CovenantBench: Navigating Corporate Debt Contracts with Litigation-Verified Rewards

Anonymous authors
Paper under double-blind review

Abstract

We present CovenantBench, a benchmark and reinforcement-learning environment for legal reasoning over corporate debt instruments, anchored to verifiable outcomes of real covenant litigation. The benchmark has three layers: (i) agentic document-navigation tasks over actual indentures and credit agreements pulled from public filings — a 4,042-task evaluation suite with audited keys — scored by deterministic checkers; (ii) outcome prediction on real covenant disputes, with a hand-audited held-out evaluation set — 101 issue prompts across 89 disputes, every record reviewed line-by-line against the deciding opinion — each labeled by the recorded outcome and the kind of motion or appeal the court was deciding; and (iii) an adversarial debate protocol in which advocate policies argue both sides of contested clauses before a judge whose reward is anchored to the true outcome. We train a 9B-parameter model (QLoRA + GRPO, under $150 of compute) through supervised tracing, navigation RL, judge anchoring, and two rounds of debate self-play. Three findings follow. First, the $150 of training pays for itself across every layer: it doubles zero-shot agentic pass@1 on 825 held-out tasks of its training families ($0.283 \to 0.558$ composite, $0.234 \to 0.485$ full-marks, thrice re-evaluated), lifts the base on every family of the audited evaluation suite, and matches frontier-scale models on outcome prediction at roughly one-fifth their inference cost. Second, the trained judge's predictions respond to the clause, not just the facts: holding the facts fixed and outcome-scrubbed, swapping a protective clause for a permissive one moves P(creditor) in the doctrinally correct direction. Third, our arms do not answer from memory, while frontier arms partly do: closed-book probes place both our base and trained models at chance, while the comparison model 44× larger in total parameters recites actual rulings from memory. We release the corpus, environment, checkers, and adapters.

1 Introduction

Covenant disputes — fights over what a bond indenture or credit agreement permits — are decided by close reading: a defined term resolved through a chain of cross-references, a basket capacity computed from interlocking definitions, an exception construed narrowly or broadly. The 2020–2024 liability-management wave (uptier exchanges, dropdown financings) turned exactly these readings into billion-dollar outcomes, and courts differed on the interpretation of clauses whose language varies by a single qualifier (Serta vs. Mitel, Appendix I).

Existing legal AI benchmarks mostly hand a model a passage and ask it to label or classify it. What practitioners do is different: locate the operative language inside a 200-page contract, resolve its defined terms, and predict how a court would rule. We built CovenantBench to test that skill with rewards grounded outside any model's opinion: mechanical checks against the contract's actual structure, and the recorded outcomes of real courts.

Our contributions:

1. A three-layer benchmark with checkable answers (§3): an 89-case hand-audited pristine registry of held-out disputes (101 issue prompts, each record reviewed line-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

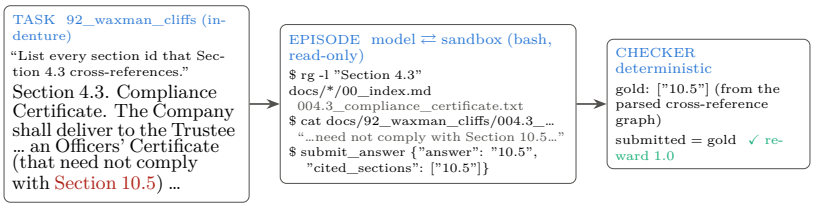


Figure 1: The task contract: real instrument, real episode, deterministic verdict. A cross-reference chase over an actual filed indenture (train-split episode, lightly truncated): single bash calls over the sectioned tree, submitted section ids, and a checker scoring against gold derived mechanically from the document's own structure — no model or human opinion enters the reward. The same contract covers every agentic task (§3).

by-line against its deciding opinion), drawn from a 424-dispute annotated corpus and held out of every training stage; 43 mounted issuer-filed contracts, each checked by a released clause-level quote audit against its deciding opinion; a 4,042-task navigation suite with audited keys (Appendix E); and 200 pre-ruling court filings with both sides' merits briefs[1] verified for 27 disputes.

2. A training environment and a 9B reference model trained for under $150 (§4), in four stages: imitating worked episodes, navigation RL, judge anchoring to real court outcomes, and debate self-play.
3. An evaluation methodology for outcome prediction (§6.1; Appendices I–J): memory-vs-reading probes, confidence-vs-accuracy checks, and clause counterfactuals — swapping the operative clause to see whether the prediction moves with the language.

2 Related work

LegalBench Guha et al. (2023), CaseHOLD Zheng et al. (2021), and contract-review datasets (CUAD Hendrycks et al. (2021), MAUD Wang et al. (2023)) test classification/extraction over supplied text; none require navigating a full instrument or ground truth in litigated outcomes. Reinforcement learning with verifiable rewards (RLVR) — training against rewards a script can check, rather than a human or model grader — underpins recent reasoning-model training; one widely used algorithm is GRPO Shao et al. (2024). We extend RLVR to document-grounded legal tasks where the check comes from document structure and court records (positioning matrix: Appendix B, Table 3). Irving et al. (2018) train agents to argue before a judge, and self-play variants report persuasion gains, but the judge is a model or human proxy with no external anchor, so reward hacking is undetectable within the protocol; our judge is trained against recorded court outcomes, giving the debate game a ground-truth-anchored reward. Agentic tool-use benchmarks (SWE-bench Jimenez et al. (2024) and successors) established pass@k over executable environments in software; CovenantBench transplants that discipline to legal documents, where the checker verifies section identity and computed capacities instead of unit tests.

3 The CovenantBench benchmark

3.1 Corpus

The corpus contains 424 issue-annotated records of U.S. covenant disputes (1982–2026, skewed 2014–2026). Each record states the facts, the operative provisions, the parties' arguments as summarized by the court, the procedural posture,[2] the governing law, and the outcome of each issue. Of these, 43 cases carry a mounted instrument parsed from

[1] A brief is a party's written argument filed with the court before it rules; it is on the merits when it argues the disputed question itself rather than scheduling or procedure.

[2] The kind of motion or appeal decided — summary judgment, plan confirmation, appeal — fixing what the court had to decide and on what record.

public SEC filings (EDGAR EX-4/EX-10 exhibits),[3] and 37 cases carry 200 court filings restricted, by construction, to documents written before each case was decided.[4] Released per-file manifests classify each filing's filer, side, and merits relevance; because collected filings sometimes address adjacent disputes on the same docket, brief-conditioned experiments draw on the manifest-verified subset only (Table 2).

Records were compiled by an LLM-assisted pipeline under human direction (see the LLM usage disclosure) from primary sources — opinions, dockets, EDGAR filings — against a released schema whose three label-leakage rules (no outcome-tainted language; faithful reconstruction of the parties' positions; equal argument bullets and length per side) appear in Appendix B, with why opinion sourcing dominates while brief coverage grows. Each issue's holding is recorded as a binary — did the creditor side or the company side win; nine mixed-outcome issues are excluded from training and kept as diagnostics (rule frozen in the released registry). An attorney-audited subsample remains future work.

3.2 The sandbox environment

Instruments are long — the median mounted document is roughly 62,000 tokens and the largest 245,000. A covenant's effect depends on the sections it cross-references rather than on the paragraph that states it. We therefore give the model a sandbox: an isolated, networkless container with the instrument mounted read-only as a tree of section files. The model writes shell commands — search, list, read, follow references — and reads their output before its next move, under a fixed per-episode tool-call cap. A median successful episode consumes about 2,700 tokens, roughly a twentieth of the whole-document alternative, because the model retrieves only the sections it needs; Appendix N lists the commands available in the container.

3.3 Task layers

L1 tests agentic navigation with a training suite and an audited evaluation suite. Both run in the sandbox over mounted real instruments, scored by deterministic checkers (normalized section-id F1, computed values, citation validity). The training suite (2,701 tasks) drives supervised tracing and navigation RL: cross-reference chase, defined-term resolution, basket capacity (the prompt stipulates formula and figures), and compliance. The evaluation suite (4,042 tasks over 41 instruments, 1,564 held out over 15) is built under an audited gate[5] in three families:

- defined-term dependencies (2,432): list every defined term a section depends on through one to three hops of definitions, with each term's depth and defining section;
- where-defined (587): name the section that defines a given term;
- explicit cross-references (1,023): list the sections a given section cites by number.

Depth levels nest (the 936 held-out dependency tasks cover 439 distinct anchor sections), and two frontier readers plus human adjudication check the depth-3 keys (Appendix E).

[3] Instrument sourcing is verified in two passes. Entity-level matching selects the exhibit: the record's issuer and trustee, and an execution date consistent with the instrument the record describes (every dated mount predates its dispute's ruling; a handful are undated printed forms). A released clause-level quote audit then checks verbatim instrument quotes from the deciding opinion against the mounted text — the test entity matching cannot make among look-alike documents by the same issuer. Of the 37 mounts with a retrievable opinion, 10 verify outright; 13 are inconclusive (issuer-matched and dated before the ruling, but the opinion quotes too little instrument text to confirm the series); 14 are proxies — a different series by the same issuer (e.g., Cash America) or a companion document of the same deal (e.g., Realogy) where the construed instrument is not publicly retrievable. Appendix B lists every verdict and why the proxies are kept.

[4] Filings come from claims-agent portals, RECAP, PACER purchases donated back to the public archive, state e-filing systems, the Supreme Court's docket, and archived merits-brief collections.

[5] Parser coverage 1.00 per instrument; every key stored with the verbatim source span it derives from; document-free constant-answer strategies score only 0.00–0.05 (Table 5).

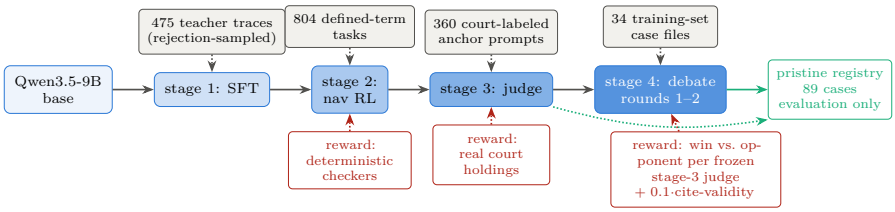


Figure 2: Four stages, one adapter chain; held-out cases touch nothing but evaluation. Each stage trains a LoRA adapter on the frozen 9B base, consuming the previous adapter, a stage dataset (top), and a verifiable reward (bottom); the debate reward is issued by the frozen stage-3 judge, and the 89-case pristine registry (right) appears in no training loop.

L2 tests holding prediction with arguments withheld, across 48 tasks. The model receives the facts and the mounted instrument, but no party arguments, and must find the operative language before predicting the outcome. Fourteen held-out tasks form the L2 evaluation set, thirteen of them evaluable (Chesapeake's construed instrument is unavailable in public filings; Appendix L); the 34 train-split case files seed the debate stage.

Judge anchoring trains on 360 prompts and evaluates on 101 pristine ones. The model receives facts, provisions, and both parties' arguments, and predicts the outcome (binary creditor/company; nine mixed-outcome issues are excluded and kept as diagnostics). A fourth released evaluation component, the 50-task instrument-verified dispute suite, composes these skills end-to-end over quote-audited mounts; it is exploratory at its size and reported in Appendix G.

3.4 Splits and the pristine registry

A fixed hash (sha256(case_id), frac 0.2) reserves one fifth of cases for evaluation and drives every layer, so a case is held out everywhere or nowhere — 89 cases, carrying the 825 held-out tasks of the training families and the 1,564 evaluation-suite tasks on the 15 held-out instruments, in no training loop of any stage. Two limits ride every held-out number: the split is pseudo-random, not temporal. Post-2020, the trained judge and Qwen3.5-397B score flat vs. pre-2020 while base, DeepSeek, and Kimi drop 5–11 points (n=20 per era), suggesting frontier familiarity with older disputes. And it hashes records rather than transactions: roughly seven multi-stage disputes straddle the split (e.g., Serta's appellate record trains while its summary-judgment record is held out). This is not an outcome-copy channel — in several straddling clusters the later stage reversed the earlier one — but held-out records are not transaction-independent of the two stages that consume records (judge anchoring, debate). Training and held-out sets share no mounted instrument, however, so the navigation stage is unaffected (Appendix B). A cluster-level split is scoped for future work.

4 Reference training pipeline

The reference system starts from Qwen3.5-9B and trains a LoRA adapter — a small set of add-on weights, the base model itself staying frozen — through four stages.[6]

1. Supervised tracing (Stage 1): a larger teacher (Qwen3.6-35B-A3B) attempts the same sandbox tasks; the 475 checker-accepted episodes are serialized as conversations with training loss on the command turns only, so the model learns to issue commands rather than to imitate the document text it was shown.
2. Navigation RL (Stage 2) on the training suite's defined-term tasks, rewarded by the deterministic checkers.

[6]Rollouts are served from an AWQ 4-bit copy of the base on one consumer GPU while an NF4 QLoRA trainer (rank 32, language-model matrices only) updates the adapter on a second, communicating over a file queue. Total compute ≈ $150 at market GPU rates.

Algorithm 1 Debate self-play (stage 4, one round). Round 1 (rebuttal) ; round 2 (blind) yields the reported stage-4 arm.

Require: train-split case files $\mathcal{C}$; advocate adapter θ; frozen judge ϕ; group size G; protocol $\in \{\text{rebuttal}, \text{blind}\}$
for step $t = 1, \ldots, T$; case c in batch do
  $s \leftarrow$ side alternating by parity of t; $\bar{s} \leftarrow$ the other side
  $a_{\text{opp}} \leftarrow \text{Episode}(\theta, c, \bar{s})$ ▷ one rollout, current policy, held fixed, no gradient
  for $i = 1, \ldots, G$ do
    $a_i \leftarrow \text{Episode}(\theta, c, s)$ ▷ sandbox, ≤10 tool calls; blind excludes a_{opp} from the prompt, rebuttal includes it
    $v_i \leftarrow \text{Judge}_\phi(c,\ a_i \,\|\, a_{\text{opp}})$; $r_i \leftarrow \mathbf{1}[v_i = s] + 0.1 \cdot \text{CiteValidity}(a_i, c)$ ▷ order seeded-random
  end for
  $\hat{A}_i \leftarrow (r_i - \bar{r})/(\sigma_r + \epsilon)$; $\hat{A}_i \leftarrow 0$ if $\sigma_r < 10^{-6}$ ▷ unanimous groups give no gradient
  update θ by GRPO on the G learner rollouts only ▷ ϕ, a_{opp} get no gradient
end for

Table 1: Results by evaluation layer. Navigation: mean checker score on the audited held-out evaluation suite, identical fixed draw and 10-tool-call budget for every arm (§5.2). L2: label accuracy on the 13 instrument-grounded holding tasks, arguments withheld. Judge: balanced accuracy on the 101 pristine prompts, majority vote. Cost: measured API (or marginal-electricity) cost of one full judge evaluation. Bold: best per column; the constant baseline answers without opening the document; API caveats in Appendix O.

	Navigation (mean score)					
Arm	where-defined	cross-references	dependency chains	L2 (13)	curated (101)	judge eval cost ($)
Untrained 9B (base)	0.657	0.841	0.087	0.615	0.910	0.17
Trained 9B (ours, $150)	0.715	0.944	0.117	0.769	0.897	0.17
Qwen3.5-397B-A17B	0.879	0.836	0.387	0.731	0.901	0.88
Kimi-K2.6	0.828	0.847	0.132	0.192	0.901	1.01
DeepSeek-V3.2	0.782	0.621	0.155	0.308	0.892	0.33
Constant baseline (no document)	0.00	0.05	0.00	—	—	—
Chance	—	—	—	0.500	0.500	—

3. Judge anchoring (Stage 3) on the 360 training prompts, rewarded by agreement with the recorded court outcome; the result is the judge.
4. Debate self-play (Stage 4): one advocate adapter argues both sides of training-set disputes before the now-frozen stage-3 judge, rewarded for winning rulings with valid citations; two rounds (rebuttal, then blind) yield the stage-4 model reported throughout (Algorithm 1).

All reinforcement-learning stages use GRPO.[7]

Two rounds of self-play lift the advocate well above its starting point on every tracked metric (win 0.202 → 0.29–0.32; cite validity 0.679 → 0.766; empty submissions 0.32 → 0.13) — with more valid citations under a frozen outcome-anchored judge, though Appendix L finds the reward bought reliability and citation validity, not content. Trajectories and the noise caveat (direction, not per-round monotonicity) are in Appendix H; the matched-compute ablation (Appendix K) reproduces the gains with plain anchoring.

5 Experiments

5.1 Curated-prompt outcome prediction

On curated prompts, neither scale nor training matters much: every arm from the untrained 9B to the 397B lands between 0.89 and 0.91 balanced accuracy with overlapping confidence

[7] Group Relative Policy Optimization Shao et al. (2024) scores a group of completions for the same prompt and uses the group mean as the baseline — no value network — reinforcing each completion by how much it beat its siblings; loss applies only to model-emitted tokens, never to sandbox output in context.

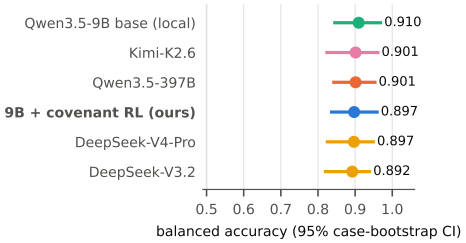


Figure 3: No clear difference across 44× scale on curated prompts. Balanced accuracy of the majority vote on the 101 pristine prompts (facts + provisions + curated arguments); whiskers: 95% case-bootstrap CIs. All arms overlap. The base's nominal high is noise: paired base−trained +0.012 $[-0.018, +0.049]$; majority verdicts differ on 3 of 101 prompts.

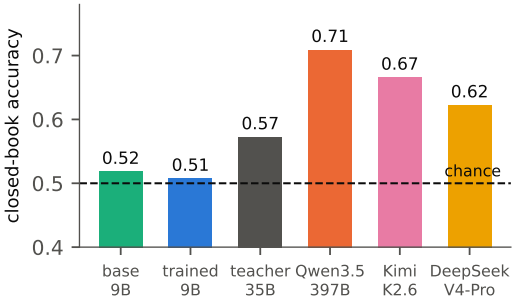


Figure 4: Our arms read; the frontier arms partly remember. Closed-book accuracy from the case caption and issue question alone, no documents in context (k=3 majority, 461 issues). Both 9B arms sit at the coin-flip line; the 35B teacher's 0.57 bounds what tracing could leak; all three frontier arms answer well above chance. Teacher bar: its pre-review probe (no v1.2 rerun).

intervals (Figure 3), and a model-free longer-argument-block rule is competitive with every model on the pre-symmetrization prompt construction. This suggests that the prompt format, not the model, carries most of the signal.

Each arm answers each of the 101 pristine prompts repeatedly, majority scored; balanced accuracy (averaged over creditor- and company-win issues; one-sided guessing scores 0.5) with 95% case-resampled CIs.[8] Every issue record underlying a held-out prompt is reviewed line-by-line against the primary record — the deciding opinion and docket. Our judge lands at 0.897 (ECE 0.057 at 16 votes; 0.074 when subsampled to the API arms' eight, level with the best API arms); the untrained base posts the nominal high (0.910) on coarser vote shares (11 recorded votes per prompt vs. 16; ECE 0.066); Qwen3.5-397B reaches 0.901, helped by a memorization advantage §6.1 measures directly (asked closed-book, the frontier arms score 0.623–0.709 where both 9B arms sit at chance).

The format is partly solvable without reading, and the record review shrank that channel. The longer curated block usually belongs to the winner: a rule reading nothing else scores 0.81 over the 459 cue-bearing issues corpus-wide, zero legal content. On the reviewed held-out records the rule drops to 0.61 on the pristine set — rewriting each side's arguments against the primary record largely evened out the lengths — and on the parties' actual pre-decision filings it fails outright (3 of 8 cleanly measurable cases), so the cue is curator emphasis inherited from the opinion. Absolute accuracy on this format remains weak evidence of reading (audit detail: Appendix C). We ship length-symmetrized prompts as the canonical evaluation set, removing the residual length channel while preserving each side's leading content and block order.[9]

We rebuilt each of the 21 pristine two-sided issues (17 cases) with equal-length verbatim excerpts from both sides' filings on the reviewed prompts (Table 4). Every 9B-scale arm drops 14–19 points — base 0.905 → 0.762, trained judge 0.905 → 0.714, Kimi-K2.6 0.857 → 0.667 — so opinion-derived arguments carry real pro-winner inflation for models that read them. The two arms that do not move are the memorization-heavy ones: Qwen3.5-397B scores 1.000 under either sourcing on these famous disputes, and DeepSeek-V4-Pro holds 0.810 — source-invariance here reads as name-keyed recall (the frontier arms answer 0.62–

[8]The public release ships 98 of the 101 prompts (three belong to the sealed private test set); aggregates differ by ≤0.005.

[9]The symmetrization and clause-counterfactual controls were run on the pre-review prompt set; they probe the prompt construction and clause sensitivity, not record content, and we read them relatively.

0.71 of issues closed-book, §6.1; the 9B arms sit at chance). On arguments written before any court ruled, no arm we trained beats 0.762, and training buys no sourcing robustness, consistent with its adding no judge-layer accuracy.

5.2 Agentic navigation: the audited evaluation suite

On the audited suite, training lifts the 9B base on every family, and the trained model is statistically level with one frontier arm and within 0.04 of a second; the largest arm leads on dependency chains, and most of that margin is explained by the tool budget rather than by answer quality. All five arms run the identical sandbox, prompts, and 10-call budget on the same fixed draw — 100 where-defined, 100 explicit cross-reference, and 300 dependency tasks (100 per depth), round-robin over all 15 held-out instruments — zero-shot, k=2, with document-clustered bootstrap CIs (tasks cluster within instruments; per-document scores in Appendix E).

Family by family (Table 1): the trained 9B leads every arm on explicit cross-references (0.944); where-defined puts all five arms between 0.66 and 0.88; dependency chains are hard for everyone — no arm exceeds a 5% exact-solution rate, the suite's clearest open headroom.

Dependency chains decompose into finishing and answer quality. An episode that exhausts its 10 calls without submitting scores zero, and that outcome dominates the family: completion rates run 16–26% for every arm but the 397B's 71%. Conditional on submitting, the ordering inverts: the trained 9B has the best answers on the board (0.630, term precision 0.84) and the 397B the weakest of the frontier arms (0.544) — it wins the mean by finishing three times as often, not by reading better (completions are each arm's own tasks; the ablation below is the like-for-like test). Paired per task, with bootstrap intervals that resample the 15 held-out instruments as clusters, the trained model beats its base (paired contrast in Appendix E, also on the stable-key subset: the 183 of the 300 drawn dependency tasks whose keys the corrected generator leaves unchanged), is inseparable from Kimi-K2.6 (-0.016 $[-0.048, +0.014]$), trails DeepSeek-V3.2 narrowly (-0.038 $[-0.069, -0.006]$, undetected on the stable-key subset) and the 397B decisively (-0.271 $[-0.302, -0.236]$, behind on all 15).

The depth-3 spread is mostly the budget. Re-running the same 50 depth-3 tasks with the cap raised from 10 to 30 calls leaves the 397B where it was (0.319 → 0.312; it already finished) and lifts every other arm by 4–8×: base 0.045 → 0.216, trained 0.030 → 0.255, Kimi 0.068 → 0.241, DeepSeek 0.184 → 0.305. At 30 calls the five arms land within 0.10 of one another and exact solutions stay rare: the 10-call protocol separates navigation economy, and three-hop resolution remains unsolved at any budget (Table 8).

On the 13 held-out L2 tasks (facts plus the mounted instrument, arguments withheld) the trained 9B submits on all 13 in a mean 6.7 calls and leads at 0.769 over the 397B's 0.731 and the base's 0.615; Kimi and DeepSeek mostly exhaust the budget unsubmitted (0.192, 0.308). These are famous disputes — asked closed-book, the 397B answers all 13 from memory (trained 0.846, base 0.692), and no arm's sandbox score exceeds its own closed-book prior — so the layer measures executing the retrieval within budget, not reading beyond memory.

The training-task families, evaluated on held-out cases, measure the same skills with the harness in the loop. The case-level hash of §3.4 holds all 825 out of every training stage. On them, training doubles zero-shot pass@1 (0.283 → 0.558 composite, 0.234 → 0.485 strict full-marks — criteria in Appendix A; three full-coverage evaluation runs within a point) while the frontier arms, seeing the sandbox for the first time, score 0.125–0.200 on its cross-reference tasks (Figure 7) — and one worked episode in context closes most of that gap for every arm (all three frontier stacks reach 0.775). Zero-shot gaps on an unfamiliar harness largely measure demonstration availability. The same control on the audited suite: given an identical worked example, all five arms lift on where-defined and cross-references (frontier arms by up to +0.28), and the trained model's cross-reference lead survives intact (0.947 vs. 0.903–0.930; Table 9).

Imitation does most of the training lift: stage 1 (teacher traces only) contributes roughly 72% of it, the two RL stages the remaining 7.6 points; stage-1's pass@8 (0.789) exceeds stage-4's (0.767), so RL adds first-sample reliability. Two checks bound the debate rounds: self-play did not erode navigation (the stage-4 model matches or exceeds the pre-debate stage-3 judge on every task type; Figure 9), and it did not buy argument content — the reward moved exactly the two dimensions it paid for, submission reliability ($0.52 \rightarrow 0.88$) and citation validity ($0.70 \rightarrow 0.94$), while coverage of the points the parties actually pressed stays flat from the untrained base through 1.6T-scale arms (Appendix L). Appendix D diagnoses failures at trace level: the base submits as often and works as hard as the trained arms and is simply wrong (accuracy given submission 0.55 vs. 0.90); training changed strategy, nearly quadrupling the reading-to-broad-search ratio. At matched judge accuracy the local judge evaluates for ≈\$0.17 per 101-prompt run, the whole pipeline ≈\$150 (Appendix F).

6 Integrity audits

This section audits memorization; counterfactuals and calibration are Appendices I and J.

6.1 Memorization: our arms at chance, frontier arms well above

Neither of our models can name a case's winner from its caption; all three frontier arms can — so our no-difference findings against them are conservative. The closed-book probe gives only the caption and issue question, no documents, facts, or arguments; accuracy above chance is prior knowledge (Figure 4). Our arms sit at chance — base 51.8%, judge 50.8% (no name-keyed recall); the 35B teacher's 57.3% bounds what tracing could leak. Same probe (k=3, 461 issues): Qwen3.5-397B 70.9%, Kimi 66.6%, DeepSeek 62.3%.

With outcome-scrubbed facts fixed and only the operative clause swapped (14 fact patterns × four clause families, k=16 per cell), both arms move P(creditor) correctly on 11 of 14 — clause sensitivity is largely base competence — and training removes wrong-direction failures (three → one; Appendix I).

7 Limitations

Two things we tried did not work: stage 4's adversarial structure is not detectably better than training the judge directly on recorded outcomes at matched compute (Appendix K) — we keep the debate protocol for its citation discipline, not a measured advantage — and training's navigation lift does not reach the deepest tasks: on depth-3 dependency chains at the 10-call budget the trained model scores at its base's level, and no arm at any scale or budget solves the family (§5.2).

The remaining limitations bound what the evaluation can establish. Small pristine outcome-eval: 101 prompts / 89 cases. Curation cue: curated arguments, largely reconstructed from the deciding opinion, may carry pro-winner framing; accuracy survives symmetrization and the brief-sourced control (§5.1). Label provenance: annotation is LLM-assisted under human review, the blind second annotation an independent LLM pipeline ($\kappa = 0.75$ on 25 issues), and no attorney audit has run; holdings are binarized. Key noise: training-suite references carry a ≈9.5% residual band; evaluation-suite key content is reader-endorsed (median 93% of terms), key recall carries a measured band (a released corrected generator grows 228 of 936 dependency keys), and the band is not arm-neutral (omissions charge arms that correctly include the term; re-keyed evaluation queued, Appendix E). Scope: U.S./NY-law instruments dominate, and the held-out navigation suite spans 15 litigated instruments, so navigation statistics are document-clustered (per-instrument scores in Appendix E); frontier agentic arms are k=2 samples under a spend cap, with closed-frontier GPT and Claude as key readers rather than sandbox arms; the L2 set is thirteen tasks; the split is pseudo-random (era check in §3.4) with ∼7 straddling disputes, scoped there.

8 Conclusion

CovenantBench tests what legal-NLP benchmarks have not: locating operative language in a real instrument, resolving its defined terms, reasoning about the dispute it produced — every reward grounded in document structure or court records. On that footing, $150 of training lifts the 9B base on every evaluation family, is not clearly separated from the frontier arms except on the deepest chains, and yields a calibrated, clause-sensitive, closed-book-at-chance judge; on the hardest tasks, tool-budget economy separates arms and none solves them.

LLM usage disclosure

This work was produced with substantial assistance from LLM-based tools (Claude Code) in implementation, experimentation, and writing. The human author originated the research idea, set the goals and constraints, made the design decisions at each stage, directed the evaluation agenda — including the contamination, memorization and prompt-signal analyses, several of which were prompted by the author's own scepticism of intermediate results — supplied credentials and data-access authorization, and reviewed the pipeline and results. LLM systems also appear inside the method itself, and are described where they are used: the annotation pipeline (§3), the blind second annotation (§3), the frontier readers that check the evaluation-suite keys and the model-drafted, human-adjudicated key audits (Appendix E), and the supervised traces distilled from a Qwen3.6-35B-A3B teacher (§4). The author assumes full responsibility for the content of this paper, and the released artifacts permit independent verification of every reported number.

References

Shao et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. 2024.

Irving, Christiano, Amodei. AI safety via debate. 2018.

Guha et al. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. NeurIPS D&B 2023.

Zheng et al. When does pretraining help? Assessing self-supervised learning for law. ICAIL 2021.

Hendrycks et al. CUAD: An expert-annotated NLP dataset for legal contract review. NeurIPS D&B 2021.

Wang et al. MAUD: An expert-annotated legal NLP dataset for merger agreement understanding. EMNLP 2023.

Jimenez et al. SWE-bench: Can language models resolve real-world GitHub issues? ICLR 2024.

In re Serta Simmons Bedding, L.L.C., 125 F.4th 555 (5th Cir. 2024). Excluded Lenders v. Serta.

Ocean Trails CLO VII v. MLN TopCo Ltd., 2024 NY Slip Op 06691 (1st Dep't 2024).

Gebru et al. Datasheets for datasets. CACM 2021.

Rein et al. GPQA: A graduate-level Google-proof Q&A benchmark. COLM 2024.

Yang et al. InterCode: Standardizing and benchmarking interactive coding with execution feedback. NeurIPS Datasets & Benchmarks 2023.

A Task example (L1 explicit cross-references, evaluation suite, held out)

> You are a credit analyst working in a shell over read-only legal documents. Navigate them with standard shell commands (rg, grep, cat, sed, ls), then submit your answer. Cite section ids exactly as they appear in the document.
> TASK: Which other sections does Section 2.8 cite by section number? Count only explicit numeric references to sections of this document (not Articles, not 'this Section', not clauses of the same section, not statutes). cited_sections = those ids.

Reference (checker xref1_v2; Petrohawk 2011 indenture): cited sections {2.7, 4.1, 11.6}, each stored with the verbatim span it derives from ("Section 11.6", "Section 2.7", "Section 4.1"); score is F1 over normalized section ids. The instrument is mounted read-only in the network-isolated container (Appendix N); the answer key lives outside the mount. Episodes allow at most 10 tool calls; the composite reward adds cite-validity (+0.2) and format (+0.1) bonuses and a −0.3 illegal-action penalty on top of task correctness. Throughout the paper, composite pass@1 counts a sample as passing when this composite reward is ≥ 0.99 (so a partially correct answer with full bonuses can pass); strict full-marks requires the maximum 1.3 — exact answer, valid citations, clean format. Both are reported in §5.2.

B Corpus provenance and additional figures

The released annotation protocol (which also governs all audits and any future human annotation) fixes three rules that matter for leakage:

- facts and issue questions must be answerable without the holding[10] (no outcome-tainted language);
- each side's arguments are faithful reconstructions of positions the parties actually took, drawn from the opinion's summary of the parties' positions, briefs quoted therein, or contemporaneous reporting;
- both sides receive the same number of argument bullets and approximately equal total length — a rule adopted after the length audit of §5.1 — so that argument volume cannot encode the outcome in any future curation.

The judge who wrote the deciding opinion of course knew the outcome, and because that opinion is the dominant curated source, arguments drawn from it may carry the winner cue; the Limitations section discusses the resulting risk for §5.1 and the brief-based control targeting it.

The opinion dominates because it is the only source retrievable at corpus scale without per-case acquisition work, not because briefing is unobtainable in principle: we estimate sustained acquisition could supply two-sided merits briefing for 150–250 of the 424 disputes. Three regimes apply.

- Genuinely unretrievable (a minority of the corpus): Cases predating federal e-filing (∼2001–2006) or New York's e-filing system exist only as paper in closed-case archives; Delaware Chancery filings sit behind a subscription service with no public portal; and some disputes (bench rulings on oral motions, omnibus-objection rulings) generated no merits briefing at all.
- Retrievable at a per-case cost (the modern majority): The corpus skews 2014–2026, squarely the e-filing era, so briefing exists on PACER, state e-filing systems, or claims-agent portals; the constraint is labor and fees — locating the decided motion on dockets that run to thousands of entries and purchasing its briefing.[11]

[10] The holding is what the court decided on the issue — here, which side won.

[11] Roughly a researcher-hour and $0–15 per case. Our per-file audit of one shallow pull found that 3 of 10 cases' collected filings addressed adjacent disputes, which is why bulk scraping does not substitute for docket work.

- Current coverage. 37 cases, 200 filings, 27 cases with on-merits briefs from both sides — two acquisition passes, not a ceiling; expanding this coverage is the roadmap's first corpus item.

This is also the honest frame for the cue controls of §5.1: they are the bridge that makes opinion-sourced prompts usable while brief coverage grows. We deliberately exclude final-hearing transcripts as an argument source even where available, because the judge's questioning at the last hearing before decision — and, in bench-ruling cases, the ruling itself — contaminates them with the outcome more severely than the opinion's after-the-fact summary does.

Table 2: Benchmark composition (top) and corpus coverage by attribute (bottom; counts over the 424 disputes, not a nested chain). Every layer's reward is checkable: L1 against document structure, L2 and anchoring against recorded court outcomes.

Layer	n	Reward source
L1 navigation, training suite	2,701	deterministic checkers
cross-reference chase	1,186	
defined-term resolution	1,115	
basket capacity	200	
compliance	200	
L1 navigation, evaluation suite	4,042	audited keys, checkers
defined-term dependencies (depth 1–3)	2,432	
where-defined	587	
explicit cross-references	1,023	
L2 holding prediction	48	court holding
Judge anchor prompts	461	court holding
training / pristine eval (hand-audited)	360 / 101	

Corpus stage	n
Annotated disputes	424
→ pristine held-out registry (hand-audited)	89
→ instruments mounted	43
quote-audited against deciding opinion	37
verified construed instrument	10
same-issuer / same-deal proxy	14
audit inconclusive	13
→ two-sided merits briefing	27
of which pristine	17

Of the 29 disputes represented by multiple records (different stages of the same transaction), roughly seven straddle the split; §3.4 gives the outcome-copy analysis and the per-stage scoping.

Beyond the id-level scoping of §3.4, we checked the text itself (script released): with 12-word shingles, the share of a held-out instrument's shingles found in any single training instrument peaks at 41.9% (Akanthos in BearingPoint, two convertible indentures of different issuers) and 40.3% (Cash America in SandRidge), then 27.9% and below, with seven of the 15 under 8%. That is the band unrelated training instruments share with one another through drafting templates (training-vs-training per-instrument maximum: median 10.7%, top 40.9%), an order of magnitude below same-issuer sibling exhibits (95–97% for the two the pool audit set aside). Nor does overlap buy the trained model anything: its per-instrument dependency-chain gain over the base (Table 7) is uncorrelated with containment (Spearman -0.15; the highest-overlap instrument, Akanthos, is one it loses ground on). The curation cue of §5.1 is a third, separate channel — a prompt-construction effect shared by every arm at evaluation, not a train/test overlap — and public filings remain a channel no split closes: any pretrained arm may have seen the raw text, an exposure shared across arms.

The quote audit resolves per mount as follows. Proxies mounting a different series by the same issuer (the construed instrument is not publicly retrievable): Cash America, Maverick, WaMu, Amylin, EFH, SandRidge, Chemtura, Cliffs. Companion documents of the

same deal: Realogy (credit agreement mounted; indenture construed), Momentive (intercreditor agreement), Trico (supplemental indenture), Quebecor (public indenture; private note agreements litigated), Sabine (indenture; gathering agreements litigated), Akanthos v. CompuCredit (affiliate securitization indenture mounted; the litigated convertible-notes indenture was located and re-mounted); the argument-level audit of Appendix L places Chesapeake v. ETC in the same class. Windstream, for example, passes the audit cleanly (18 quotes matched).

The 13 inconclusive mounts are issuer-matched and, where dated, executed before the ruling, but the opinion quotes too little instrument text to confirm the series. For six of them (Sears/ESL, Motors Liquidation, Chesapeake/ETC, Med Diversified, Tribune, WestPoint Stevens) the pre-mount document-type screen had also flagged a possible supplement or shelf form, and that flag was never closed out in writing; they are the weakest provenance class. Nine look-alikes have verified executed replacements in public filings, included in the corpus; the rest have none (private placements, pre-EDGAR paper, 144A-only issues, unexecuted forms) — a scarcity bounding any public-filings corpus. We keep the proxy mounts deliberately: a same-issuer, same-era instrument preserves the drafting conventions navigation tasks exercise, and dropping the cases would bias the corpus away from the private-credit disputes that dominate recent practice. Checkers derive from the mounted document itself, so structural rewards stay internally valid; case–instrument-dependent evaluations report the verified subset alongside the full set.

Table 3: Positioning: the first legal benchmark that is simultaneously agentic, outcome-grounded, checker-scored, and memorization-audited, with an RL training recipe against those rewards. Compared: LegalBench Guha et al. (2023), CUAD Hendrycks et al. (2021), MAUD Wang et al. (2023), CaseHOLD Zheng et al. (2021), SWE-bench Jimenez et al. (2024), InterCode Yang et al. (2023).

	Litigated instruments	Outcome labels	Agentic environment	Deterministic checkers	Memorization audit	RL recipe
LegalBench	✓	—	—	—	—	—
CUAD / MAUD	✓	—	—	—	—	—
CaseHOLD	—	✓	—	—	—	—
SWE-bench	—	—	✓	✓	✓	—
InterCode	—	—	✓	✓	—	✓
CovenantBench	✓	✓	✓	✓	✓	✓

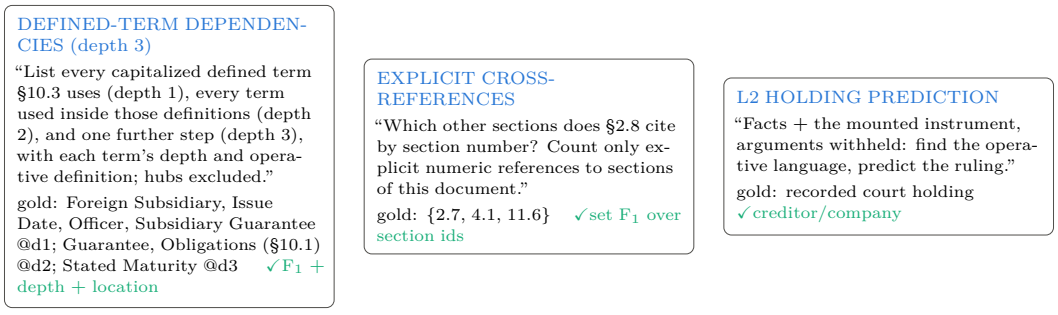


Figure 5: Three more task families keep the same contract. Real task instances from the evaluation suite (Petrohawk 2011 indenture, held out) and the L2 contract; every gold value derives from the document's parsed structure or the court record, never from a model.

C Curated-prompt cue: audit detail

Two mechanisms could produce the length asymmetry of §5.1 — curator emphasis inherited from the deciding opinion, which elaborates the winner's arguments as its own reasoning, or the winning side genuinely having more to say.

Two re-evaluations, run on the pre-review prompt set, remove the cue outright. (i) Length-symmetrized prompts: truncating both argument blocks to identical length (residual gap $\leq$19 characters) does not reduce accuracy — base 89/101 and judge 91/101 at k=8, against

base 9B 9 turns · reward 0.30 ✗

1 $ cat docs/00__index.md
  [stderr] No such file or directory
2 $ ls docs/ → 02__cash__america
4 $ cat docs/…/001.4__acts__of__holders.txt
  "…(subject to Section 6.1)…"
5–8 $ rg "Section" …001.4….txt ×4, same file
9 $ submit_answer {"answer": ["1.4","6.1","14.6"],
    "cited_sections": ["1.4"]}

includes the self-reference; cites a section it does not claim

trained 9B (stage 4) 7 turns · reward 1.30 ✓

1 $ ls docs/ → 02__cash__america
2 $ cat docs/02__cash__america/00__index.md
  → section index: 1.1 Definitions … 1.4 Acts of Holders …
5 $ cat docs/…/001.4__acts__of__holders.txt
6 $ grep -o "Section [0-9]*[0-9]*" …001.4…
  → Section 1.4 · Section 14.6 · Section 6.1
7 $ submit_answer {"answer": ["6.1","14.6"],
    "cited_sections": ["6.1","14.6"]}

excludes the self-reference; citations match the claim (cite-validity bonus)

Figure 6: Both arms make the same discovery; the discipline differs. Both arms on the same held-out task (02__cash__america___xref___0002: sections cross-referenced by §1.4; gold {6.1, 14.6}), captured at T=0.2 (excerpted; step numbers are turn indices). Both locate the operative text and surface the same three section mentions; the base model then violates the task contract — submitting the self-reference and citing a section it does not claim (reward 0.30) — while the trained model filters the self-reference and cites exactly what it submits (1.0 + 0.3 cite-validity). The gap of §5.2 is largely this: discipline about the task's rules, not search.

88/101 and 91/101 on the unmodified prompts at k=16 — and the cue-wrong subset improves (base 15/21 → 18/21). (ii) Arguments-only prompts, with facts and provisions removed: the judge still scores 89/101 (base 83/101), so the argument sections carry essentially all of the signal this format measures.

Table 4: Swapping curated arguments for the parties' real pre-decision briefs separates the arms that read from the arms that remember. Majority-vote accuracy on the 21 pristine two-sided issues, reviewed prompts (k=16 local, k=8 API). Every 9B-scale arm drops 14–19 points; the two memorization-heavy arms do not move. Paired case-bootstrap 95% CIs on the drop: base [0.00, 0.30], trained [0.05, 0.36], Kimi [0.04, 0.37]; DeepSeek-V4-Pro [−0.14, +0.15], Qwen3.5-397B exactly 0.

Arm	curated args	real briefs	Δ
Untrained 9B (base)	0.905	0.762	−0.143
Trained 9B (ours)	0.905	0.714	−0.190
Qwen3.5-397B-A17B	1.000	1.000	±0.000
Kimi-K2.6	0.857	0.667	−0.190
DeepSeek-V4-Pro	0.810	0.810	±0.000

D Trace anatomy: where navigation fails

The base model's failures are neither a compliance problem nor an effort problem: it submits as often and works as hard as the trained arms, and is simply wrong. Three findings from 780 instrumented episodes over a stratified held-out sample across all three local arms (Figure 8):

- The base model's deficit is not protocol compliance. It submits answers at the trained arms' rate (0.73 vs. 0.72 of episodes), but its submissions are wrong: accuracy given submission is 0.55, against 0.90 for stage 4.
- Failure is not under-effort. Base failures consume 8.7 calls on cross-reference tasks and still succeed in only 2 of 80 episodes, while the trained arms modulate effort with difficulty (successes ≈7.4 calls, failures ≈10.2).
- What training changed is strategy. Relative to base, the stage-4 model nearly quadruples its reading-to-broad-search ratio (cat 578 → 772 calls while rg falls 421 → 126), moving from scattershot string matching toward opening a section and following its references, which is the procedure cross-reference resolution actually requires (Appendix N describes each command).

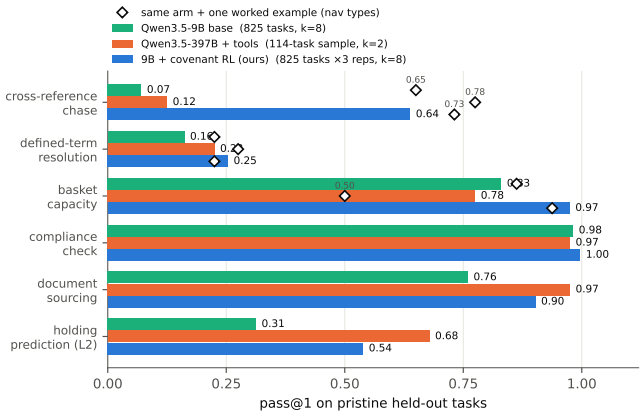


Figure 7: On the training suite, the zero-shot gap lives in navigation-heavy families and one demonstration closes most of it. Training-family performance on held-out tasks, by type. Bars: zero-shot. Diamonds: the same arm with one worked navigation example in context (§5.2). The cross-reference gap that looks like a 9× capability difference zero-shot mostly closes with a single demonstration — for the untrained base (0.07 → 0.65) as much as the frontier model (0.12 → 0.78) — while the same demonstration degrades the frontier model's basket computation (0.78 → 0.50) but neither 9B arm. Training's durable delta on this suite is demonstration-free competence.

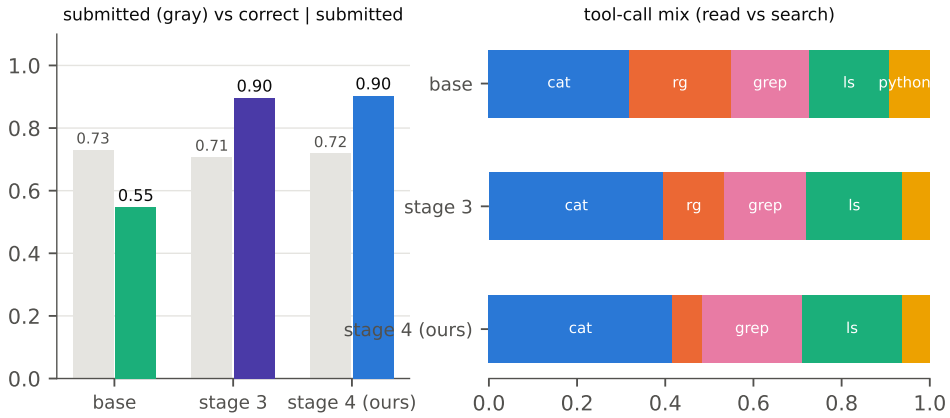


Figure 8: The base model submits confidently and wrongly; training converts search into targeted reading. Trace anatomy across arms. Left: episodes with a submission (gray) vs. accuracy given submission (color) — the base model submits confidently and wrongly. Right: tool-call mix — training shifts behavior from broad search (rg) toward targeted reading (cat, grep).

E The audited navigation evaluation suite: construction and key verification

The evaluation suite of §3 is built so that every key can be checked without re-deriving it: the parser is gated on coverage, every key carries the verbatim span it derives from, every family ships with a constant-answer baseline, and the hardest family's keys are read independently by two frontier models and adjudicated by hand where they disagree.

The parser turns each filed exhibit into the tree the sandbox mounts. It takes the raw HTML or text of the exhibit and emits one file per section, keeping the document's own numbering, plus the defined-term and cross-reference maps from which every key derives. An instrument passes the gate when at least 95% of the 200-character windows of its raw body

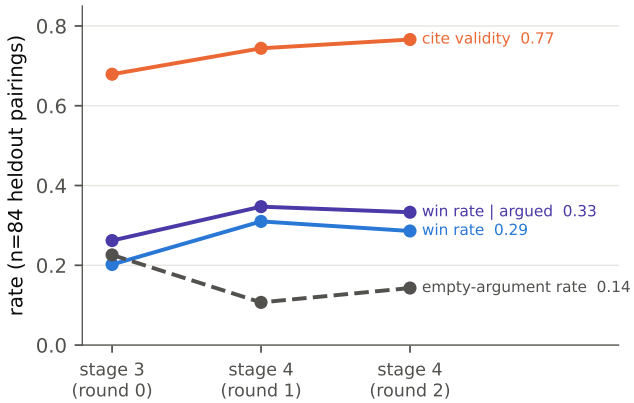


Figure 9: Self-play compounds across rounds (direction, not per-round monotonicity). Debate self-play rounds compound on the tracked metrics (held-out pairings, frozen outcome-anchored judge, blind protocol).

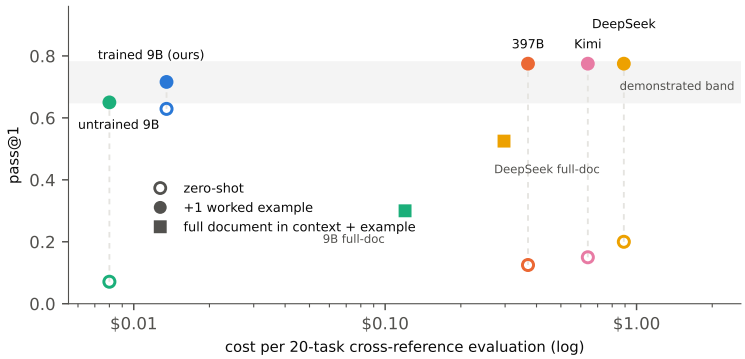


Figure 10: Cost against cross-reference pass@1 (20 tasks, k=2). Hollow markers are zero-shot; solid markers add one worked example. API costs are measured run tokens at posted rates; local cost is marginal electricity. Frontier arms are Qwen3.5-397B, Kimi-K2.6 and DeepSeek-V4-Pro. Squares: the long-context control — the full instrument in context plus the same worked example — which stays below the demonstrated band at 20–40× the tokens per attempt.

text (sampled every 600 characters, from the first real section to the exhibit cut) appear verbatim in the emitted sections. All 41 instruments used score 1.00. Sections keep the document's own numbering, table-of-contents duplicates are rejected rather than resolved by length, and front and back matter mount as separate files.

Every defined term enters a typed graph. Each is recorded with its kind (definitions-article entry, pointer to another section, inline coinage, delegation to another document, external statute, recitals), its operative span and quote, and the sections that use it. A hub is a term used in more than a quarter of a document's operative sections or feeding more than 15% of its definitions; hubs are listed in the prompt and excluded from every key. A dependency key is the breadth-first closure from the anchor section's non-hub terms through their definitions to depth 1, 2, or 3 (shortest path), capped at 25 terms; a depth is emitted only when it adds a term, so no depth-2 or depth-3 task restates a shallower one.

Verbosity does not pay under the dependency score. A submission lists terms, each with a depth and a defining section. Score $= 0.7\,F_1 + 0.15\,\text{depth} + 0.15\,\text{location}$: F_1 is symmetric over canonicalized terms, and every submitted term outside the key — a hub, an undefined phrase, an over-traversed term — counts against precision; depth accuracy is measured on

Table 5: Evaluation-suite census and constant-answer baselines. Held-out counts by depth over 41 instruments (26 train, 15 held out); every baseline is the mean checker score of a document-free strategy on the held-out tasks.

Family	train	held out	d1 / d2 / d3	constant baseline
Defined-term dependencies	1,496	936	439 / 323 / 174	prompt's hubs: 0.00
Where-defined	358	229	94 / 135 / –	definitions article: 0.00
Explicit cross-references	624	399	330 / 69 / –	every section id: 0.05
Total	2,478	1,564		41 instruments

matched terms and location accuracy over all locatable key terms, so a missing location counts as wrong. Verbosity does not pay. On the held-out depth-3 tasks the key itself scores 1.00; the key plus the prompt's hub terms 0.51; the key plus ten extra defined terms 0.51; the full closure to depth 12 (over-traversal) 0.51; half the key 0.48; the depth-1 terms alone 0.41; every defined term in the document 0.10.

Two frontier readers independently check the depth-3 keys — they are a check on the keys, never their source. The keys derive mechanically from the parsed document alone. For the 174 held-out depth-3 tasks, GPT-5.6-Sol (eight questions per call, and separately one per call) and Claude Opus 5 receive the full instrument text and the task prompts outside the sandbox and return term lists. Each reproduces the benchmark key exactly on about a quarter of the tasks (24.7%, 26.4%, 24.7%; Jaccard 0.70, 0.73, 0.72), and the two readers agree with each other more than either agrees with the key (Jaccard 0.88). By agreement pattern: key and both readers agree on 35 tasks; both readers agree against the key on 40; the readers disagree on 91; one reading is missing on 8. Where two independent readers agree against a mechanical key, the key is the suspect, and that set forms the adjudication queue. The 91 disagreements bound what they can impeach: even there, both readers endorse a median 93% of the key's terms, and the gaps concentrate in terms the readers propose beyond the key (jointly a median of two). The contested question is the key's recall, not its content. Symmetric F1 does not make omissions arm-neutral: an arm that correctly includes a term the key omitted is charged a false positive while an arm that misses it is not, so the 228-key band can move arms differentially; the re-keyed evaluation of the next revision is the test that settles it.

Contested keys go to human adjudication, which yields a defect taxonomy. Twenty contested tasks are adjudicated by the author against written span rules, released with the annotation protocol; the two load-bearing ones: the exclusion list is exactly the prompt's, and depth is the shortest path. Drafts are prepared by an LLM over the filed PDF and every line is adjudicated by the author, so the labels are model-drafted and human-adjudicated. Sixteen of the twenty keys were wrong, three were right against both readers, and one was mixed. Every error traces to a generator defect class (Table 6). The corrected generator ships with the release, developed under a four-gate regression harness (the five original hand-built keys reproduce, no term disappears from any graph, changed-key counts are reported, the adjudicated keys reproduce); its keys differ from the benchmark's on 228 of the 936 held-out dependency tasks (170 grow, 52 shrink, mean +2.5 terms), quantifying the recall band above. The benchmark evaluates on the keys as generated — the audited object — and re-keyed evaluation under the corrected generator is the next revision.

A post-hoc sensitivity check bounds what the corrected keys could change for the trained-versus-base contrast. Restricting the fixed 300-task dependency draw to the 183 tasks whose scoring is equivalent under the published references and the corrected generator (all 15 held-out instruments; depths 78/62/43), trained and base mean correctness are 0.156 and 0.098 — paired difference +0.058, 95% document-cluster bootstrap CI $[+0.022, +0.096]$, ahead on 13 of 15 instruments; +0.049 $[+0.021, +0.083]$ with depths weighted equally. On the 117 generator-changed tasks, scored on the original keys, the difference is -0.015 $[-0.035, +0.006]$: the aggregate lift comes from tasks whose keys the correction leaves alone, not from defective ones. This establishes robustness to excluding generator-changed tasks — not label validity, corrected-key performance on the full draw, or any depth-3 advantage (-0.018 $[-0.065, +0.049]$). The frontier arms on the same 183 tasks: trained 0.156

vs. DeepSeek-V3.2 0.152, Kimi-K2.6 0.149, and Qwen3.5-397B 0.425 — paired differences $+0.004$ $[-0.042, +0.054]$, $+0.007$ $[-0.039, +0.051]$, and -0.269 $[-0.313, -0.220]$, the same under equal-depth weighting and a three-comparison Bonferroni adjustment. No aggregate difference from DeepSeek or Kimi is detected. On the 43 retained depth-3 tasks DeepSeek (0.232) and Qwen (0.345) remain clearly ahead of the trained model (0.072). The 117 changed-key tasks, unrescored, show original-score deficits to DeepSeek (-0.104) and Kimi (-0.051), so excluding them is a sensitivity check, not a substitute for the full-draw table.

Table 6: Defect classes surfaced by adjudication, with corpus-wide incidence. The released corrected generator fixes them, changing 228 of the 936 held-out dependency keys (170 grow, 52 shrink) and removing no term from any graph; the benchmark evaluates on the keys as generated.

Class	Incidence
Defining verb not recognized ("shall be deemed to have occurred", "shall occur", "shall be determined")	18 terms, 21 instruments
Paired definition ("A" and "B" have the respective meanings)	4 instruments
Backward coinage ("… referred to as 'Restricted Payments' ")	18 sites, 13 instruments
Digit-initial term ("2012 Notes")	29 sites, 19 instruments
Page number interleaved in a multi-word term	101 occurrences, 30 instruments
Heading text absorbed into a term; hyphen and all-caps plural boundaries; unpaired quote; "hereinafter" coinage	measured per class before landing

Every arm runs one fixed, seeded draw. All reported navigation numbers come from a fixed, seeded, document-round-robin draw (500 held-out tasks; draw ids ship with the release) run identically for every arm: zero-shot, k=2, temperature 0.8, 10 tool calls, 3,000 characters of output per call. Bootstrap intervals resample instruments, not tasks. Table 7 gives each arm's dependency-family mean per held-out instrument.

Table 7: Dependency-chain mean score by held-out instrument (10-call budget, fixed draw). Paired per task with a 95% instrument-cluster bootstrap, the trained 9B beats its base by $+0.029$ $[+0.004, +0.058]$, ahead on 10 of 15 instruments ($+0.058$ $[+0.022, +0.096]$ on the 183 stable-key tasks, above); the 397B leads on all 15.

Instrument (case)	base	trained	397B	Kimi	DeepSeek
Cash America	0.08	0.15	0.47	0.30	0.16
Petrohawk	0.10	0.09	0.42	0.06	0.17
Akanthos/CompuCredit	0.14	0.08	0.46	0.09	0.22
Sears/ESL	0.11	0.10	0.39	0.09	0.12
WaMu	0.09	0.14	0.41	0.12	0.19
Chesapeake/ETC	0.17	0.28	0.40	0.20	0.22
UMB/Neiman Marcus	0.10	0.11	0.36	0.17	0.15
Fletcher/ION	0.07	0.19	0.49	0.14	0.22
AMR/American	0.10	0.14	0.26	0.14	0.00
Calpine	0.06	0.09	0.32	0.10	0.12
Premier/Biloxi	0.07	0.07	0.29	0.11	0.14
Maverick Tube	0.08	0.14	0.39	0.16	0.12
Charter	0.00	0.03	0.32	0.05	0.12
Young Broadcasting	0.06	0.05	0.36	0.08	0.17
Sanofi CVR	0.03	0.15	0.35	0.13	0.08

The demonstration control uses one identical worked example for every arm. Each arm re-runs the where-defined and cross-reference draws with one worked example prepended — the identical, family-matched episode for every arm, scripted on a training-split instrument (Windstream), verified to earn full marks in at most four calls, and shipped verbatim with the release. Budget and protocol are otherwise unchanged.

The two examples, outputs elided (full transcripts ship with the release):

EXAMPLE (cross-references; Windstream §11.05):
$ ls docs/ $ ls docs/01_windstream | grep -i 11.05

Table 8: Tool-budget ablation: the same 50 held-out depth-3 tasks at a 10- and a 30-call cap. Completion = episodes ending in a submission (an unsubmitted episode scores zero); every unsubmitted episode at either budget ran to its cap. The 397B spends whatever budget it is given (mean calls 9.5 → 28.7) and completes less often at 30; the other arms convert the added budget into finished answers. Score given submission covers completed episodes only — each arm's own subset, answer quality, not a ranking.

	10-call budget				30-call budget			
Arm	score	compl.	calls	score | subm.	score	compl.	calls	score | subm.
Untrained 9B (base)	0.045	11%	10.8	0.411	0.216	43%	28.9	0.502
Trained 9B (ours)	0.030	7%	10.9	0.434	0.255	44%	28.3	0.579
Qwen3.5-397B-A17B	0.319	72%	9.5	0.443	0.312	61%	28.7	0.512
Kimi-K2.6	0.068	15%	10.7	0.451	0.241	39%	30.2	0.618
DeepSeek-V3.2	0.184	33%	10.6	0.558	0.305	49%	30.2	0.622

Table 9: One worked example lifts every arm on the training-related families, and the trained model's cross-reference lead survives the control. Mean checker score, zero-shot → with the example (100 tasks per family, 10-call budget).

Arm	where-defined	cross-references
Untrained 9B (base)	0.657 → 0.821	0.841 → 0.925
Trained 9B (ours)	0.715 → 0.828	0.944 → 0.947
Qwen3.5-397B-A17B	0.879 → 0.912	0.836 → 0.930
Kimi-K2.6	0.828 → 0.860	0.847 → 0.911
DeepSeek-V3.2	0.782 → 0.860	0.621 → 0.903

$ cat docs/01_windstream/11.05_release_of_guarantor.txt
$ submit_answer {"answer": ["4.10","12.01","4.18"], "cited_sections": […]}

EXAMPLE (where-defined; "Registrar"):
$ grep -n "Registrar" docs/01_windstream/01.02_other_definitions.txt
$ grep -n -A1 '(" Registrar ")' docs/01_windstream/02.04_registrar_and_paying_agent.txt
$ submit_answer {"answer": "2.04", "cited_sections": ["2.04"]}

F Cost

At matched judge accuracy (overlapping CIs with DeepSeek-V4-Pro and Kimi-K2.6), our locally served judge costs ≈$0.17 per full 101-prompt evaluation — 5–6× cheaper than Kimi-K2.6 and the top-scoring 397B at Table 1's measured costs, with DeepSeek-V3.2 nearer at 2× — and the entire training pipeline cost ≈$150. Figure 10 plots the same cost basis against cross-reference pass@1 on the navigation suite, with the long-context control. Admittedly, the bases differ: the local figure is measured wall power, marginal electricity only with hardware amortization excluded, while API retail prices embed provider capital and margin, so read this as a marginal-cost comparison. On that basis the $150 of training is recovered after roughly 180–210 full 101-prompt evaluations against Kimi-K2.6 or the 397B (about 940 against DeepSeek-V3.2).

G Exploratory: the verified dispute suite (50 tasks)

Reported here rather than in the body: with 50 tasks over ten cases, the paired intervals of Table 10 cannot distinguish parity from a ten-point gap. On a 50-task suite built entirely over quote-audited mounts, the debate-trained 9B shows no clear difference from the frontier arms, the untrained base separates cleanly, and no arm at any scale reliably reconstructs the citation set a litigating side pressed. The suite turns the corpus audit of §3 into an evaluation instrument: every example lives on a held-out case whose mount passes the quote audit, with machine-checkable gold.

Seven scoring families, each with verbatim provenance: quote localization (19 — locate the court-quoted passage from the document), advocacy citations (8 — recover the section set a named side pressed, from the opinion's own attributions), the staged-episode operative-set, contrast, and amount tasks (15), clause discrimination (5 — pick the governing section among four real same-document candidates), and numeric thresholds (3, doc-verified). An integrity gate checks every gold id against its mounted corpus before release; evaluation uses the L1/L2 sandbox and tool budget, k=4 with uniform forced final submission — 200 episodes per arm.

	quote	advoc.	oper.	contr.	disc.	num.	amt.	macro	micro
DeepSeek-V4-Pro	0.961	0.393	0.744	0.679	0.950	1.000	1.000	0.818	0.802
Kimi-K2.6	0.934	0.457	0.678	0.571	1.000	1.000	1.000	0.806	0.783
stage-4 (9B)	0.816	0.422	0.734	0.750	0.900	1.000	1.000	0.803	0.755
Qwen3.5-397B	0.934	0.450	0.713	0.857	1.000	0.917	0.500	0.767	0.812
base (9B)	0.447	0.312	0.598	0.607	0.950	0.917	0.500	0.619	0.549

Table 10: No clear difference from the frontier arms; clean separation from the base; advocacy open for everyone. Verified dispute suite, mean over k=4; task counts 19/8/7/7/5/3/1. Macro = equal-weight mean of the seven family means; micro = mean over 200 episodes. Case-clustered 95% CIs on the macro (resamples average the families present): base 0.619 [0.562, 0.685], stage-4 0.803 [0.665, 0.860]; paired stage-4−base +0.18 [+0.05, +0.25]. Paired gaps to DeepSeek/Kimi/Qwen (−0.015/−0.003/+0.036; 95% CIs ±0.07–0.11 span zero): no detectable difference at ten cases, not equivalence. Under micro, Qwen3.5-397B leads stage-4 by 0.057 [−0.008, +0.135]; the one-task amount family carries 1/7 of macro weight.

Three readings stand out. Quote localization is the cleanest discriminator (frontier 0.93–0.96 vs. base 0.447), resistant to outcome-keyed recall by construction — knowing the holding does not reveal which section of this exhibit contains the quote — though the exhibits are public, so pretraining exposure to their text (shared by every arm) is untested. The 9B's standing is earned agentically: it leads all but the 397B on contrast and submits unforced more reliably than two of the three frontier arms (forced 6.5% vs. 5.0–23.5%). Advocacy citation: no arm exceeds 0.46 at recovering a side's actually-argued provisions. Statistics stay case-clustered.

H Debate dynamics: per-round trajectories

We score each advocate on held-out disputes against a fixed opponent (arguments generated once by the stage-3 model), with the frozen stage-3 judge deciding (n=84, blind protocol; Figure 9). From the stage-3 initialization through debate rounds 1 and 2:

- win rate 0.202 → 0.286 → 0.321 (a fresh k=3 re-evaluation under provenance-guarded logging, this revision: 0.202 → 0.310 → 0.286);
- win rate given that an argument was produced 0.298 → 0.333 → 0.370;
- citation validity, the share of citations that point to real sections, 0.628 → 0.694 → 0.728 (re-evaluation: 0.679 → 0.744 → 0.766);
- empty-submission rate 0.321 → 0.143 → 0.131.

Both trained rounds sit well above the stage-3 baseline (0.29–0.32 across rounds and evaluations), but the round-1 → round-2 ordering inverts between evaluations because per-round increments are inside n=84 evaluation noise (±0.05–0.10), so we claim direction, not per-round monotonicity.

I Clause counterfactuals: does the model read the provision?

Both arms move their prediction in the doctrinally correct direction when only the clause changes; what training adds is not direction sensitivity but the removal of wrong-direction failures (three → one) and sharper predictions where the facts are genuinely contested. The

design follows Serta (5th Cir.) and Mitel (1st Dep't), which reached opposite results on materially different buyback clauses — "open market purchases" (narrow) vs. "purchase by way of assignment" (broad): we hold outcome-scrubbed facts fixed and swap only the operative clause between a protective and a permissive variant, over 14 real fact patterns × four clause families,[12] k=16 votes per cell, both arms (Figure 11).

- Both real-world diagonal cells are predicted correctly by the trained judge (Serta facts + narrow clause: $P(\text{creditor}) = 1.00$; Mitel facts + broad clause: 0.00).
- On direction the result is honest rather than flattering: both arms move $P(\text{creditor})$ the doctrinally correct way in 11 of 14 patterns, so clause-direction sensitivity is largely a base-model competence, consistent with §5.1; on roughly half the patterns the base shows larger correct-direction shifts than the trained judge.
- What training changes is the failure profile and the contested middle: wrong-direction cells drop from three (base) to one (the trained judge's sole reversal, TriMark buyback, is a blemish the base does not share), mean clause effect rises from +0.33 to +0.36, and the sharpening concentrates where facts are genuinely contested (Mitel +0.39 → +0.60; Hertz make-whole +0.11 → +0.38; Windstream dropdown −0.05 → +0.12).

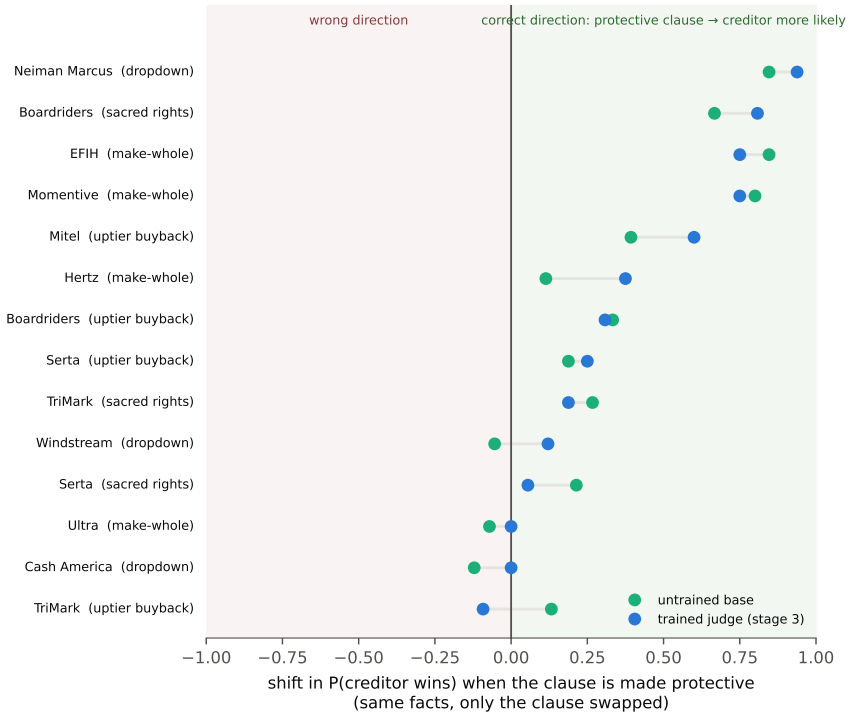


Figure 11: Predictions move with the provision, in the doctrinally correct direction. Clause counterfactuals: $\Delta P(\text{creditor})$ when swapping a permissive for a protective clause under fixed, outcome-scrubbed facts (right of zero = doctrinally correct direction). Training removes wrong-direction failures (3 → 1) and sharpens contested patterns rather than creating direction sensitivity from scratch.

[12]The four families are the recurring flashpoints of recent restructurings: whether a majority of lenders may buy back debt on terms that leave the minority behind (uptier buyback); protections that every lender must consent to waive (sacred rights); whether early repayment triggers a premium (make-whole on acceleration); and moving assets out of creditors' reach (dropdown/asset transfer).

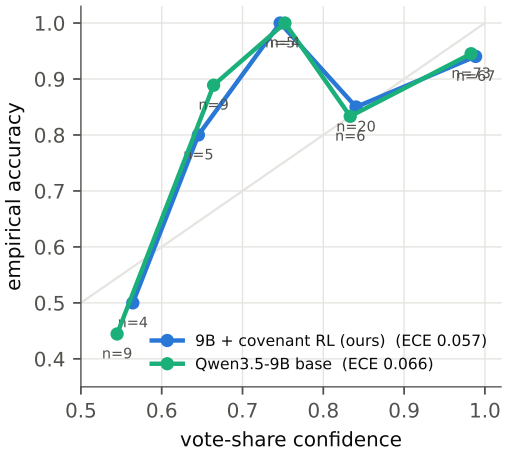


Figure 12: Vote share works as a probability. Reliability of judge vote-share on the pristine prompts (trained ECE 0.057 at 16 votes, 0.074 at the API arms' 8; base 0.066 at 11).

J Calibration: vote share is a usable probability

The judge answers by majority over k votes, so the fraction of votes for the winning side doubles as a confidence estimate: 13 of 16 votes for the creditor reads as 0.81. Calibration asks whether that number means what it says — across all the prompts where the judge reports roughly 80% confidence, does it get roughly 80% of them right? We bucket the pristine prompts by reported confidence and compare each bucket's mean confidence against its observed accuracy (Figure 12).

Aggregated over buckets — each bucket's observed mean confidence against its observed accuracy — the trained judge's expected calibration error is 0.057 and the untrained base's is 0.066 on the identical harness; the API arms run 0.069–0.113. The trained judge posts the lowest ECE on the table — as a point estimate at unmatched vote counts: the paired case-bootstrap gap over the base is -0.009 $[-0.052, +0.042]$ (Brier 0.085 vs. 0.088, Δ $[-0.024, +0.018]$), inside noise at n=101, and it does not survive matching: subsampling the trained judge to the base's 11 votes per prompt gives ECE 0.068 [0.044, 0.094] over 500 subsamples, so the 0.057 rests on finer vote-share resolution, not better-placed confidence. Subsampling to the API arms' eight votes gives 0.074 [0.045, 0.108], level with DeepSeek-V4-Pro (0.074) and Kimi-K2.6 (0.069): the "lowest ECE" is a vote-count artifact against every arm on the table, not only the base. It is also a curated-format result: on the brief-sourced prompts of Table 4 (n=21), calibration collapses for every reading arm (ECE 0.23–0.28, Brier 0.19–0.25 across both 9B arms and the two reading frontier arms) while the 397B stays near zero error by reciting. Vote share is a usable probability on curated prompts; no reading arm's confidence yet survives real filings.

What differs is how confidence is spent. The base crowds 73 of 101 prompts into its top bucket — its shares are quantized over 11 recorded votes per prompt (0.91 parseable) against the judge's 16 (0.98), so they run extreme — and is overconfident where it hedges (4 of 9 right at a reported ≈0.55). The trained judge spreads its confidence and tracks it: 17 of 20 right at a reported ≈0.85 and 63 of 67 in the top bucket. The low-confidence buckets hold few prompts, so we read the pattern, not the per-bucket margins.

K Ablation: adversarial self-play vs. matched-compute training

The forty extra steps pay on the argument side — debate and plain anchoring both lift the challenger win rate from the stage-3 initialization's 0.202 to 0.286, and debate lifts citation validity 0.679 → 0.766 — while every +40 arm lands at 90±2% judge accuracy. What is not detectable at matched compute is the vehicle: from the same initialization, curated pairs, fresh adversarial pairs, and advocacy-side training land together (Table 11).

The debate rounds' compounding curves (Figure 9) invite a causal reading, that adversarial structure improves the model, and this ablation tests it. From the same stage-3 initialization we trained a control for exactly the debate rounds' budget — 40 GRPO steps, same pool, hyperparameters, and optimizer schedule (moments reset at the round boundary) — but training the judge directly against recorded outcomes throughout; the comparison is compute-conservative in the ablation's favor, since debate additionally spent opponent and judge inference.

- Plain anchoring (anchor40) matches debate on every axis: judging 91.0% vs. 90.1% balanced, calibration ECE 0.073 vs. 0.078, navigation retention on identical held-out tasks 0.658 vs. 0.644 (neither arm forgets), and — most strikingly — argumentation, a skill the control never once practiced (challenger win rate 0.286 vs. 0.286, cite validity 0.719 vs. 0.766, regenerated cells). Learning to predict outcomes appears to implicitly teach producing winning arguments.
- Judge alternation lands on the same plateau. The design's last open cell trains the judge for the same 40 steps and labels on fresh adversarial argument pairs generated per step by the frozen debate-trained advocates, refreshing the input distribution that anchor40 had saturated (anchor40's training reward sat at ceiling, ≈0.96; the alternation arm's stayed hard throughout, ≈0.29). Result: judging 89.7%, ECE 0.064, navigation 0.631, and challenger win rate 0.202 — exactly the stage-3 initialization's level, so rating adversarial arguments, unlike rating curated ones, produced no argumentation transfer (a directionally interesting asymmetry within noise at n=84).
- All three +40 arms trade a small calibration loss against the stage-3 initialization (ECE 0.044 → 0.06–0.08 under the binning of §J; pairwise ΔECE CIs straddle zero) — a shared cost of the extra forty steps, not a debate-specific one.

"Null" here means bounded and undetectable at n=101, not proven zero: the design bounds any input-engineering effect at roughly five points of judging accuracy.[13] The equivalence also holds only at this reward specification: the judge rules on the argument texts without opening the document, so no arm's reward prices whether a cited section is the operative one — debate's directional edge on citation validity, the one dimension only it exercises in the sandbox, is exactly where a content-priced reward would look next. At this scale and label budget the curated-prompt judging task is saturated, and the residual error appears label-limited or genuinely contested rather than input-limited. Whether any of these mechanisms pays at larger scale or label count remains open.

L Argument-level audit: what the self-play reward bought

The debate stage's reward is $\mathbf{1}[\text{win}] + 0.1 \cdot$citation-validity against the frozen judge (§4). On training cases that judge was itself trained on the recorded outcome, so a verdict may be memory rather than argument; under GRPO's group-relative baseline (Algorithm 1, line 8) a verdict that is the same for every rollout in a group — as a verdict decided from the memorized outcome is — contributes no win-gradient, only the citation term does, so memorized outcomes cannot reward an argument the judge would not also reward on content. This audit asks what that reward purchased at the level of argument content, and the answer is reliability rather than content: the two rewarded dimensions moved sharply, the two unrewarded ones did not, and no arm at any scale produces the doctrine-level points the litigants led with. Two of the 13 audited issues sit on look-alike mounts (§3); on the instrument-verified remainder the trained arms' point recall is slightly higher (round 1 +0.06, round 2 +0.04) with the base flat, so the conclusions here are, if anything,

[13] Because all four arms share the same 101 prompts, we report paired statistics rather than independent proportions: no pairwise judging contrast is distinguishable (McNemar's paired test, exact $p \approx 1.0$; at most five prompts per pair on which the two arms disagree), case-bootstrap 95% CIs on pairwise balanced-accuracy differences span at most ±0.075 (debate vs. anchoring: $[-0.051, +0.037]$), and an equivalence test (TOST) certifies the difference lies within a ±0.05 margin for the three contrasts among the +40 arms (not for those against the stage-3 initialization, whose CIs graze the margin).

	stage 3 (init)	+40 debate	+40 anchoring	+40 alternation
Judge accuracy (101, k=16, balanced)	0.897	0.901	0.910	0.897
ECE	0.044	0.078	0.073	0.064
Challenger win rate (n=84)	0.202	0.286	0.286	0.202
Win rate | argued	0.262	0.333	0.320	0.266
Citation validity	0.679	0.766	0.719	0.688
Navigation pass@1 (130 tasks)	0.632	0.644	0.658	0.631

Table 11: The forty extra steps pay; which vehicle delivers them is not detectable at equal compute. Matched-compute ablation on the pre-review prompt set, all four arms from the same stage-3 initialization: debate self-play, plain anchoring on curated pairs (anchor40), and judge alternation on fresh adversarial pairs. No pairwise contrast on any axis is statistically distinguishable (McNemar $p \approx 1$; paired case-bootstrap CIs within ± 0.075), and TOST certifies equivalence at ± 0.05 among the +40 arms: input engineering did not detectably move the plateau. Provenance: judge-accuracy and ECE cells regenerate from the released per-prompt vote files and script.

conservative. Each arm drafts both sides' arguments over the mounted instrument for the held-out L2 disputes (13 of the 14 held-out issues — Chesapeake is excluded, its construed instrument being unavailable in public filings; 2 sides × 2 samples = 52 arguments per arm). Arguments are graded against the argument points each party actually pressed in the litigation — distilled from the record into 152 per-side reference points, withheld from every prompt — with per-point credit 1 / 0.5 / 0 and flags for empty and wrong-side arguments. The grader is gpt-oss-120b, an open-weight model from a family that appears in no arm; a second, independent in-context grading pass agrees at 78.8% exact / 98.8% within half a point, with near-identical wrong-side flags.

Arm	condition	clean args	point recall	recall | clean
Untrained 9B	zero-shot	24/52	0.319	0.693
stage 3 (judge-anchored)	zero-shot	34/52	0.401	0.616
stage 4, round 1 (self-play)	zero-shot	39/52	0.500	0.673
stage 4, round 2 (self-play)	zero-shot	45/52	0.612	0.705
Qwen3.5-397B	demo+forced submit	40/52	0.329	0.431
Kimi-K2.6	demo+forced submit	43/52	0.576	0.694
DeepSeek-V4-Pro	zero-shot	34/52	0.516	0.801
DeepSeek-V4-Pro	demo+forced submit	41/52	0.595	0.754

Table 12: Coverage of the litigants' real argument points. "Clean" arguments are non-empty and argue the assigned side; empty and wrong-side arguments score zero. Frontier arms are shown at their best-faith harness: zero-shot, 397B and Kimi exhaust the 10-call budget without submitting (9/52 and 10/52 clean), so a worked example plus a forced final submission is added; DeepSeek is the only frontier arm that submits reliably unaided.

The reward moved exactly what it paid for. From base to stage 4, the two rewarded dimensions move sharply: submission reliability 0.52 → 0.88 and citation validity 0.70 → 0.94. The two unrewarded dimensions do not: the fraction of citations landing on the provisions actually in dispute stays at ≈0.27–0.29 throughout, and point recall conditional on a clean argument is flat — 0.62–0.80 for every arm in the study, local or frontier, except the 397B's forced submissions (0.43). Training's measured contribution to argumentation is reliability, not content. No arm at any scale produces the doctrine-level points the litigants led with (Sharon Steel uniformity, demand futility, §502(b)(2)/§506(b) theory) — knowledge that is not in the mounted document.

A one-shot worked example does not substitute for training here: base+demo submits no more often than base (27/52 vs. 27/52 non-empty), unlike navigation (§5.2), where one demonstration recovers most of the zero-shot gap. The demo also narrows arguments for the one frontier model that did not need it (DeepSeek 0.516 → 0.428 paired on shared clean points, $p \approx .001$) and roughly doubles wrong-side arguments, concentrated on the side matching the exemplar's label — scaffolding is not free, in either direction.

Citation validity is saturated; citation alignment and side conformity are the measured headroom, and both admit verifiable reward terms (alignment-F1 against a provision-derived reference built without the court's reasoning; an assigned-side conformity check). Per-verdict justifications, both graders' logs, and capture scripts ship with the release.

M General-capability regression check

Domain training converts sampling diversity into first-sample reliability (pass@8 falls while pass@1 rises, §5.2), which raises the question of what the adapters cost in general capability. The answer is: essentially nothing measurable. On a 500-question stratified MMLU sample and 200 GSM8K problems (greedy decoding, no extended reasoning, identical prompts per arm):

Arm	MMLU (500)	GSM8K (200)
Untrained base	0.764	0.910
stage 3 (judge-anchored)	0.752	0.915
stage 4 (full pipeline)	0.752	0.925

Table 13: General-capability check, base vs. adapters. The full pipeline costs 1.2 points of MMLU and gains 1.5 points of GSM8K — within sampling noise at these sizes. The RL-stage narrowing is domain-internal mode concentration, not capability loss; and because every stage trains a LoRA adapter on a frozen base, the base model is recoverable exactly.

N Sandbox commands

The container runs a standard POSIX shell over the mounted document tree. The commands the model uses in practice are:

Command	Purpose	Typical use
ls	list directory contents	see which sections exist
cat	print a whole file	read one section in full
grep	print lines matching a pattern	find a term inside a section
rg	recursive pattern search (ripgrep)	find which sections mention a term
sed	print or edit selected lines	read a line range of a long section
python3	run a short script	arithmetic
submit_answer	submit the structured answer	ends the episode

The instrument is mounted read-only; the working directory is a fresh writable temporary directory; there is no network access; each call is time-limited and its output truncated; and the episode ends at the tool-call cap or on submit_answer.

O Inference settings by arm

Every evaluation arm's decode regime, for the fairness audit the cross-model comparisons require.

P Datasheet

Following the practice of Gebru et al. (2021), abridged; the release archive carries the full version.

The dataset exists to evaluate — and train for — the skill legal-NLP benchmarks do not test: locating operative language inside full debt instruments, resolving defined terms, and predicting litigated outcomes, with every reward grounded in document structure or court records rather than model opinion.

Arm	Engine	Thinking	Max new	Temp	k	Coverage
Agentic, local (base/st. 3/st. 4)	vLLM AWQ-4bit	on	1024	0.8	8	825 (stage-4 ×3)
Agentic, 397B (API)	DeepInfra	off	700	0.8	2	114, budget-capped
Demo-injected cells	as parent	as parent	as parent	0.8	8 / 2	20 tasks/type
Judge, local (adapters + base)	vLLM AWQ-4bit	on	1024	0.8	16/11	101 prompts
Judge, API (397B/9B/Kimi)	DeepInfra	off	900–1200	0.8	8	101 prompts
Judge, API (DeepSeek-V4-Pro)	DeepInfra	n/a	—	0.8	8	101 prompts

Table 14: Every arm's decoding regime is disclosed. The base judge run recorded 11 votes per prompt (0.91 parseable) vs. 16 (0.98) for the adapter arms; DeepSeek-V4-Pro exposes no hybrid-thinking switch. Navigation tables use DeepSeek-V3.2 (the tool-calling arm evaluated there); judge and dispute-suite tables use DeepSeek-V4-Pro (both judge arms appear in Figure 3). "Thinking off" on API arms is forced by the serving provider: with hybrid thinking enabled, the reasoning budget is routed to a separate reasoning_content field and the message content (which must carry the tool call or the ruling line) returns empty. API judge prompts additionally append a suffix capping stated reasoning at 120 words. The agentic API arm truncates transcripts at 26,000 characters (≈6.5k tokens) vs. the local arms' 8,192-token context; both hit the same 10-tool-call episode cap. The thinking asymmetry favors the local arms and is one reason the zero-shot 397B agentic rows are not the load-bearing comparison (§5.2).

The dataset comprises: 424 annotated covenant disputes (461 binary issue prompts; the 89-case / 101-prompt held-out evaluation set is hand-audited line-by-line); 43 mounted instruments, each quote-audited against its deciding opinion where retrievable (§3); a 4,042-task audited L1 evaluation suite over 41 instruments (the 2,701-task L1 training suite is regenerable from the released corpus and generators); 48 L2 case files; a 50-task instrument-verified dispute suite (Appendix G); a 200-document pre-decision filing corpus (27 disputes two-sided, 17 pristine). Splits are case-level by content hash; an 89-case pristine registry appears in no training loop (Table 2).

Everything is collected from public primary sources. Instruments: public SEC EDGAR filings located by full-text search and confirmed against the record's instrument description. Court records: published opinions and public dockets (CourtListener/RECAP); brief documents purchased from PACER under the quarterly fee waiver are donated back to the public RECAP archive. Annotation: LLM-assisted under human direction with released per-task audit verdicts; checker gold references independently re-derived on a 30-task sample (§3); evaluation-suite keys derive mechanically from an audited parse, with depth-3 keys checked by two frontier readers and human adjudication (Appendix E); an attorney-audited subsample remains future work.

Exhibits are parsed to sectioned trees. HTML/text exhibits are parsed with defined-term maps and cross-reference graphs (parser and per-document quality metrics released); prompts ship in named variants (canonical symmetrized, caption-anonymized, party-scrubbed).

The intended use is research evaluation and RL training for document-grounded legal reasoning. Outputs are not legal advice; the labels record how specific courts ruled on specific records, not legal truth.

Distribution is versioned; maintenance is by the author. Versioned release archive (v1.1) with append-only result logs, an embedded canary string, and per-version change notes. All underlying documents are public records; no non-public personal data is introduced. The distribution comprises the evaluation benchmark, corpus, generators, environment, checkers, an adapter manifest (SHA-256-pinned; weights ship via the artifact archive), and evaluation result logs; training-split task rows and annotation-derived training prompts are not distributed. Known limitations are documented in-release: a residual reference-noise band on the training-suite generators; per-mount quote-audit verdicts (§3); and a case-level (rather than dispute-cluster-level) train/eval split. Maintained by the author; issues via the release repository.

Q Reproducibility

A canary string tags the release against training contamination. The archive embeds covenantbench:canary:7bb0404a-e18e-4439-b543-e57421cbca9b at its root, with a request that benchmark items not be reproduced in plain text online and that documents containing it be excluded from training corpora, following the practice of GPQA Rein et al. (2024).

All figures regenerate from the released result files via a released script; splits derive from sha256(case_id) (frac 0.2) and the frozen released registry. Training ran on two consumer GPUs: AWQ-4bit vLLM rollouts on one and an NF4 QLoRA trainer on the other, over file-queue IPC. Adapters for stages 1–4 (SFT through debate). Rollout regime for all reported evals: thinking enabled, temperature 0.8.