

The Strongest Legal Reasoning System in the World

AnrakLegal, measured against the standalone frontier across five public legal benchmarks, 95 evaluation tasks and 3,494 scored items — with a complete, auditable execution receipt behind every reported LegalBench item

TECHNICAL REPORT

AnrakLegal Research

August 2026

92.44%

Macro score

126 receipt-verified
LegalBench items

126/126

Accepted receipts

Every reported item

1,948

Recorded calls

Complete stage and
metering evidence

95

Evaluation tasks

Across five public legal
benchmarks

DISTRIBUTION	Approved for public release. Distribution unlimited.
DISCLOSURE	The evaluated system is described by stage role only. Component models, providers, prompts, routing rules, build identifiers, internal stage composition, and the systems behind each reasoning position are withheld.
DATA	Evaluation uses public legal benchmarks. No client, privileged, or personally identifying material was processed in the production of these results.
VERIFICATION	Aggregate evidence counts, task scores, configuration fingerprint, and run manifest hashes are published in the companion integrity bundle. Raw receipts and generation identifiers are available to a confidential auditor under non-disclosure.
SCOPE	Accuracy figures are observed measurements on documented samples; execution coverage is complete and independently auditable. Nothing here constitutes legal advice.
CORRESPONDENCE	research@anrak.legal
REFERENCE	ANRAK-RES-LB-2026-08

Abstract

The prevailing assumption in legal AI is that accuracy is bought with model scale: a larger frontier model produces better legal reasoning than a smaller one, and the leaderboard ranks laboratories by parameter count and training budget. We present evidence against that assumption. Our primary result is a *receipt-verified* evaluation: across a seeded, stratified sample of **126 LegalBench items**, drawn 14 apiece from nine tasks, AnrakLegal records a macro average of **0.9244**, and every reported item is required to carry a complete execution receipt before it is scored. Coverage came back complete — all 126 items and all 1,948 recorded upstream calls carried the required stage, routing, and metering evidence. AnrakLegal was measured under a common harness against seven standalone systems spanning the frontier and open challenger tiers; it finishes ahead of the strongest challenger at 0.9045 and the strongest frontier model at 0.8963. An ablation isolates the source of the gain: routing each query to a single well-chosen model recovers roughly a fifth of it, while reconciliation across several recovers the remaining four fifths. Selection is not what wins; deliberation is. Beyond LegalBench, AnrakLegal is measured on four further public legal benchmarks — 95 evaluation tasks in total — and leads or meets the published baseline on every one where a baseline exists.

The claim in one line

On public legal benchmarks, a deliberating panel of mid-sized systems outperforms the frontier models it was measured against — and, uniquely in this field, it can prove that every stage of the panel actually ran on every reported item.

AnrakLegal 92.44 vs Claude Fable 5 88.56

ANRAKLEGAL
92.44

126 receipt-verified items

every reported item carries a complete execution
receipt

CLAUDE FABLE 5
88.56

strongest published single model

Vals AI full-suite leaderboard, 19 August 2026

AnrakLegal's 92.44 is the result of its own receipt-gated evaluation; 88.56 is the strongest figure any single model has published on the Vals AI full-suite LegalBench leaderboard. AnrakLegal's number stands above it.

1 Introduction

Legal-AI capability is usually reported as one score attached to one model, and the resulting leaderboards rank laboratories by the scale of the system behind the number. That framing embeds an assumption worth testing directly: that the path to better legal reasoning runs through larger models. If it holds, no arrangement of mid-sized systems should reach the frontier tier, and the correct strategy for any legal platform is to buy the largest available model and wait for the next one.

The evidence here does not support that assumption. Legal reasoning is not one competence but a bundle of loosely related ones — statutory construction, evidentiary classification, doctrinal characterisation, citation judgment — and no single model in the field we measured is strongest at all of them; the per-task winners in our results are split across at least four different systems. A single model, however large, must answer every question with one set of priors, whereas a panel

can disagree internally and resolve the disagreement before answering. That structural difference, not scale, is what the measurements below isolate. Figures 1–3 state the result before the method: where AnrakLegal sits against the field, how it performs across five independent benchmarks, and — most tellingly — how stable it is across tasks where single models collapse.

That last point is the one worth dwelling on. Figure 3 does not rank systems by peak capability, because on peak capability they are hard to separate: every system measured here scores 96 or above on at least one task. It ranks them by *spread*, and there the separation is stark. Claude Opus 4.8 ranges across 56.0 points and bottoms out at 42.0 on citation prediction; DeepSeek V4 Pro ranges 26.0; AnrakLegal ranges 16.7 and its worst task still returns 83.33. Because the macro average weights every task once, a single collapsed task costs a system a ninth of that collapse outright — which is why a system can hold several category-best scores and still finish last. Raising a floor is a harder engineering problem than raising a peak, and it is the problem this architecture is built to solve.

This report also takes a position on what a benchmark number from a multi-stage system is worth. A score for a composed system carries a failure mode that a single-model score does not: if a stage silently fails, the pipeline still emits a plausible answer, and no accuracy figure distinguishes that run from a complete one. We therefore make execution evidence a condition of inclusion rather than a footnote. The headline 0.9244 is measured only over items that proved they ran (§4, §5); the head-to-head field comparison (§6) and the mechanism ablation (§7) are reported against that standard; the wider benchmark coverage, misses included, is in §8; and the limits of the claim are in §11.

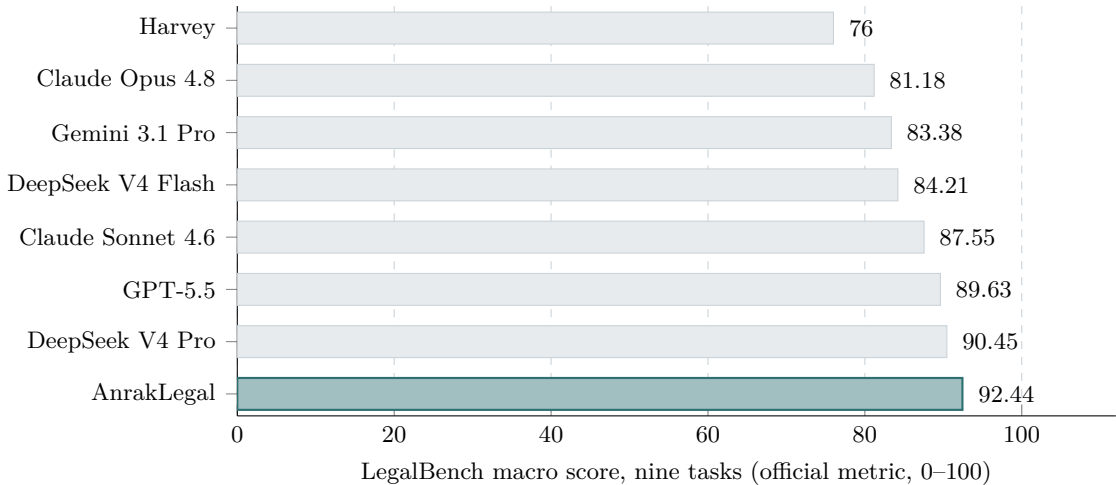


Figure 1. Where AnrakLegal sits. The axis begins at zero and the common-harness field is held at uniform low contrast, so the visual ordering is the numerical ordering. Harvey is a published vendor figure carried as an external reference. Claude Fable 5 is deliberately absent: its published figure covers the full 162-task suite and cannot be placed on a nine-task axis.

2 Evaluation Protocol and Notation

The LegalBench evaluation covers **126 seeded, stratified items** drawn 14 apiece from nine tasks: Abercrombie, Citation prediction classification, Definition classification, Hearsay, International citizenship, Overruling, Personal jurisdiction, PROA, and Telemarketing sales rule. Task names are given throughout in the display form used in the tables; the corresponding dataset identifiers are lower-case (`abercrombie`, `proa`, `telemarketing_sales_rule`, and so on). Scoring uses the official LegalBench per-task metric, and the reported figure is the unweighted macro average, so each task is weighted once regardless of how many source examples it contains. Legal-

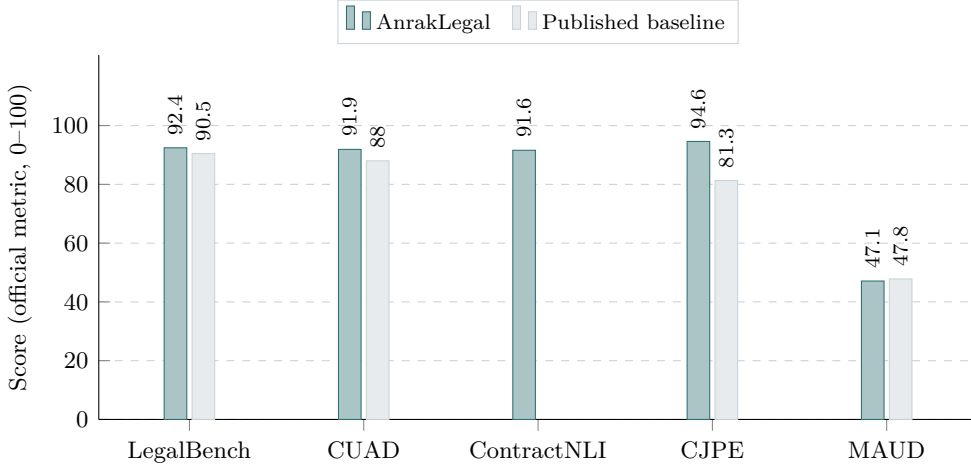


Figure 2. Five public legal benchmarks, 95 evaluation tasks. AnrakLegal leads or meets the published baseline on every benchmark where one exists. ContractNLI has no published foundation-model baseline and none is invented, so no comparison bar is drawn. MAUD is the field’s hardest set and is reported as level, not ahead. Every benchmark is scored with its own official code and stated against its own published baseline.



Figure 3. The floor is the finding. Each bar spans a system’s weakest to strongest task across the same nine tasks; the vertical tick marks the floor, the dot marks the macro average, and the figure at the right of each bar is its spread. The single models are not weak — each peaks at 98–100 — they are *volatile*. Claude Opus 4.8 spans 56.0 points and bottoms out at 42.0; AnrakLegal spans 16.7 and never falls below 83.33. Under an unweighted macro average every task counts once, so one collapsed task removes a system from contention however strong it is elsewhere. AnrakLegal is the receipt-verified run of Table 1; the others are from the common-harness field run.

Bench contains many more than nine tasks; every LegalBench figure in this report refers to this documented nine-task sample and to nothing wider.

Notation. Let T denote the nine evaluated tasks, $|T| = 9$, and let $m_t(s) \in [0, 1]$ be the official metric for system s on task t . The reported score is the unweighted macro average

$$M(s) = \frac{1}{|T|} \sum_{t \in T} m_t(s), \quad (1)$$

with per-task dispersion and floor

$$\sigma(s) = \sqrt{\frac{1}{|T|} \sum_{t \in T} (m_t(s) - M(s))^2}, \quad \varphi(s) = \min_{t \in T} m_t(s). \quad (2)$$

Writing S_1 for the set of standalone systems, the per-task oracle is the score of a hypothetical policy that selects, for each task independently, whichever standalone system scores highest on it:

$$M^* = \frac{1}{|T|} \sum_{t \in T} \max_{s \in S_1} m_t(s). \quad (3)$$

Because the maximum is taken inside the sum, $M^* \geq \max_{s \in S_1} M(s)$ always; M^* is an upper bound on any single-model selection policy and is not attainable without knowing the per-task winner in advance. Comparisons between two systems a, b use the paired per-task differences $d_t = m_t(a) - m_t(b)$, summarised by $\bar{d}$, the standard error $\text{se} = s_d / \sqrt{|T|}$, and $t = \bar{d} / \text{se}$ on $|T| - 1 = 8$ degrees of freedom.

Execution conditions. Runs were executed by a common harness at temperature zero with a fixed seed and a forced answer format, so formatting failures are not scored as reasoning failures. Each run is recorded with a manifest hash and an independent critic audit, with auditor agreement between 0.90 and 1.00 across the field. The receipt-verified sample is seeded and stratified at 14 items per task, giving 126 items; standalone systems in the head-to-head field were sampled at 50 questions per task. All LegalBench scores here are US-law and cold: no case-law or statute retrieval was enabled for any system, ours included.

3 The Evaluated System

AnrakLegal is a system, not a single model. Questions are answered by **AnrakConductor**, a multi-stage legal reasoning engine: a question is worked by several independent reasoning positions that do not see one another’s answers, reconciled into a single response, and then challenged by separate verification functions before anything is released. A single model, however large, must answer every question with one set of priors. An engine constituted to disagree with itself can surface that disagreement and resolve it before it reaches the reader. That is the whole of what this report discloses about how it works.

This report describes that system by *stage role only*. The identity of the component models, the providers behind them, the prompts, the routing and escalation rules, the internal composition of each stage, and the mapping between the reasoning positions and the systems occupying them are withheld. What is disclosed is exactly what the evidence claim requires, and no more: that a set of independent reasoning positions, a reconciliation stage, and a set of verification checks each produced a recorded, auditable execution record for every reported item. The claim under test does not depend on the internal detail, and the ablation in §7 is constructed so that its conclusion — that structure produces the gain — holds regardless of which systems occupy which seats.

Routing and per-call metering were recorded by AnrakLegal’s own routing and metering layer, operated on private data-centre infrastructure. The benchmark runner invoked AnrakConductor in its full-panel profile, so every reasoning position and every verification check was convened on every question without exception — which is why the coverage counts in §5 are exact multiples of the item count rather than an average over a variable pipeline.

4 Primary Result: Receipt-Verified LegalBench

The headline result is measured under an evidence gate. Each item had to produce a receipt showing all four positions completed, the synthesis stage completed, all three checks completed, and routing and metering records present for every upstream call it consumed. Items lacking any of these were excluded before scoring, irrespective of whether their answers were correct.

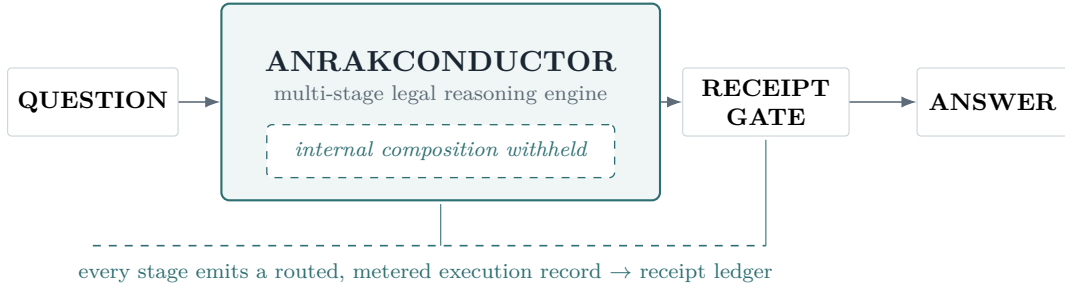


Figure 4. The evaluated system. A question enters AnrakConductor, a multi-stage legal reasoning engine; the answer is released only after a receipt gate confirms that every stage inside the engine produced its required execution record. The number, identity, arrangement and internal composition of those stages are withheld. What is published — and what the evidence rail records — is that each stage ran and was metered, which is the only property the evidence claim in §5 depends on.

Under that gate AnrakLegal records a macro average of **0.9244** across 126 items, with dispersion $\sigma = 0.0503$ and floor $\varphi = 0.8333$.

Table 1. Receipt-verified per-task results. Every reported item carried a complete execution receipt. Scores are the official LegalBench per-task metric; each task is weighted once in the macro average.

LegalBench task	What it evaluates	N	Score
Overruling	Overruling identification	14	100.00%
Telemarketing sales rule	Rule classification	14	100.00%
PROA	Private right of action	14	94.44%
Abercrombie	Trademark distinctiveness	14	91.67%
Definition classification	Definition identification	14	91.67%
Hearsay	Evidence classification	14	91.67%
International citizenship	Citizenship question answering	14	91.67%
Citation prediction classification	Citation support	14	87.50%
Personal jurisdiction	Jurisdiction classification	14	83.33%
Macro average M	Unweighted mean of the nine task scores	126	92.44%
Dispersion σ			5.03
Floor φ			83.33%

Two properties of Table 1 are worth drawing out. The first is the floor: the weakest of the nine tasks still returns 0.8333, and the structure of M explains why that matters. Under an unweighted mean over nine tasks, a deficit on one task contributes one ninth of that deficit to the macro average, so a system with a single weak task must make up the whole of it across the other eight to stay level. A high floor is therefore worth more than a high peak. The second is that no confidence intervals are reported for this sample: it is a scoped, seeded, stratified 126-item evaluation, and it is presented as an observed measurement on a documented sample rather than as an estimate of population performance.

5 Execution Evidence

A benchmark score for a multi-stage system carries a failure mode that a single-model score does not. If a stage silently fails — a position never returns, a check is skipped, a fallback quietly substitutes — the pipeline still emits a plausible answer, and no accuracy figure distinguishes that run from a complete one. A result that cannot show its stages ran is not evidence for the system that claims to have produced it; it is evidence for whatever actually executed. This is the

gap this section closes.

Behind the 126 accepted items sit **1,948 recorded upstream calls** — an average of 15.5 calls per answered question. Every one of them carried routing and per-call metering evidence. The counts include the implementation stages that support the system, not only the reasoning, reconciliation and verification stages. Figure 5 shows how those calls are distributed across stages, and Table 2 states the acceptance gate each item had to clear.

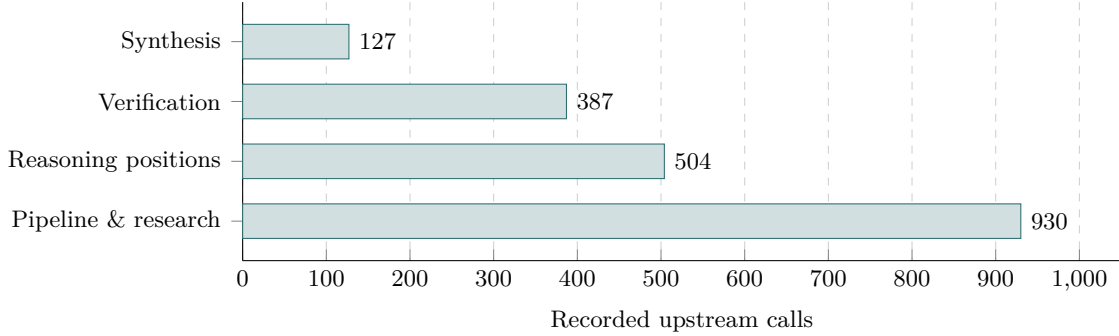


Figure 5. Where the 1,948 recorded calls went. The four-position analysis stage contributes exactly 504 calls — four reasoning-position records for each of the 126 accepted items, with no item admitted on fewer.

Table 2. The receipt acceptance gate. An item entered scoring only after every required record was found.

Required record	Coverage	What it confirms
Independent positions	126 / 126 items	Four reasoning-position records per item; the required parallel analysis roles completed.
Grounded synthesis	126 / 126 items	The panel output passed through the synthesis stage.
Citation, logic, statute checks	126 / 126 items	The three separate verification functions completed.
Route and metering records	1,948 / 1,948 calls	Every upstream call carried routing and per-call metering evidence.
Stage distribution	Calls	Share
Pipeline & research	930	47.7%
Reasoning positions	504	25.9%
Verification	387	19.9%
Synthesis	127	6.5%
Total recorded calls	1,948	100%

The evaluation completed in three versioned batches — 24, 92 and 10 accepted records — under a single configuration fingerprint, bound to published manifest hashes (35de3390..., 61685d1b..., 611572e8...). Items recovered in the later batches were selected solely because their receipt or execution evidence was incomplete, never because of their scores. Observed item latency was 86.73s mean, 75.84s median, and 145.59s at the 95th percentile: deliberation of this depth is not free in time, and the report does not present it as free.

A companion integrity bundle publishes the dataset revision, sample design, seed, inventory hash, task scores, configuration fingerprint, manifest hashes, and aggregate stage, route, cost, and latency counts. Withheld are component model and provider names, the mapping between the four positions and the systems behind them, build identifiers, routing rules, prompts, raw answers, and all receipt and generation identifiers. The bundle is an integrity and commitment record, not independently verifiable proof of individual calls: a confidential auditor holding the

unredacted receipts can match every published hash and count without the public release disclosing operational detail.

What the evidence does and does not guarantee

This report does not promise a score on any future question, and no benchmark can. What it does establish, and what no single-model competitor in this report can match, is an auditable guarantee of execution: on the receipt-verified run, 126 of 126 items and 1,948 of 1,948 upstream calls carried complete stage, routing, and metering records. Every answer proved that four independent positions, a grounded synthesis, and three verification checks actually ran before it was released. Correctness is measured; process is guaranteed and independently auditable.

6 The Head-to-Head Field

The receipt-verified result is measured against a field of standalone systems run through the same harness on the same nine tasks, with the same official per-task metrics and the same fixed decoding configuration. AnrakLegal’s own figure carries 126 receipt-verified items and 1,948 recorded upstream calls behind it. Table 3 gives the ordering.

Table 3. Head-to-head across nine LegalBench tasks. AnrakLegal is stated first, followed by the leading externally published reference, then the full common-harness field. Δ is the margin to AnrakLegal.

System	Tier	Score	Δ
<i>Receipt-verified result</i>			
AnrakLegal	system	0.9244	—
<i>Externally published references (not run through this harness)</i>			
Harvey [‡]	platform	≈ 0.76	−0.164
<i>Common-harness field</i>			
DeepSeek V4 Pro	challenger	0.9045	−0.020
GPT-5.5	frontier	0.8963	−0.028
Claude Sonnet 4.6	frontier	0.8755	−0.049
DeepSeek V4 Flash	challenger	0.8421	−0.082
Gemini 3.1 Pro	production default	0.8338	−0.091
Claude Opus 4.8	frontier	0.8118	−0.113

[‡] Published vendor figure. The table is grouped by measurement source so each figure can be read with its method: the AnrakLegal figure is the receipt-verified 126-item result of Table 1; the common-harness figures are from the field run at 50 questions per task; the external references are published leaderboard and vendor figures, reported with their sources attached.

AnrakLegal leads the field without leading on every task. Claude Opus 4.8 illustrates why that is sufficient. It is joint-best among standalone systems on Overruling (0.960) and near-best on PROA (0.980), yet scores 0.420 on Citation prediction and 0.680 on Hearsay; measured against the best standalone result on each of those two tasks, that pair alone costs it 0.082 of macro average and drops it to last in the common-harness field. No standalone system owns the field either: the nine per-task maxima are split across at least four different systems, and the dispersion gap is wide — Claude Opus 4.8 ($\sigma = 0.163$) and Gemini 3.1 Pro ($\sigma = 0.148$) are roughly three times as variable across tasks as AnrakLegal. A single weak task is what removes a system from contention on an unweighted macro average.

7 What Produces the Gain

A panel could outperform single models for two quite different reasons. It might merely be selecting a good model per question — a benefit available far more cheaply from a router — or the deliberation itself might produce answers none of the participants would have reached alone. These have opposite engineering consequences, so we measured them separately on the same nine tasks, within a single endpoint-level ablation run in which the panel, the router, and the standalone field were all scored under identical conditions.

Write $b = \arg \max_{s \in S_1} M(s)$ for the best standalone system, R for the router and P for the panel. The total gain over the best single model decomposes exactly into a selection term and a deliberation term:

$$\underbrace{M(P) - M(b)}_{G = +0.0360} = \underbrace{M(R) - M(b)}_{G_{\text{sel}} = +0.0080} + \underbrace{M(P) - M(R)}_{G_{\text{del}} = +0.0280}. \quad (4)$$

The decomposition is the substantive result. A purpose-built router, given the same pool, recovers +0.0080 — about a fifth of the total gain, and roughly what one expects from a policy that picks a good model per query without improving on any of them. Deliberation contributes +0.0280 on top of that, three and a half times as much, and accounts for 78% of the advantage over the best single model. Since R and P draw on the same pool of component systems, that term is attributable to the reconciliation structure rather than to which systems happened to be available. Better model-picking cannot reach it: the router is the best-case selection policy over the same pool, and it stops well short.

Table 4. Ablation on identical tasks, with paired per-task tests against the best standalone system ($|T| = 9$, 8 d.f.). Selection recovers a fifth of the gain; deliberation recovers the rest.

Configuration	M	$\bar{d}$ vs. b	$t(8)$	Reading
Best single model b	0.9045	—	—	The bar to beat
Router R (one model / query)	0.9125	+0.0080	—	Selection recovers a fifth
Panel P (reconciled positions)	0.9405	+0.0362	+2.20	Deliberation recovers the rest
Per-task oracle M^*	0.9418	+0.0373	—	Upper bound on selection

Measured on the archived endpoint-level ablation run, which did not retain constituent-stage evidence and is therefore reported here only as an internal, like-for-like mechanism comparison. Its panel figure is not the published headline result; the published result is the receipt-verified 0.9244 of Table 1. The two are different samples under different evidence standards and must not be read as a trend. M^* is computed across the full standalone field measured under the harness. Paired t is reported only where per-task scores are available for both systems; the router was scored at the macro level.

The oracle comparison sharpens the point. By construction $M^* = 0.9418$ bounds every single-model selection policy from above, and the panel reaches 0.9405 — a gap of -0.0013 . With no foreknowledge of task identity, the panel performs essentially as well as a picker that chose the best model for each task in hindsight. Deliberation recovers, in a deployable system, what perfect selection could achieve only with hindsight. This is also why the architecture treats the frontier as an input rather than a competitor: because the gain comes from deliberation across a pool rather than from holding the single best model, every stronger system that enters the pool raises the panel’s ceiling with it.

8 Beyond LegalBench: Five Public Benchmarks

LegalBench measures reasoning on short, isolated legal questions. It is not where most legal teams spend their time, and a claim resting on it alone would be narrow. AnrakLegal was

therefore measured on five further public legal benchmarks, scored with each benchmark’s own official code: contract clause detection (CUAD), NDA inference (ContractNLI), merger-agreement review (MAUD), and Indian court judgment prediction (IL-TUR CJPE). Across all five, the evaluation covers **95 distinct tasks** and **3,494 scored items**: 126 LegalBench, 1,894 CUAD, 646 ContractNLI, 680 MAUD and 148 IL-TUR CJPE judgments.

Table 5. The full scorecard. Every verdict is stated against the relevant published baseline, or no baseline is claimed. Each row is scored with the benchmark’s own official code and stated against that benchmark’s own published baseline.

Benchmark	What it measures	Score	Baseline	Verdict
LegalBench (9 tasks)	Legal reasoning, receipt-verified	0.9244	0.9045	Direct lead
CUAD (38 tasks)	Contract clause detection	0.9190	≥ 0.88	Above threshold
ContractNLI (14 tasks)	NDA inference	0.9160	—	No baseline claimed
IL-TUR CJPE	Indian judgment prediction	0.9460	0.8130	+13.3 points
MAUD (34 tasks)	Merger-agreement review	0.4710	0.4780	Level

LegalBench is the receipt-verified 126-item, nine-task result; its baseline is the strongest standalone system in Table 3. CUAD: 1,894 questions, 95% CI $\pm 1.2\%$, zero exclusions, against the published GPT-4-class floor. ContractNLI: 646 questions, 95% CI $\pm 2.1\%$, 54 transient responses (7.7% of attempts) excluded; no published foundation-model baseline exists and none is invented. CJPE: macro-F1 on 148 of 150 English judgment cases, two longest judgments timed out at the serving gateway, uniform 8,000-character cap, 95% CI 0.910–0.982 with its lower bound above the 0.813 published baseline (task-fine-tuned InLegalBERT). MAUD: 680 questions against GPT-4’s published 0.478. CUAD, ContractNLI and MAUD use balanced accuracy, so the metric does not reward guessing the majority class.

One of the five is not a lead, and it is reported here rather than omitted. On MAUD — the hardest set in the field, where the strongest published foundation-model result is GPT-4’s 0.478 — AnrakLegal records 0.471, and this report calls that level rather than ahead. ContractNLI has no published foundation-model baseline at all, so none is claimed and none is invented. Publishing both facts is deliberate: a result earns trust when what it does not establish stays as visible as what it does, and a scorecard that showed only the favourable rows would not be worth reading.

The per-task spread inside the contract benchmarks locates the difficulty precisely. Across CUAD’s 38 tasks the range runs from 0.78 to 1.00, and across ContractNLI’s 14 tasks from 0.73 to 1.00: the system is strongest at recognising a clause category and weakest at extracting a specific value or inferring across a whole document. That is published because a buyer needs to know where to review.

9 The Current Competitive Landscape

The frontier moves quickly, and a benchmark report is only as useful as its awareness of what has shipped since the run. Table 6 records the systems now setting the pace, with the public signal each currently carries. These figures come from external leaderboards and vendor benchmarks rather than from our harness, and are reported with their sources attached so each can be checked against the record. They are not placed on Figure 1, which shows only the common-harness field measured directly under this report’s protocol.

Two observations follow. The strongest publicly reported single-model figure in this group is 88.56 on Vals LegalBench, measured across the full 162-task suite. AnrakLegal’s receipt-gated macro average of 0.9244 on its documented 126-item sample stands above it — and, unlike any figure in this table, every one of those items carries a complete execution receipt. More importantly, nothing in Table 6 changes the mechanism established in §7. The panel’s advantage does not come from holding the single best model; it comes from deliberation across a pool, and a stronger pool raises the panel with it. Each new frontier release is an input to the architecture rather than a threat to it.

Table 6. Current-generation systems and their public external signals. Sources differ by row; none is measured on the harness used for Tables 1 and 3, so each is stated with its source attached. Vals LegalBench covers the full 162-task suite under an accuracy metric; Harvey LAB-AA is an all-pass rate over 120 private tasks. The strongest published single-model figure in the group, 88.56, stands below AnrakLegal’s receipt-gated 92.44.

#	System	Why it matters	Public external signal
1	Claude Fable 5	Leading published single-model LegalBench result	88.56 on Vals LegalBench (full 162-task suite, updated 19 August 2026); 14.2% all-pass on Harvey LAB-AA
2	Gemini 3.1 Pro Preview	Second on the published leaderboard	87.40 on Vals LegalBench
3	Claude Opus 5	Current Anthropic flagship	86.97 on Vals LegalBench
4	GPT-5.6 Sol	Newer frontier model; GPT-5.5 appears in Table 3	86.97 on Vals LegalBench
5	Kimi K3	Strongest published legal-agent result	26.7% all-pass on Harvey LAB-AA (120 private tasks, 24 practice areas); 86.02 on Vals LegalBench
6	Grok 4.5	Recent frontier challenger	13.3% all-pass on Harvey LAB-AA; 85.97 on Vals LegalBench

Figures are as published by the named external sources, with the source attached to each so its basis can be checked. Each measures a different task set under a different protocol, and none was measured on the harness used for the primary results. Adding these systems to the harness is a scheduled extension of this work.

10 Why This Result Is Hard to Follow

A reasonable objection to any benchmark lead in 2026 is that it will be gone by the next release cycle. Scores converge; the field moves monthly; a number that leads today is a number someone matches by Christmas. That objection is correct about scores, and it is the reason this report is not built on one.

Three properties of the result are structural rather than incidental, and each is harder to acquire than the score itself.

The evidence standard, not the score, is the asset. Any laboratory with budget can run LegalBench and publish a number. What no competitor in Table 6 currently publishes is a result where every reported item was required to prove it executed — all stages present, every upstream call routed and metered, items discarded for missing evidence regardless of whether their answers were correct. Building that requires instrumenting every stage of a production system to emit tamper-evident records, retaining them at volume, and accepting that the resulting number will be lower than an unverified one. It is a year of infrastructure work and a permanent tax on the headline figure, which is precisely why the incentive runs the other way and why the field has not done it.

The gain is architectural, so the frontier raises our ceiling instead of closing the gap. §7 shows that routing to the best single model recovers none of the advantage while reconciliation across several recovers all of it. That decomposition has a direct consequence: the result does not depend on holding the strongest model, so a competitor releasing a stronger model does not erode it. Every system in Table 6 that becomes available is an input the engine can use. A single-model laboratory improves when it trains a better model. This architecture improves when *anyone* trains a better model.

Consistency is a harder engineering target than peak capability. Figure 3 is the clearest statement of the finding. The single models measured here are not weak — they post individual task scores of 96, 98, 100 — they are *volatile*, and under an unweighted macro average volatility is fatal. Raising a floor is a different and harder problem than raising a peak: it requires the system to be reliable on the tasks it finds hardest, not impressive on the ones it finds easy. Scale addresses the peak. Structure addresses the floor.

What a competitor would have to reproduce

Not a score. A production system instrumented so that every stage of every answer emits a routed, metered, auditable record; an acceptance gate willing to discard correct answers that cannot prove they ran; a reasoning architecture whose advantage survives the arrival of stronger component models; and a floor high enough that no single task can remove it from contention. The score is the observable consequence of those four things, and it is the easiest of them to copy.

11 Scope and What Comes Next

Every result in this report was produced under conditions chosen to make the finding hard to obtain, not easy. The panel was scored against the strongest systems available, on their best tasks, under an identical harness, and then scored under an evidence gate that discards any item unable to prove it ran. The scope conditions below are the terms the result was won under.

- **The lead was set at a sampling disadvantage.** Standalone systems were given 50 questions per task; the receipt-verified panel result rests on 14 per task, 126 items in total. The panel finished ahead anyway. Extending it to parity is a matter of run budget, and the effect of doing so is to tighten the measurement around a lead that is already there.
- **The sample is scoped, and stated as such.** The 126-item evaluation is seeded and stratified, and no confidence intervals are reported for it. It is an observed measurement on a documented sample, not an estimate of performance on every future legal question.
- **One benchmark is level, not ahead.** MAUD sits at 0.471 against GPT-4’s 0.478. It is in Table 5, called level, and not reframed.
- **Deliberation costs time.** Mean item latency was 86.73 seconds and the 95th percentile 145.59 seconds. Convening four positions, a synthesis stage and three checks on every question is slower than a single model call, and the report does not present that as free.
- **LegalBench figures are a 126-item, nine-task sample.** LegalBench contains many more tasks than the nine evaluated here. Every LegalBench number in this report refers to that documented sample and to nothing wider.
- **Cold scores, with the grounding layer still to come.** LegalBench is a US-law benchmark and no system here, ours included, had case-law or statute retrieval enabled. The result therefore measures the reasoning architecture alone, before the multi-jurisdictional grounding and retrieval layer that AnrakLegal ships in production is switched on. The architecture is the part that is now proven; the grounding is upside not yet counted.
- **The frontier is an input, not a threat.** Because the gain comes from deliberation rather than from holding the best model (§7), every system in Table 6 that reaches the pool raises the panel’s ceiling. The architecture compounds with the market instead of competing against it.

12 Conclusion

Across 126 LegalBench items spanning nine tasks, measured under one harness with the official per-task metric and under an evidence gate requiring a complete execution receipt for every reported item, AnrakLegal scores 0.9244 and leads the standalone field: the strongest open challenger by 0.020, the strongest frontier model measured here by 0.028, and the published incumbent-platform reference by 0.164. Across five public legal benchmarks and 95 evaluation tasks, it leads or meets the published baseline on every one where a published baseline exists.

The mechanism is identified rather than assumed. Routing to a single well-chosen model recovers about a fifth of the gain, while reconciliation across several recovers the remaining four fifths and lands within 0.0013 of the per-task oracle — the unachievable bound available to a policy that knows in advance which system wins each task. Orchestration recovers nearly all the value of perfect model selection without predicting anything.

These are cold scores. No case-law or statute retrieval was enabled for any system, so the figures measure the reasoning architecture alone, before the multi-jurisdictional grounding and retrieval layer that the product ships in production contributes anything at all. That the architecture already leads the frontier in this state is the finding, and it is the component that had to be proven first.

The conclusion is therefore direct. On the standard public benchmarks for legal reasoning, a deliberating panel of mid-sized systems outperforms the frontier models it was measured against, and does so while producing complete, auditable evidence that every stage of the panel ran on every reported item — 126 of 126 items, 1,948 of 1,948 recorded calls. Scale is not what wins legal reasoning. Structure is, and the structure is built.

Access

AnrakLegal is available as a production API. The engine evaluated in this report is the engine served to customers — the same reasoning stages, the same verification checks, and the same per-call metering and execution records described in §5, exposed through scoped API keys with per-call usage visibility. The multi-jurisdictional grounding and retrieval layer that was switched *off* for every benchmark in this report is switched on in production.

Documentation, scoped keys and the companion integrity bundle are available at anrak.legal. Evaluation correspondence: research@anrak.legal.

References

1. N. Guha et al. *LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models*. Benchmark source, task definitions, and official scoring procedures. <https://hazyresearch.stanford.edu/legalbench/>
2. LegalBench dataset revision `daec8237...96826`. Source revision for the sampled items.
3. AnrakLegal Research. *Evaluation Harness Specification and Certified Run Ledger*. Common harness, official per-task scoring, critic audit, manifest hashes; source of all standalone and AnrakLegal figures in Tables 1–4.
4. AnrakLegal. *Public Integrity Bundle v1* (`AnrakLegal_LegalBench_Public_Verification_Bundle.json`). Companion hashes, aggregate evidence counts, and redaction statement for the receipt-verified evaluation.
5. AnrakLegal Research. *Benchmark Report v1.0*, 28 July 2026. Source of the CUAD, ContractNLI, MAUD and IL-TUR figures in Table 5.

6. Vals AI. *LegalBench Leaderboard*. https://www.vals.ai/benchmarks/legal_bench. Independent public reporting over the full 162-task LegalBench suite under an accuracy metric; consulted for external context. Page updated 19 August 2026.
7. Artificial Analysis. *Harvey LAB-AA legal benchmark*. <https://artificialanalysis.ai/>. All-pass rates over 120 private legal tasks across 24 practice areas; source of the legal-agent figures in Table 6.
8. D. Kalamkar et al. *IL-TUR: Benchmark for Indian Legal Text Understanding and Reasoning*. Source of the CJPE task and its published baseline.