence 4–5 and flagged **zero** as irreducible — itself a reliability observation: presented with two conflicting rationales, the model does not hedge, it commits. Whether it commits *correctly* is the question the gold benchmark (§7) must answer.

Cross-model check. To test whether the resolution of these hard cases is specific to one model, we re-classified the 194 substantive-swap spans with an independent model from a different family (Claude Haiku 4.5, `claude-haiku-4-5-20251001`) under the identical prompt. Both it and DeepSeek (v4-flash) are lightweight serving-tier models decoded at temperature 0, so matching the capability tier isolates the *model-family* difference from raw capability. Claude sides with DeepSeek's Pass 1 on 82 spans (42.3%) and Pass 2 on 43 (22.2%), and matches *neither* pass on 69 (**35.6%**; Figure 4). Judged against the adjudicated label set specifically, it reproduces that set on only **71 spans (36.6%)**; Table 5 breaks this

down by which reading the adjudicator adopted. Because the adjudicated label is itself a DeepSeek silver label, this low agreement measures *how ambiguous the boundary is*, not which model is correct — a question only human gold can settle. With that caveat the point still sharpens: the adjudicator resolved every one of these spans with high confidence, yet an independent family reproduces its final labelling on fewer than four in ten.

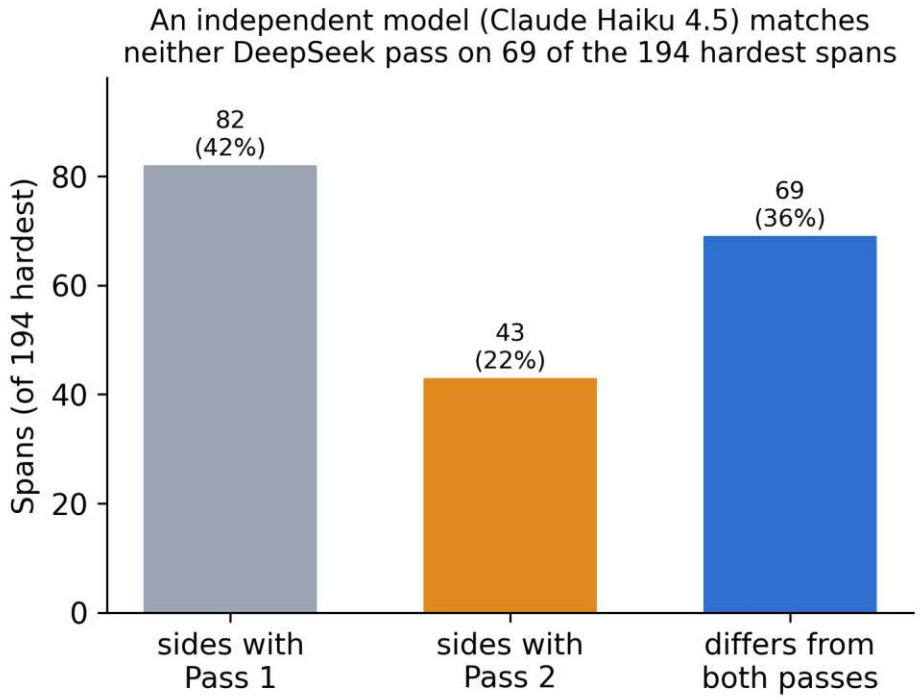


Figure 4. Cross-model check on the 194 substantive-swap spans. An independent model from a different family (Claude Haiku 4.5), under the identical prompt, sides with DeepSeek Pass 1 on 82 spans and Pass 2 on 43, and matches neither pass on 69 (35.6%). Judged against the adjudicated label set specifically (Table 5), it reproduces that set on 71 spans (36.6%).

Table 5. Whether Claude's label set matches the adjudicated final set, cross-tabulated with which reading the adjudicator ultimately adopted, on the 194 hardest spans. Claude reproduces the adjudicated set on 71 spans (36.6%): 57 where the adjudicator kept Pass 1's labels, 12 where it kept Pass 2's, and 2 where the adjudicator and Claude independently arrived at the same third labelling.

Adjudicator's final reading ↓	Claude = adjudicated	Claude ≠ adjudicated	total
kept Pass 1's labels	57	73	130
kept Pass 2's labels	12	36	48
a third labelling	2	14	16
total	**71 (36.6%)**	123 (63.4%)	194

Table 6. Three-way disagreements on the hardest spans: DeepSeek Pass 1, Pass 2, and an independent model (Claude Haiku 4.5) each assign a different label set to the same clause.

Clause (filer · exhibit)	Pass 1	Pass 2	Claude
Alien Technology · EX-10.27 (stacking provision requiring third-party evidence)	NDC, RSC, TPRC	NDC, RSC	NDC, TPRC
Sionna Therapeutics · EX-10.11 (credit against revenue share for third-party royalties)	RSC, TPRC	NDC, TPRC	NDC, RSC, TPRC
BiomX · EX-10.24 (stacking floor plus sublicensing revenue)	RSC, SRO	RSC	RSC, SRO, TPRC

Model–model agreement. The 36.6% above is measured on a *deliberately hard* subset; to gauge agreement on the corpus at large we take a separate stratified sample of 311 spans across all labels (overlapping the 194 hard spans in only 18 cases). There the two independent LLM annotators (DeepSeek and Claude) reach *substantial* agreement — macro Cohen's κ = **0.78** (Table 7; computed per label as binary presence/absence, then macro-averaged), from 0.54 for the gating condition NDC (again the hardest label) to 0.97 for PCO. The two figures are complementary rather than contradictory: κ = 0.78 reports broad-sample stability, while 36.6% reports failure precisely at the hardest boundaries. This is *model–model* agreement, reported in lieu of the human inter-annotator agreement a fully human-labelled version would give; we treat it as an agreement-based check and leave a human κ study as the natural next step.

Table 7. Model–model agreement (Cohen's κ) between two independent LLM annotators, DeepSeek vs. Claude, on a stratified sample (n=311).

TPRC	RSC	RBE	SRO	MRF	PCO	NDC	O	macro
0.86	0.77	0.86	0.75	0.80	0.97	0.54	0.71	**0.78**

Human validation. A human reviewed a stratified spot-check of 36 spans, sampled across adjudicator-resolved disagreements, per-label examples, and spans at different inter-pass agreement levels so that the check is not confined to easy cases; all 36 were accepted as consistent with the model's final labels. We position this as an early *sanity check* on the protocol — deliberately not a gold benchmark — and note that at 0.7% of the corpus it complements rather than replaces full human adjudication (Limitations). Of 5,111

released labels, 4,886 come from Pass 1, 189 from adjudication, and 36 are human-confirmed.

Reference baselines. To characterise the task we train supervised baselines to predict the eight functional/housekeeping labels from the clause text, splitting **by contract** (accession) so no exhibit appears in both train and test (4,058 train / 1,053 test spans across 619 contracts). Because the labels are silver, these measure *learnability* of the LLM's labelling from surface text, not accuracy against human ground truth. A TF-IDF (word + bigram) one-vs-rest logistic regression reaches **micro-F1 0.84 / macro-F1 0.80** (Table 8), far above a keyword-rule baseline (0.57 / 0.39) and the multi-label majority floor (0.00). Per-label F1 tracks difficulty: keyworded functions are near-solved (TPRC 0.90, MRF 0.89, RSC 0.85) while the gating condition **NDC is hardest at 0.64**, followed by the low-frequency labels (SRO 0.71, RBE 0.76). A fine-tuned legal encoder is the natural next baseline and is deferred to a GPU run.

Table 8. Reference baselines (contract-level split; predicting silver labels).

Model	micro-F1	macro-F1
B0 majority	0.00	0.00
B1 keyword rules	0.57	0.39
B2 TF-IDF + OvR logistic	**0.84**	**0.80**

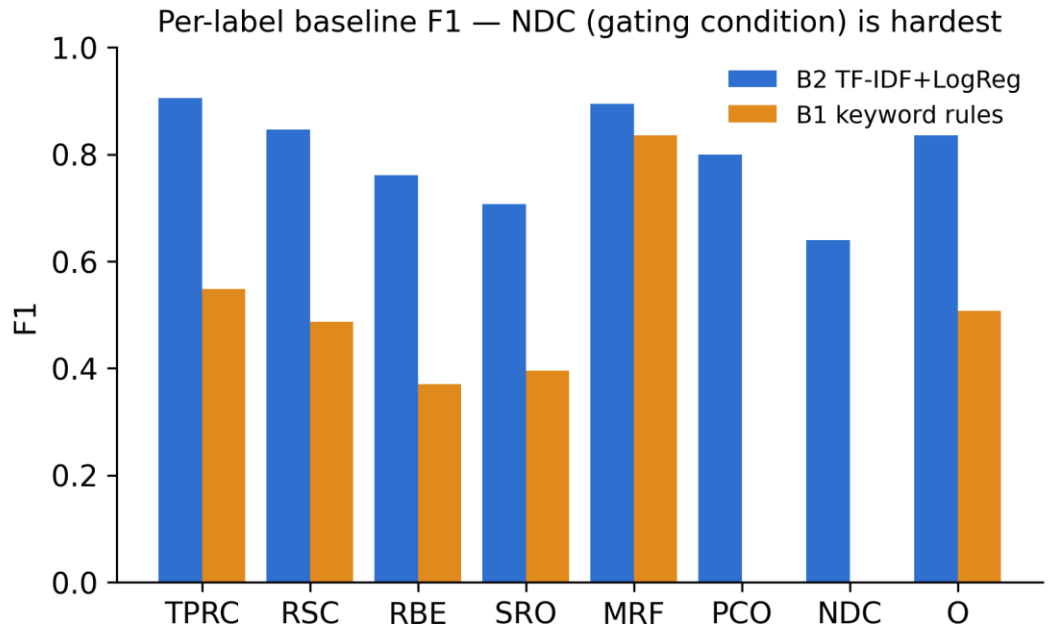


Figure 5. Per-label baseline F1. A TF-IDF logistic-regression classifier (blue) recovers the keyworded functions almost perfectly but struggles with the NDC gating condition; hand-written keyword rules (orange) lag throughout.

7. Discussion and limitations

What the paper delivers. A new silver-labelled corpus of royalty-risk clauses, a reproducible LLM-assisted annotation protocol, and a released set of artifacts — together with a quality-control analysis of that protocol showing where an LLM's confidence can and cannot be trusted. The labels are LLM-generated with a human spot-check (*silver*); the Limitations set out the concrete path to a human-gold version.

Over-confidence is the risk, self-consistency is the mitigation. Because the model floods the top confidence bin and almost never volunteers uncertainty, any single-pass LLM labelling would silently inherit its confident errors; double-pass disagreement recovers a small, interpretable hard set (concentrated on the TPRC/RSC boundary) at a fraction of full-annotation cost — a practical protocol for silver annotation when full human labelling is infeasible.

Threats to validity. (i) Labels are silver; consistent-but-wrong errors survive both passes and adjudication. (ii) Full-corpus labelling uses one model family; a cross-model check on the hardest spans (§6) shows an independent family agrees with the adjudicated label only 36.6% of the time, so multi-model annotation over the full corpus is an important extension. (iii) Human validation covers 0.7% of spans; we report model–model agreement between two independent LLM annotators (macro $\kappa = 0.78$) in its place, but a human inter-annotator study remains the standard we defer to. (iv) The corpus over-represents pharmaceutical and biotechnology filers and, being drawn from *material* exhibits, is an evidentiary sample rather than a representative cross-section. Pharma/biotech licences carry characteristic royalty structures (milestone-linked payments, tiered sublicensing) that differ from semiconductor, software, or bulk-chemical licences; this skew may also have shaped which functional boundaries the taxonomy foregrounds, so both the label distribution and the taxonomy's emphasis should be re-examined when the corpus is extended to other sectors.

The path to gold. A defensible benchmark requires a human-coded subset (target 300–500 spans) with a second coder or expert audit and inter-annotator agreement (κ), against which the LLM labels are scored — turning this silver corpus into both a scale benchmark and a genuine LLM-reliability evaluation.

8. Conclusion and follow-on work

We contribute a sizable EX-10 royalty-clause corpus, an LLM-assisted annotation protocol, a calibration and reliability analysis whose main lesson is that LLM confidence and LLM difficulty are decoupled for these clauses, and learnability baselines. **We make the resource openly available** — the corpus (redistribution-safe excerpts and an EDGAR reconstruction script), the per-pass and adjudicated labels, the baseline and analysis code, and a web labelling interface — available from the author during review and to be deposited in a public repository with a DOI upon acceptance, so that researchers and practitioners can inspect, correct, and extend it rather than re-collect it. Three follow-ons

are natural and independent: a human-gold benchmark and full reliability evaluation built on top of this release; a *computable-contracts* model mapping clauses to effective royalty rates; and an empirical atlas of how royalty-risk clauses have evolved across two decades of SEC filings.

Data and code availability

The corpus (source identifiers and redistribution-safe excerpts), the three-pass and adjudicated labels, and the baseline/analysis code are available from the author during review and will be deposited in a public repository with a DOI upon acceptance. Raw exhibit text is not redistributed; a reconstruction script retrieves it from public SEC EDGAR.

AI-use disclosure

DeepSeek was used as the primary LLM annotator described in §5, and Claude Haiku 4.5 (`claude-haiku-4-5-20251001`, temperature 0) for the cross-model check (§6); general drafting-support tools assisted manuscript preparation. The author reviewed and is accountable for all analysis, framing, and final wording.

Limitations

This work has four principal limitations, discussed in §7 and summarised here. (1) *The labels are silver*. They are produced by an LLM protocol with human validation on only 0.7% of spans; errors that are consistent across both passes and the adjudicator survive undetected, so the released labels should be used as a resource and a reliability probe, not as human ground truth. (2) *Primarily one model family*. The full-corpus labelling uses DeepSeek, so the self-consistency signal is *within-family*. We add a cross-model check on the 194 hardest spans (§6): an independent family (Claude Haiku 4.5) reproduces the adjudicated label set on only 36.6% (71 of 194 spans) and matches neither DeepSeek pass on 35.6% (69 of 194), empirically confirming that within-family agreement understates difficulty. A multi-model annotation of the whole corpus remains future work. (3) *Baselines predict silver labels*. The reported F1 measures learnability of the LLM's labelling from surface text, not accuracy against a gold standard; a fine-tuned encoder baseline is also deferred to a GPU run. (4) *Sampling*. The corpus is drawn from EX-10 *material* exhibits and over-represents pharmaceutical and biotechnology filers, so it is an evidentiary sample, not a representative cross-section of technology licensing. A human-coded subset with inter-annotator agreement is the concrete path to a gold benchmark (§7).

Ethics Statement

The data are drawn exclusively from public SEC EDGAR filings (material-contract exhibits) and contain no personal or sensitive data. We do not redistribute exhibit text; the release contains only source identifiers (accession/exhibit numbers), character offsets, and short excerpts, together with a script that reconstructs spans from the public EDGAR system, respecting quotation limits and confidential-treatment redactions (numeric terms already

masked as [***] in the source). The intended use is scholarly analysis of licensing-contract language and of LLM behaviour on it; the released labels are explicitly marked as model-generated ("silver") to prevent their being mistaken for authoritative legal classifications. The work is descriptive and does not provide legal advice.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D. M., Aletras, N. (2022). LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. ACL.

Dahl, M., Magesh, V., Suzgun, M., Ho, D. E. (2024). Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. Journal of Legal Analysis, 16(1), 64–93.

Gilardi, F., Alizadeh, M., Kubli, M. (2023). ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks. PNAS, 120(30).

Guha, N., et al. (2023). LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. NeurIPS Datasets & Benchmarks.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. ICML.

Hendrycks, D., Burns, C., Chen, A., Ball, S. (2021). CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. NeurIPS Datasets & Benchmarks.

Kadavath, S., et al. (2022). Language Models (Mostly) Know What They Know. arXiv:2207.05221.

Katz, D. M., Bommarito, M. J., Gao, S., Arredondo, P. (2024). GPT-4 Passes the Bar Exam. Philosophical Transactions of the Royal Society A, 382(2270).

Koreeda, Y., Manning, C. D. (2021). ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts. Findings of EMNLP.

Parr, R. L. (2007). Royalty Rates for Licensing Intellectual Property. Wiley.

Razgaitis, R. (2003). Valuation and Pricing of Technology-Based Intellectual Property. Wiley.

Surden, H. (2012). Computable Contracts. UC Davis Law Review, 46, 629.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., Manning, C. D. (2023). Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models. EMNLP.

Tuggener, D., von Däniken, P., Peetz, T., Cieliebak, M. (2020). LEDGAR: A Large-Scale Multi-label Corpus for Text Classification of Legal Provisions in Contracts. LREC.

Wang, S. H., et al. (2023). MAUD: An Expert-Annotated Legal NLP Dataset for Merger Agreement Understanding. EMNLP.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D. (2023). Self-Consistency Improves Chain of Thought Reasoning in Language Models. ICLR.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., Yang, D. (2024). Can Large Language Models Transform Computational Social Science? Computational Linguistics, 50(1).

Appendix A: Example clauses and sources

Each functional label and the worked disagreement (§6) is illustrated below by a fuller excerpt from a public SEC EX-10 exhibit; numeric terms appear redacted as [***] in the filing, and bracketed text is our clarification. Every clause is verifiable at the cited URL.

TPRC — third-party royalty credit — ATAI Life Sciences, EX-10.8. > "Third Party Royalties" shall mean any amounts paid by the Company, an Affiliate or a Sublicensee to a third party … for the right or license to use the technology … [creditable against royalties due under this Agreement]. Source: https://www.sec.gov/Archives/edgar/data/1840904/000119312521123009/d39052dex108.htm

RSC — stacking cap — T2 Biosystems, EX-10.13. > (b) Royalty Stacking. … In the event that COMPANY is required to make [***] payments to one or more third parties in order to practice the [***] PATENT RIGHTS for any PRODUCT … Source: https://www.sec.gov/Archives/edgar/data/1492674/000104746914005988/a2220654zex-10_13.htm

RBE — base exclusion — Vir Biotechnology, EX-10.28. > "Net Sales" means the gross amount [***], less Permitted Deductions. "Permitted Deductions" means the following amounts incurred by [***] any Licensed Product. Source: https://www.sec.gov/Archives/edgar/data/1706431/000119312519236592/d755217dex1028.htm

SRO — sublicensing offset — Islet Sciences, EX-10.1. > … under the Kissei Patents, after first deducting any amounts incurred by ISLT in settlement or defense of, or damages related to, [***], including but not limited to any royalties that may be payable to [***]. Source: https://www.sec.gov/Archives/edgar/data/1425919/000135448815001017/islt_ex101.htm

MRF — minimum floor — Smart for Life, EX-10.37. > "Guaranteed Minimum Royalty(ies)" (also referred to herein as "GMR(s)") shall be defined as non-returnable advances recoupable against Royalties earned in the same Contract Year … Source:

https://www.sec.gov/Archives/edgar/data/1851860/000121390022002202/fs12022a1ex10-37_smartforlif.htm

PCO — patent-cost offset — Liquidia Technologies, EX-10.18. > … [documented patent costs arising from Licensee's use of University Technology] or PATENT RIGHTS shall be offset against future royalties, fees, or other payments LICENSEE owes to UNIVERSITY hereunder. Source: https://www.sec.gov/Archives/edgar/data/1330436/000104746918004801/a2235907zex-10_18.htm

NDC — notice/documentation — Wyndham Hotels & Resorts, EX-10.4. > … (the "Report"), together with any supporting documentation reasonably requested by SpinCo with respect to the calculation by RemainCo of the Royalty amount or the Minimum Annual Royalty Shortfall, as applicable. Source: https://www.sec.gov/Archives/edgar/data/1722684/000110465918037816/a18-14429_1ex10d4.htm

Worked confusion (TPRC vs. TPRC+RSC) — Aclaris Therapeutics, EX-10.3. > … amounts paid to Third Parties shall be the "Excess Third-Party Royalties," and with respect to the Net Sales on which Excess Third-Party Royalties are due, Assignee will be entitled to deduct up to [***] of the Excess Third-Party Royalties … [but] where the Third-Party royalty obligation is less than [***], Assignee has no right [to] offset any portion of it. Source: https://www.sec.gov/Archives/edgar/data/1557746/000104746915006870/a2225650zex-10_3.htm