ALawBench: Benchmarking Frontier Agents in IPO Legal Due Diligence

Anonymous authors
Paper under double-blind review

Abstract

Legal due diligence for an initial public offering (IPO) requires reviewing extensive records across an issuer's corporate history, consumes thousands of lawyer hours, and demands professional judgement. Advances in frontier agents raise the prospect of conducting such matters end to end, but testing this prospect requires a realistic workspace, materials grounded in real disclosures, and rubrics grounded in legal practice. We introduce ALawBench, the first benchmark to evaluate agents on comprehensive legal due diligence for IPOs in the Chinese mainland and Hong Kong, a task usually undertaken by legal teams. Its legal workspace combines scanned documents, corporate databases with controlled distractors, and professional tools for the full IPO due diligence workflow. Built by 37 legal experts from public disclosures, ALawBench comprises 63 matters, 12,151 PDFs, 30,243 pages, and 15,557 rubrics. The rubrics measure diligence through review, verification, and report delivery, and judgement through risk identification and professional assessment. Across 21 frontier models, the best net score rate is 40.39%, and 12 models score below zero. Review, delivery, and risk identification improve with capability, but verification errors barely decline and professional assessment improves only marginally. Reading nearly all key documents leaves verification errors unchanged and adds little assessment credit. Frontier agents cannot yet deliver issuable IPO legal due diligence reports, as their diligence falls short at verification and their judgement stalls at assessment.

1 Introduction

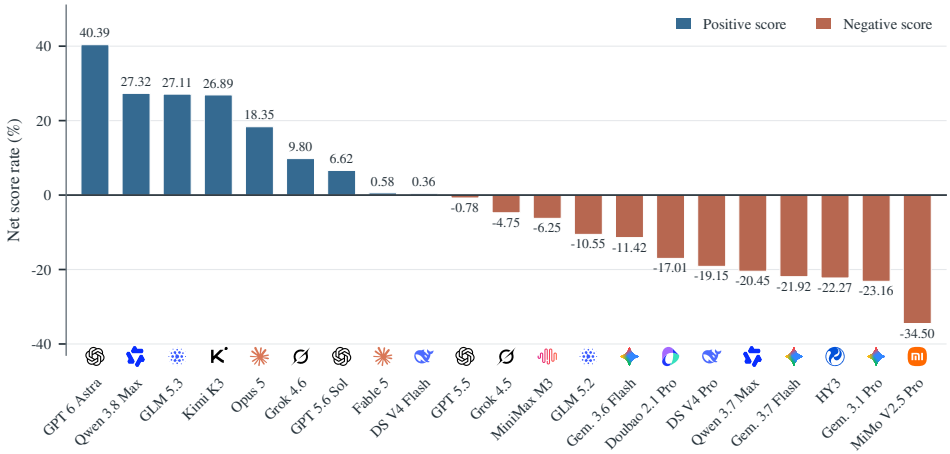


Figure 1: **Net score rates across frontier agents.** Blue and red indicate positive and negative scores; 12 of 21 models score below zero (Equation 1); Table 2 gives the action decomposition.

Anonymous repository: https://anonymous.4open.science/r/ALawBench-D0AA/README.md

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.

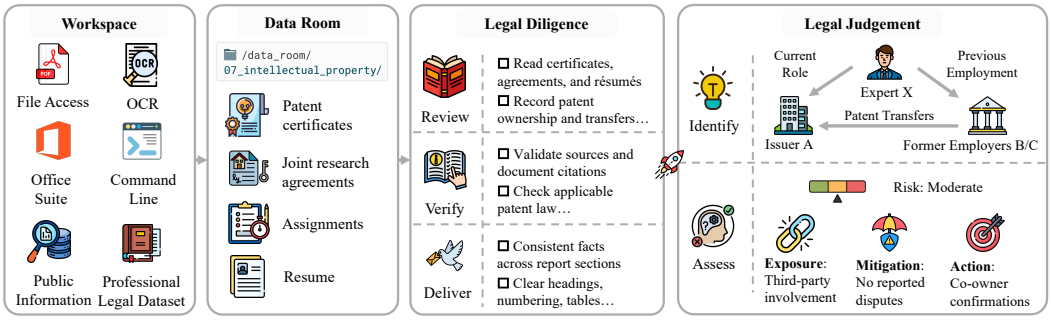


Figure 2: **One component of IPO legal due diligence: a running example.** Diligence requires reviewing Issuer A's patent certificates, joint-research agreements, assignments, and Expert X's résumé, verifying sources and applicable law, and recording consistent findings (Review, Verify, Deliver). Judgement requires connecting X's employment at B/C with transfers to A, identifying possible third-party interests, weighing the absence of reported disputes, and recommending co-owner confirmations (Identify, Assess). Both dimensions inform the report's patent-ownership analysis.

Legal due diligence (LDD) for an initial public offering (IPO) requires lawyers to reconstruct corporate history, reconcile extensive records, and assess risks that bear on disclosure and issuance. This work consumes thousands of lawyer hours and carries substantial costs and legal liability (Beatty & Welch, 1996; Hanley & Hoberg, 2010). As large language model (LLM) agents move toward complete professional workflows (Xu et al., 2024; Vidgen et al., 2026), IPO diligence tests whether they can connect evidence work with the legal analysis and advice that lawyers contribute to transactions, from the first document request to the final opinion (Gilson, 1984; Liu, 2006).

Drawing on lawyer-effectiveness research and theories of professional inference (Shultz & Zedeck, 2011; Abbott, 2014), we distinguish *diligence*, reliable investigation and reporting, from *judgement*, interpreting evidence to identify legal risks and recommend appropriate responses. Figure 2 illustrates the distinction through patent ownership. Five actions operationalize the two dimensions: Review, Verify, and Deliver for diligence; Identify and Assess for judgement (Section 2.2; Appendix A.1). Our central question is: *can frontier agents combine diligent investigation with the legal judgement needed to deliver a complete IPO due diligence report?*

Existing evaluations leave two connected gaps. **First, end-to-end diligence from raw records to an issuable report.** Legal benchmarks commonly supply facts directly or evaluate litigation workflows (Guha et al., 2023; Li et al., 2025; Akyürek et al., 2025; Zuo et al., 2026); Harvey LAB and Legora BAR reach data-room scale but target intermediate drafts for lawyer review (Grupen et al., 2026a; Pereyra, 2026; Lauritzen et al., 2026). ALawBench requires agents to find the facts themselves in scanned documents and external databases with controlled distractors, and grades the final report against an issuable-product standard (Table 1). **Second, diagnosing which capability limits an agent.** Task completion and error localization do not separate failing to find evidence from failing to check it or to reason from it. ALawBench decomposes each report score into five action contributions and pairs each report with its document-access record.

We introduce ALawBench, the first benchmark to evaluate agents on comprehensive IPO LDD in the Chinese mainland and Hong Kong, work usually performed by legal teams. Its 63 matters span 12 industries, with 12,151 PDFs, 30,243 pages, and 15,557 rubrics developed with 37 legal experts. We ask whether frontier agents can deliver issuable reports (RQ1), which actions limit them (RQ2), and whether broader evidence access removes these limits (RQ3).

Across 21 frontier models, the best net score rate is 40.39%, and 12 models score below zero. Review, Deliver, and Identify contributions rise steeply with capability. Verify contributions barely change, and Assess contributions rise only from 3.4 to 12.4 points and plateau among the five leading agents. Broader access removes neither limit. Reports that access at least 95% of key documents make as many verification errors as those that access fewer than half, and earn only 3.50 more Assess points. Frontier agents cannot yet deliver issuable IPO legal due diligence reports, as their diligence falls short at verification and their judgement stalls at assessment.

Table 1: Legal benchmarks by evaluated output and execution setting. Task gives the evaluated output. Long horizon denotes sustained work across connected steps within a matter. Data-room scale denotes a matter record of several hundred pages or more that requires selective search and cross-referencing. An issuable product is an end-to-end agent output assessed as the final deliverable, against the standard of a document the firm would issue to clients and regulators. ✓ marks coverage in every task, △ coverage in only a subset of tasks, and ✗ no coverage.

Benchmark	Primary jurisdiction	Task	Legal reasoning	Tool use	Long horizon	Data-room scale	Issuable product
Static legal benchmarks							
LEXam	Switzerland	Exam answers	✓	✗	✗	✗	✗
PRBench-Legal	U.S./Europe	Advisory answers	✓	✗	✗	✗	✗
JuDGE	China (mainland)	Court judgments	✓	✗	✗	✗	✗
Legal agent benchmarks							
LegalAgentBench	China (mainland)	Research answers	✓	✓	✗	✗	✗
LongJud-Bench	China (mainland)	Filings, advocacy	✓	✓	✓	✗	✗
Harvey LAB	U.S./Europe	Working drafts	✓	✓	✓	△	✗
Legora BAR	U.S./Europe	Working drafts	✓	✓	✓	△	✗
ALawBench (ours)	**China (mainland/HK)**	**IPO LDD reports**	✓	✓	✓	✓	✓

Our contributions are threefold. **(1) A complete long-horizon IPO diligence benchmark** combines scanned workspaces, database distractors, and expert-authored, evidence-linked rubrics. **(2) An action-based evaluation framework** decomposes the net score rate into five contributions and connects document access with report analysis. **(3) Diagnostic findings** show that verification and professional assessment limit even the strongest agents, including reports that access nearly all key documents, which points to reasoning rather than retrieval.

2 ALawBench

2.1 Task Definition

The evaluation unit is a complete IPO legal due diligence matter. The agent receives matter-specific instructions, a workspace of company documents, external corporate databases, investigation and report-production tools, and a report date that sets the legal cutoff. It must examine the evidence, reconcile external records, and deliver an evidence-supported DOCX report (Section 3.1). A median matter contains 182 PDFs, 457 pages, and 12 investigation items. These items connect entities, facts, and timelines across files and report sections, requiring repeated cross-referencing (Appendix F.5). Patent ownership in Figure 2 is one component of this larger task.

2.2 Material Construction and Expert Annotation

Workspaces. Figure 3 summarizes the workflow from public disclosures to workspaces, expert rubrics, and evaluation. We select 63 IPO matters across 12 industries and five listing categories in the Chinese mainland and Hong Kong (Appendix B.2). Public prospectuses and disclosures supply company structures, ownership, assets, transactions, and legal matters. Experts extract fact cards and timelines, anonymize identities while retaining legally relevant relationships, and generate supporting documents from templates using a 14-category checklist. At least two legal experts review each workspace, and a third checks quality, before documents are rendered as scans (Appendix B.1).

The same process constructs external corporate records covering registration, industrial intellectual property, copyright, domain names, judicial records, credit information, and leases and procurement. Controlled document defects and database distractors reproduce the need to reconcile evidence and distinguish entities. Distractors include similar names, altered case descriptions, and expired records. The workspaces contain 12,151 PDFs and 30,243 pages; 99.0% of PDFs have no text layer. Across the benchmark, seven database domains span 23 table types; availability varies by matter. The external records include 1,963 distractors, 62.0% involving identical or similar entity names that agents must tell apart (Figure 4; Table 5; Appendix C.1).

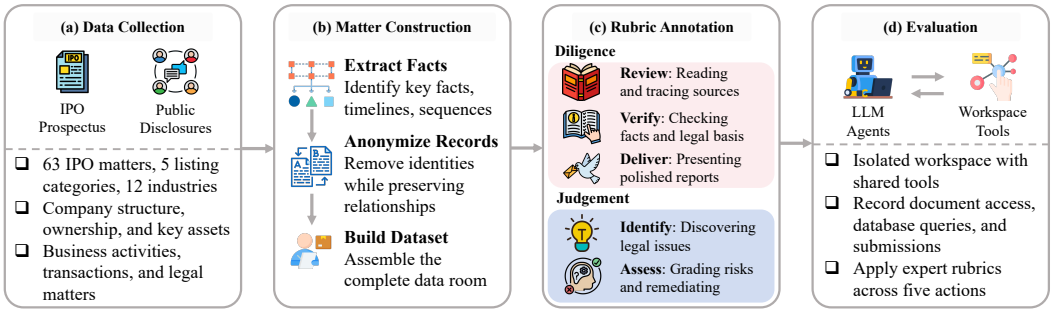
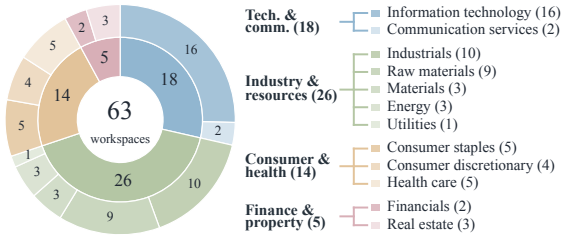


Figure 3: **Benchmark construction pipeline.** (a) Select IPO matters and collect public disclosures. (b) Extract facts, anonymize entities, and construct workspace documents and external records. (c) Legal experts develop evidence-linked rubrics; rubrics are organized into five actions for analysis. (d) Agents investigate through professional tools and submit reports for independent evaluation.

(a) Industry distribution



(b) Workspace statistics

Metric	Total	Mean	Min.	Max.
PDFs	12,151	192.87	119	311
Pages	30,243	480.05	311	1,493

(c) Scoring-rubric statistics

Action	Count	Share	Mean
Review	1,699	10.92%	26.97
Verify	5,435	34.94%	86.27
Deliver	2,328	14.96%	36.95
Identify	4,224	27.15%	67.05
Assess	1,871	12.03%	29.70
Total	**15,557**	**100.00%**	**246.94**

Figure 4: **Benchmark characteristics.** (a) Industry distribution of 63 matters: four groups and twelve recorded industries; numbers count matters. (b) Per-workspace file and page statistics. (c) Counts, shares, and mean counts per matter for the 15,557 scoring rubrics. Means are per workspace.

Rubric construction. Legal experts develop rubrics specifying a requirement, its supporting evidence, a pass/fail decision rule, and a signed weight. A second legal expert reviews the annotations; 37 legal experts contributed overall (Appendix B.4). Each rubric specifies explicit, independently verifiable pass/fail conditions, including any facts that must hold jointly. Partial mention receives no partial positive credit. Repeated error occurrences are individually assessed and counted under the rubric's trigger rule, so each error carries its own penalty.

The evaluation uses **15,557 scoring rubrics: 6,467 positive and 9,090 negative**. Appendix B.3 details their structure and distribution. For analysis, we organize eligible rubrics into five actions, assigning each to exactly one action. **Review** checks that required sources are read and cited as supporting evidence. **Verify** checks for factual, legal, and consistency errors. **Deliver** checks that the report is complete, with substantive conclusions and professional presentation. These form diligence, the requirements of reliable investigation and reporting. **Identify** rewards establishing each legal issue with its required facts, legal grounds, and their application. **Assess** rewards grading the significance of each risk and proposing remedies suited to the facts. These form judgement, the professional analysis a report delivers.

Weights have magnitudes 3, 7, or 10, reflecting expert-assigned importance. Errors require legal professionals to revisit evidence and correct conclusions, motivating penalties alongside rewards. Figure 5 pairs rubrics with contrasting report excerpts. Naming supporting files makes the evidence traceable, while the illustrated distractor adoption receives a recorded three-point penalty. Identifying co-ownership risk earns three points only when all four required facts appear in the risk analysis; mentioning shared patents and transfer restrictions without the total patent count earns none. An appropriate co-ownership remedy earns three. These examples show how binary decisions separate reliable evidence work from substantive analysis.

Company A: intellectual-property section

Hong Kong IPO of an LLM developer; 18 registered patents, four co-owned with University X.

Diligence

Review Name the supporting files.

Cites IP certificates and database records without naming a supporting file. (✗ No specific supporting file)	Names the *Joint R&D Agreement.pdf* and patent and copyright database tables. (✓ Traceable supporting evidence)

Verify Exclude non-issuer patents.

Lists ZL2022XXX456.7 as issuer-owned, while marking its validity as unverified. (✗ Incorrect patent attribution)	Excludes this record from the issuer's patent portfolio. (✓ Non-issuer record excluded)

Deliver State a substantive IP conclusion.

Reports only that document and database identifiers match. (✗ No substantive IP conclusion)	Concludes: 18 valid patents, four co-owned; IP protection covers core LLM technologies. (✓ Ownership, validity, and coverage)

Judgement

Identify Establish co-ownership risk with all four facts.

Four patents co-owned with University X; transfers require co-owner consent. (✗ Portfolio total missing from risk analysis)	18 patents; four co-owned with University X; co-owner consent restricts transfers and exclusive licensing. (✓ All four facts and the associated risk)

Assess Remedy the co-ownership restrictions.

Recommends repaying debt and releasing patent pledges. (✗ Remedy addresses a different issue)	Confirm use and revenue-sharing arrangements with University X; obtain an exclusive license if needed. (✓ Remedy addresses co-ownership)

Figure 5: **Rubric assessment of the intellectual-property section.** Each action contrasts a weaker excerpt (left) with a stronger excerpt (right). Appendix E.3.2 presents the corresponding model outputs, scores, and failure analysis; Appendix F provides the workspace inventory and rubric details.

3 Experimental Setup and Evaluation

3.1 Workspace, Protocol, and Models

Each agent receives matter-specific LDD task instructions, shared investigation requirements, tool guidance, and a report date that sets the legal cutoff. The ALawBench engine provides a workspace for investigation and report production. Rubrics, key-document checklists, and distractor provenance are reserved for evaluation and hidden from the agent. An independent rubric grader, separate from the evaluated agent, determines the benchmark score.

The standard interface comprises seven general-purpose tools and eleven Model Context Protocol (MCP) tools for workspace investigation, corporate and legal research, and report production. These include OfficeCLI's `officecli` and the commercial PKULaw service's `search_article` and `get_article`. The harness collects report files after each investigation-and-submission cycle. Appendix C.1 lists every tool and the operation it performs in the workspace.

We evaluate 21 frontier models from 12 model families. Appendix C.2 specifies the task profile, and Appendix C.2.1 lists their exact checkpoints.

Submitted reports are graded by Gemini 3.8 Flash against expert-authored rubrics. Against a three-expert majority on 3,321 rubric decisions, the grader achieves Cohen's $\kappa = 0.72$ (95% confidence interval 0.69–0.76) and accuracy 0.93. Appendices C.3 and C.5 detail scoring and agreement.

3.2 METRICS

Net score rate. For run r on matter c, let $\mathcal{P}_c$ and $\mathcal{N}_c$ be the eligible positive and negative rubrics, w_j the weight magnitude, $s_{rj} \in \{0, 1\}$ positive satisfaction, and $t_{rj} \in \mathbb{N}_0$ the negative trigger count. With $P_c = \sum_{j \in \mathcal{P}_c} w_j$, we compute

$$S_r = 100 \frac{\sum_{j \in \mathcal{P}_c} w_j s_{rj} - \sum_{j \in \mathcal{N}_c} w_j t_{rj}}{P_c}. \quad (1)$$

Matter budgets range from 207 to 556 points. Model scores average all 63 matter-level scores. Appendix C.3.2 specifies run handling. A negative rate means the weighted penalties for a report's defects outweigh the credit for its analysis.

Action contributions. Because all actions share the matter denominator P_c, the five action contributions sum to the net score rate. Review, Verify, and Deliver sum to the diligence total; Identify and Assess sum to the judgement total. We report contributions in percentage points. Appendix C.4.1 gives the aggregation formulas for each action and dimension.

Source-access diagnostics. For source-access analysis, we match each scored report to its tool-access records and compute key-document coverage as accessed key files divided by the expert-designated checklist size. We group reports by coverage, average each action contribution within each group, and select reports with coverage of at least 95%. Key files comprise 77.2% of workspace PDFs. Recorded access bounds the evidence an agent could have used, so shortfalls at high coverage cannot be attributed to missing evidence. Risk-identification and remediation satisfaction rates give the share of eligible positive rubrics in each category that a report satisfies, averaged across reports. Appendix C.6 defines both rubric categories and the averaging.

4 RESULTS AND DISCUSSION

4.1 CAN AGENTS DELIVER ISSUABLE REPORTS? (RQ1)

Figure 1 ranks models by net score rate. GPT 6 Astra leads at 40.39%, followed by Qwen 3.8 Max at 27.32% and GLM 5.3 at 27.11%, and 12 of 21 models score below zero. Net scores span 74.89 points, so the benchmark separates agents widely. The leader earns 50.53% of the positive budget and loses 10.15 percentage points to penalties, so even the strongest agent leaves about half of the creditable analysis unmet. An issuable report would need to earn most of the available credit while incurring few penalties, and no agent approaches this standard. For the 12 models below zero, Review and Deliver contribute -26.2 points on average, losses that stronger agents largely recover, whereas Verify losses of -5.1 persist at the top (Section 4.2). Table 2 gives the action contributions, and Appendix E.2 reports results by listing category and industry.

4.2 WHICH ACTIONS LIMIT AGENTS? (RQ2)

To locate the actions behind this gap, Figure 6 plots each action contribution against the net score rate. An action that limits agents improves little as overall performance rises.

Diligence. Review and Deliver contributions rise with overall performance (Spearman $\rho = 0.90$ and 0.96). GPT 6 Astra loses 0.07 points to Review and gains 0.17 from Deliver, so the leader nearly eliminates losses from unread documents and incomplete reports. Verify behaves differently. Its contributions stay between -1.8 and -8.5 points and show little association with overall performance ($\rho = 0.13$). Qwen 3.8 Max ranks second overall yet has the largest verification loss, -8.51 points, twice that of GPT 6 Astra, and Verify is the largest remaining diligence loss for each of the four leading agents. Because Verify counts every error occurrence, it measures the correction burden a reviewing lawyer faces, and stronger agents do not reduce it. Verify is also nearly unrelated to Review and Deliver ($\rho = 0.20$ and 0.11), so reading more and reporting more completely do not reduce errors. The observed errors include citations to files absent from the workspace, misattributed entities, and misleading records adopted as facts, and heavy tool use can coexist with them (Appendix E.3.1). Within diligence, the limit is verification rather than coverage or completeness.

Table 2: **Action contributions to the net score rate.** Contributions are in percentage points (pp): Review + Verify + Deliver = Diligence Total; Identify + Assess = Judgement Total; Diligence Total + Judgement Total = Net. Bold marks column leaders; shading indicates higher totals.

Model	Diligence				Judgement			Net (%)
	Review	Verify	Deliver	Total	Identify	Assess	Total	
GPT 6 Astra	**-0.07**	-4.23	**0.17**	**-4.13**	**32.13**	**12.38**	**44.51**	**40.39**
Qwen 3.8 Max	-3.68	-8.51	-2.75	-14.94	30.83	11.42	42.25	27.32
GLM 5.3	-2.89	-4.60	-2.53	-10.02	25.88	11.25	37.13	27.11
Kimi K3	-2.95	-3.28	-2.19	-8.42	23.97	11.35	35.31	26.89
Claude Opus 5	-8.88	-5.04	-5.67	-19.59	26.53	11.41	37.94	18.35
Grok 4.6	-8.44	-3.29	-7.57	-19.30	19.86	9.24	29.09	9.80
GPT 5.6 Sol	-5.23	-3.31	-7.74	-16.28	14.07	8.82	22.89	6.62
Claude Fable 5	-9.94	-3.52	-8.01	-21.48	12.79	9.27	22.06	0.58
DeepSeek V4 Flash	-11.78	-5.83	-10.01	-27.62	18.25	9.73	27.99	0.36
GPT 5.5	-8.61	-3.23	-11.15	-23.00	14.18	8.04	22.21	-0.78
Grok 4.5	-9.90	-3.68	-10.53	-24.11	10.86	8.50	19.36	-4.75
MiniMax M3	-8.78	-8.28	-11.07	-28.14	14.45	7.44	21.89	-6.25
GLM 5.2	-12.01	-4.06	-11.17	-27.24	9.98	6.70	16.69	-10.55
Gemini 3.6 Flash	-8.71	-3.41	-10.06	-22.18	6.31	4.45	10.76	-11.42
Doubao Seed 2.1 Pro	-12.80	-6.63	-14.15	-33.59	9.67	6.90	16.58	-17.01
DeepSeek V4 Pro	-14.79	-3.92	-16.19	-34.90	9.12	6.63	15.75	-19.15
Qwen 3.7 Max	-14.29	-7.32	-14.49	-36.10	8.67	6.98	15.65	-20.45
Gemini 3.7 Flash	-16.00	**-1.79**	-16.77	-34.56	9.21	3.43	12.65	-21.92
HY3	-13.78	-5.62	-14.80	-34.20	6.24	5.69	11.93	-22.27
Gemini 3.1 Pro	-13.28	-4.39	-15.42	-33.09	5.41	4.53	9.93	-23.16
MiMo V2.5 Pro	-16.59	-8.28	-19.34	-44.20	5.67	4.03	9.70	-34.50

Judgement. Identify contributions rise steeply, from 5.4 to 32.1 points ($\rho = 0.95$). Assess contributions also rise with overall performance ($\rho = 0.94$), but only from 3.4 to 12.4 points, a third of the 26.7-point rise in Identify. Among the five leading agents, Identify spans 8.2 points, from 24.0 to 32.1, while Assess spans only 1.1, from 11.3 to 12.4. No action varies less among these leaders. From Claude Opus 5 to GPT 6 Astra, the net score rate rises by 22.0 points while Assess rises by 1.0. The gains that separate the strongest agents therefore come from Identify, Review, and Deliver rather than from Assess. Assess rubrics require weighing the significance of each risk and proposing remedies suited to the facts, the standard an issuing lawyer must meet. Stronger agents therefore find more risks without assessing them better, and within judgement the limit is assessment.

4.3 Does Evidence Access Remove These Limits? (RQ3)

If agents verify poorly or assess superficially because they miss key records, both limits should shrink as agents read more of the key documents. Figure 7 tests this by grouping all reports by key-document coverage. Access does help most actions. From the lowest to the highest coverage group, Review, Deliver, and Identify improve by 10.25 to 13.22 points, and risk-identification satisfaction rises from 13.60% below 50% coverage. The two limits barely move. Verify stays within 1.1 points across the four groups, with a report-level correlation of $\rho = 0.00$ with coverage, and Assess gains only 3.50 points, a quarter of the Identify gain.

Near-complete access leaves both limits in place. Reports that access at least 95% of key documents, with a mean coverage of 99.41%, satisfy only 28.03% of risk-identification requirements and 54.73% of remediation requirements, and risk identification levels off near 28% once coverage exceeds 90%. Reports that access every key document still lose 4.35 points to Verify and earn 10.53 from Assess. The pattern also holds at the model level. GPT 6 Astra exceeds 95% coverage on all 63 matters yet loses 4.23 points to Verify. Among models with at least 20 reports above 95% coverage, Qwen 3.8 Max loses the most to Verify, at -9.42 points, while Assess ranges only from 8.1 to 12.4.

The limits therefore lie in checking evidence and acting on it rather than in finding it. Better retrieval would help agents find risks, but it would not by itself make them check records or propose remedies that fit the facts. Pairing each report with its access record is what makes this separation possible.

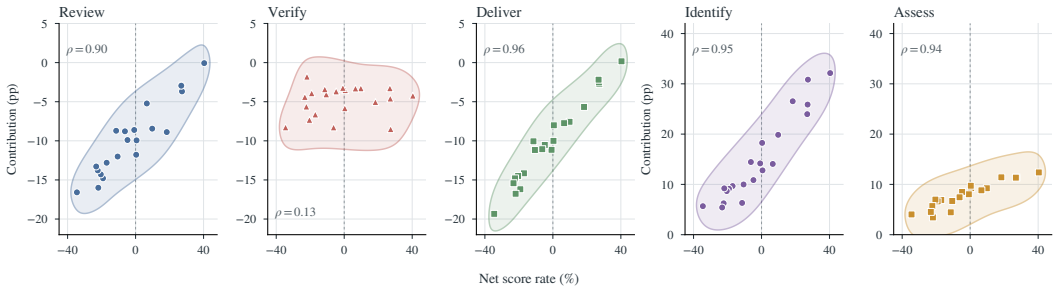


Figure 6: **Action contributions against net score rate.** Each point is one of the 21 models, and shaded regions enclose each action's points. Spearman ρ is computed across models.

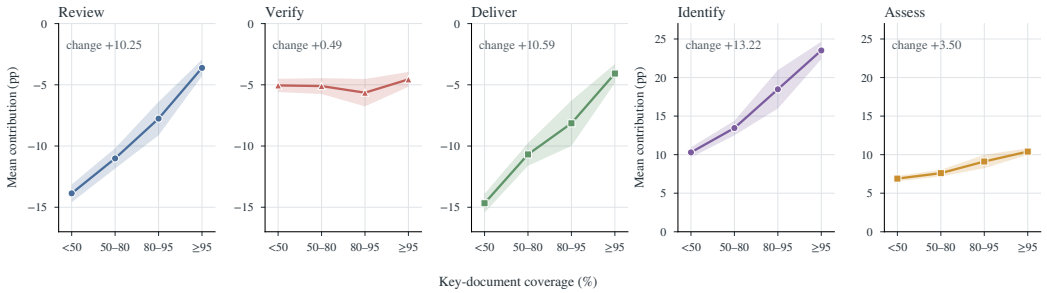


Figure 7: **Action contributions by key-document coverage.** Each scored report is assigned to one coverage group. Lines show mean contributions in percentage points, bands show 95% confidence intervals across reports, and change is the highest group minus the lowest. Table 15 gives the values.

4.4 An Illustrative Failure

The patent example in Figure 5 shows both limits on one matter. Kimi K3 qualifies a non-issuer patent as unverified yet still includes it in the issuer's portfolio, a Verify error. HY3 recommends releasing patent pledges instead of addressing co-ownership restrictions, so its remedy misses the actual risk. The two limits thus appear within one issue analysis, one in the facts the report asserts and the other in the response it recommends. By contrast, MiMo V2.5 Pro names the joint-research agreement as evidence. Appendix E.3.2 attributes each excerpt and score to its model, and Appendix E.3.3 shows a second Verify failure, in which DeepSeek V4 Flash reports a conflicting public record as an additional case without checking it against the workspace.

5 Related Work

Legal benchmarks. LegalBench, LEXam, LexGLUE, and LawBench assess legal knowledge, reasoning, and language understanding (Guha et al., 2023; Fan et al., 2025; Chalkidis et al., 2022; Fei et al., 2023). BenGER evaluates subsumption on fixed German-law tasks (Nagl et al., 2026); contract benchmarks cover review, interpretation, and inference (Hendrycks et al., 2021; Wang et al., 2023; Koreeda & Manning, 2021). CLERC studies retrieval and grounded analysis (Hou et al., 2025); JuDGE generates Chinese court judgments from supplied facts (Su et al., 2025). LegalAgentBench evaluates tool use, DLawBench studies client dialogue and fact elicitation, and PRBench-Legal scores advice and drafting from fixed histories (Li et al., 2025; Zhang et al., 2026; Akyürek et al., 2025). LongJud-Bench connects Chinese civil-litigation stages; J1-ENVS and LawFlow study procedural action and legal workflows (Zuo et al., 2026; Jia et al., 2025; Das et al., 2025). PLawBench uses expert legal rubrics, while BigLaw Bench evaluates drafting and review with positive and negative scoring (Shi et al., 2026; Harvey Team, 2024).

The gap is *end-to-end diligence from raw records to an issuable report*. Most legal benchmarks hand the relevant facts to the model, and connected workflows target litigation or adjudication. Harvey

LAB's M&A diligence extension and Legora BAR's longest tasks reach data-room scale, but they grade intermediate drafts that a lawyer still reviews (Pereyra, 2026; Grupen et al., 2026a; Lauritzen et al., 2026). ALawBench closes this gap. Agents must find every fact themselves in scanned documents and in external databases seeded with distractors, and we grade the final report against the standard of a document a firm would issue (Table 1).

Long-horizon agent evaluation. TheAgentCompany, WorkBench, WorkArena, and Odyssey-Bench evaluate workplace operations (Xu et al., 2024; Styles et al., 2024; Drouin et al., 2024; Wang et al., 2025); WorkspaceBench and SpreadsheetBench cover multi-file and spreadsheet tasks (Tang et al., 2026; Ma et al., 2024b). APEX-Agents, JobBench, OccuBench, and GDPval broaden professional coverage (Vidgen et al., 2026; Li et al., 2026; Hu et al., 2026; Patwardhan et al., 2025), while PaperBench and DeepResearch Bench assess research outputs (Starace et al., 2025; Du et al., 2025). Domain-specific benchmarks cover finance (Bigeard et al., 2025; Lau et al., 2026; Wang et al., 2026a; Yen et al., 2026; Wang et al., 2026b), clinical workflows (Jiang et al., 2025), and software engineering (Deng et al., 2025). AgentBoard measures intermediate progress, TRAIL localizes trajectory errors, and LAB relates reading and actions to performance (Ma et al., 2024a; Deshpande et al., 2025; Grupen et al., 2026b). To score professional outputs, HealthBench uses positive and negative rubrics written by physicians, ResearchRubrics evaluates research reports, and ALCE assesses text and citations (Arora et al., 2025; Sharma et al., 2025; Gao et al., 2023). Meta-evaluations examine LLM judges for open-ended responses, rubric decisions, and multi-step agent trajectories (Zheng et al., 2023; Pan et al., 2026; Lù et al., 2025).

The gap is *diagnosing which capability limits an agent*. Task-completion scores report whether an agent succeeds, and step-level error localization reports where a trajectory breaks, but neither separates failing to find evidence from failing to check it or to reason from it. ALawBench provides this diagnosis. It decomposes every report score into five action contributions and pairs every report with its document-access record, so each shortfall is traced to a specific action and tested against what the agent actually read. It shows that verification and assessment stay weak even at near-complete access to the key documents of each matter.

6 Conclusion

ALawBench evaluates IPO legal due diligence end to end, from investigating scanned records and external databases to delivering an issuable report, and thus tests whether agents can take on comprehensive and complex legal work. Even the best of 21 frontier agents reaches 40.39%, and 12 agents score below zero, so the task remains far from solved. Review, delivery, and risk identification improve with capability, but verification errors barely decline and assessment improves only marginally, even when agents read nearly all key documents. Across all 21 agents, Verify stays between -1.8 and -8.5 points and Assess never exceeds 12.4. The stronger agents read more, report more completely, and find more risks, but they neither check records more reliably nor propose remedies that fit the facts better. ALawBench makes both limits measurable by scoring each action in every matter and pairing every report with its document-access record. Professional legal work rests on records that are incomplete, inconsistent, or in conflict, so agents must verify each fact across records before they enter a report. Clients and regulators also need proportionate assessments and workable remedies, which longer contexts or broader retrieval on their own are unlikely to supply. Frontier agents cannot yet deliver issuable IPO legal due diligence reports, as their diligence falls short at verification and their judgement stalls at assessment.

7 Limitations and Future Work

ALawBench covers matters under the legal regimes of the Chinese mainland and Hong Kong, with all workspaces and instructions in Chinese, so results may differ for other jurisdictions and languages. Its workspaces and rubrics are built and reviewed by legal experts, which makes each matter costly to produce. The benchmark also focuses on IPO due diligence, one of many transactional settings in legal practice. Future work can extend ALawBench to other languages and jurisdictions, and to private equity and venture capital investments, mergers and acquisitions, and other transactions, reusing the same construction pipeline and action-based rubrics.

AI USE STATEMENT

LLMs assisted legal experts in generating supporting documents from templates and facts extracted from public disclosures. At least two legal experts reviewed every constructed workspace, and a third expert performed quality control. LLMs also serve as the evaluation judge, and we report LLM-judge agreement with human annotators (Appendix C.5). In addition, LLMs were used to edit the project code and polish the writing. All AI-generated content was reviewed and revised by legal experts at each construction stage, and the authors take full responsibility for the final content of this work.

ETHICS STATEMENT

The benchmark adapts real IPO matters from publicly available filings into constructed workspaces. Selected identifying details are replaced consistently across documents while preserving legally relevant relationships and dates (Appendix B.1). Legal experts were compensated for construction, annotation, review, and grader calibration, with roles and compensation detailed in Appendix B.4. The materials and generated reports are research artifacts and do not constitute legal advice.

REPRODUCIBILITY STATEMENT

We describe the benchmark construction pipeline in Section 2.2, with full construction guidelines and de-identification rules in Appendices B and D. The evaluation protocol and agent scaffolds are detailed in Section 3 and Appendix C. We use each model's default reasoning effort. The LLM-judge prompt is provided in Appendix G.2, and the exact model checkpoints evaluated are listed in Appendix C.2.1. The human–judge agreement study is described in Appendix C.5. Anonymized supplementary material is available at Anonymous repository.

REFERENCES

Andrew Abbott. *The system of professions: An essay on the division of expert labor*. University of Chicago Press, 2014.

Afra Feyza Akyürek, Advait Gosai, Chen Bo Calvin Zhang, Vipul Gupta, Jaehwan Jeong, Anisha Gunjal, Tahseen Rabbani, Marian Mazzone, David Randolph, Mohammad Mahmoudi Meymand, Gurshaan Chattha, Paula Rodriguez, Diego Mares, Pavit Singh, Michael Liu, Subodh Chawla, Pete Cline, Lucy Ogaz, Ernesto Hernandez, Zihao Wang, Pavi Bhatter, Marcos Ayestaran, Bing Liu, and Yunzhong He. PRBench: Large-scale expert rubrics for evaluating high-stakes professional reasoning. *arXiv*, abs/2511.11562, 2025. URL https://arxiv.org/abs/2511.11562.

Anthropic. Claude Fable 5 and Claude Mythos 5, 2026a. URL https://www.anthropic.com/news/claude-fable-5-mythos-5. Accessed: 2026-09-22.

Anthropic. Introducing Claude Opus 5, 2026b. URL https://www.anthropic.com/news/claude-opus-5. Accessed: 2026-09-22.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health. *arXiv*, abs/2505.08775, 2025. URL https://arxiv.org/abs/2505.08775.

Randolph P. Beatty and Ivo Welch. Issuer expenses and legal liability in initial public offerings. *Journal of Law and Economics*, 39(2):545–602, 1996. doi: 10.1086/467359.

Beijing Stock Exchange. Rules for reviewing public offerings of shares to unspecified qualified investors and listing, April 2024. URL https://www.bse.cn/fxrz_list/200021750.html. Announcement No. 26 of 2024, April 30.

Antoine Bigeard, Langston Nashold, Rayan Krishnan, and Shirley Wu. Finance agent benchmark: Benchmarking llms on real-world financial research tasks. *arXiv*, abs/2508.00828, 2025. URL https://arxiv.org/abs/2508.00828.

Katerina Bohle Carbonell, Renée E. Stalmeijer, Karen D. Könings, Mien Segers, and Jeroen J.G. van Merriënboer. How experts deal with novel situations: A review of adaptive expertise. *Educational Research Review*, 12:14–29, 2014. ISSN 1747-938X. doi: https://doi.org/10.1016/j.edurev.2014.03.001. URL https://www.sciencedirect.com/science/article/pii/S1747938X14000116.

ByteDance Seed Team. Seed2.1 Officially Released: Advancing AI Productivity, 2026. URL https://seed.bytedance.com/en/blog/seed2-1-officially-released-advancing-ai-productivity. Accessed: 2026-09-22.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. LexGLUE: A benchmark dataset for legal language understanding in English. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4310–4330, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.297. URL https://aclanthology.org/2022.acl-long.297/.

China Securities Regulatory Commission. Standards for sponsor due diligence, May 2022. URL https://www.csrc.gov.cn/csrc/c101954/c3048267/content.shtml. Announcement No. 36 of 2022, May 27.

China Securities Regulatory Commission, Ministry of Justice of the People's Republic of China, and All China Lawyers Association. Regulatory rules application guidelines—legal no. 2: Practice rules for law firms engaged in initial public offering and listing legal services, 2022. URL https://www.sse.com.cn/lawandrules/regulations/guidelines/c/c_20230301_5717199.shtml.

Debarati Das, Khanh Chi Le, Ritik Sachin Parkar, Karin de Langis, Brendan Madson, Charles Berryman, Robin M. Willis, Daniel H. Moses, Brett H. McDonnell, Dan Schwarcz, and Dongyeop Kang. Lawflow: Collecting and simulating lawyers' thought processes on business formation case studies. *arXiv*, abs/2504.18942, 2025. URL https://arxiv.org/abs/2504.18942.

DeepSeek-AI, Anyi Xu, Bang Lin, Bing Xue, Bing-Li Wang, Bin Xu, Bo Wu, Bowei Zhang, Chao Lin, Chengyao Dong, Chen Ling, Cheng-Da Lu, Chen Zhao, Chengqi Deng, Cheng Hou, Chen Xu, Chen-Ze Shao, Chong Ruan, Conner Sun, Da-Mai Dai, Daya Guo, De-Bin Yang, Deli Chen, Dong-Hui Li, Dong-Li Ji, Erhang Li, Fang-Chen Wei, Fangyun Lin, Fang Yuan, Fei Xia, Fucong Dai, Guangbo Hao, Guanting Chen, Guo Cao, Guo-Hui Meng, Guowei Li, Han Yu, Han Zhang, Hanwei Xu, Hao Li, Hao Liang, Haoling Zhang, Haoming Luo, Haoran Wei, Hao Yuan, Haowei Zhang, Hao-Wen Luo, Hao Chen, Haozhe Ji, Heng Zhang, Honghui Ding, Hongxuan Tang, Huanqi Cao, Hua-Zuo Gao, Hui Qu, Hui Zeng, J. Yang, J.-Q. Zhu, Jianming Luo, Jia Song, Jia Yu, Jialiang Huang, Jia-Lu Cai, Jian Liang, Jiangting Zhou, Jiasheng Ye, Jia-Shi Li, Jia-Xin Xu, Jiewen Hu, Jieyu Yang, Jin Chen, Jin Yan, Jing-Chang Chen, Jing Zhou, Jing Xiang, Jingyang Yuan, Jing Cheng, Jing-Zi Zhou, Jinhua Zhu, Jiping Yu, Joseph Sun, Junliang Ran, Jun Jiang, Jun-Jie Qiu, Jun-Long Li, Junming Zheng, Jun-Mei Song, Kai Dong, Kaige Gao, Kang Guan, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Wang, Leyi Xia, Li Zhang, Liang Zhao, Lihua Guo, Lin Luo, Lin Ma, Linyan Zhu, Li-Tong Wang, Li-Yu Cai, Liyue Zhang, Long-Hao Chen, Mingxi Di, My Xu, Max Mei, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Mingxu Zhou, Min Han, Ning Wang, Pan Huang, Pan Wang, Peixin Cong, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qing-Yang Li, Qinyu Chen, Qiushi Du, Qi Jiang, Rui Tian, Rui qin Xu, Rui-Jie Lu, Ruiling Xu, Ruiqi Ge, Rui-Song Zhang, Ruizhe Pan, Runji Wang, Runqian Chen, Runqiu Yin, Runxin Xu, Ruo-Han Shen, Ruoyu Zhang, Ruyi Chen, Sh. L. Liu, Shanghao Lu, Shangmian Sun, Shang-Yan Zhou, Shan-Shan Chen, Shaofei Cai, Shao Nie, Shao-Ping Wu, Shao-Yuan Chen, Shengding Hu, Sheng Liu, Shiqiang Hu, Shi-Rong Ma, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, Shuying Yu, Songyang Zhou, Tao Ni, Tao Yun, Tian Jin, Tian-Hong Pei, Tian Ye, Tian-Le Lin, Tian Ji, Tianyi Cui, Tianyuan Yue, Tingting Yu, Tu-Tu Wang, W. Y. Zhang, WL Xiao, Wangding Zeng, Wei An, Weilin Zhao, Wen

Liu, Wen-Feng Liang, Wenjie Pang, Wenjing Luo, Wenjin Yao, Wenjun Gao, Wenkai Yang, Wenlve Huang, Wenqing Hou, Wentao Zhang, Wen-Ting Ma, Xi Gao, Xiang He, Xiangwen Wang, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xing-Chen Liu, Xing-Kai Yu, Xing-You Li, Xin-Yu Yang, Xinyu Zhang, Xu Chen, Xuanyu Wang, Xuecheng Su, Xue-Yin Chen, Xuheng Lin, Xu Fu, Y.-C. Yan, YQ Wang, Yw Ma, Yanfen Luo, Yang Zhang, Yan-Hong Xu, Yanru Ma, Yanwen Huang, Yao Li, Yao Xu, Yao Zhao, Yao-Feng Sun, Yao-Hui Wang, Yi Qian, Yingxia Shao, Yi Yu, Yichao Zhang, Yifan Ding, Yifan Shi, Yi-Jia Wu, Yi-Yu Xiong, Yi Ma, Ying He, Ying Tang, Ying Zhou, Yingjia Luo, Yinmin Zhong, Yishi Piao, Yisong Wang, Yixiang Zhang, Yixiao Chen, Yi-Xuan Tan, Yixuan Wei, Yiyang Ma, Yiyuan Liu, Yong-Lu Yang, Yong-Qiang Guo, Yong-Tong Wu, Yu Wu, Yukun Li, Yuan Cheng, Yuan Ou, Yuanfan Xu, Yuanhao Li, Yuduan Wang, Yuehan Yang, Yue Xu, Yuhan Wu, Yu Meng, Yu Zou, Yukun Zha, Yunfan Xiong, Yu-Peng Chen, Yuping Lin, Yu Cao, Yuqian Wang, Yushun Zhang, Yu-Ting Yan, Yu-Tong Lin, Yu-Xian Gu, Yu-Wei Luo, Yu-Mei You, Yuxuan Liu, Yuxuan Zhou, Yuyang Zhou, Yuzhen Huang, Zhanghua Wu, Ze Wang, Ze Zhao, Zehui Ren, Ze-Kai Zhang, Zhangli Sha, Zhe Fu, Zhe Ju, Zhean Xu, Zhen-Da Xie, Zhengyan Zhang, Zhe Gao, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhi-Xiang Huang, Zhi-Xuan Chen, Zhiyu Wu, Zhizhou Ren, Zhongyu Wu, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihua Qu, Zihui Gu, Zi jia Zhu, Zi-Long Li, Zipeng Zhang, Zi-Wei Xie, Ziyi Gao, Zi-Yi Wan, Zi-Zheng Pan, and Zong xiang Yao. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv*, abs/2606.19348, 2026. URL https://arxiv.org/abs/2606.19348.

Xiang Deng, Jeff Da, Edwin Pan, Yan He, Charles Ide, Kanak Garg, Niklas Lauffer, Andrew Park, Nitin Pasari, Chetan Rane, Karmini Sampath, Maya Krishnan, Srivatsa Kundurthy, Sean M. Hendryx, Zifan Wang, Chen Bo Calvin Zhang, Noah Jacobson, Bing Liu, and Brad Kenstler. Swebench pro: Can ai agents solve long-horizon software engineering tasks? *arXiv*, abs/2509.16941, 2025. URL https://arxiv.org/abs/2509.16941.

Darshan Deshpande, Varun Gangal, Hersh Mehta, Jitin Krishnan, Anand Kannappan, and Rebecca Qian. Trail: Trace reasoning and agentic issue localization. *arXiv*, abs/2505.08638, 2025. URL https://arxiv.org/abs/2505.08638.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam Hadj Laradji, Manuel Del Verme, Tom Marty, L'eo Boisvert, Megh Thakkar, Quentin Cappart, David Vázquez, Nicolas Chapados, and Alexandre Lacoste. Workarena: How capable are web agents at solving common knowledge work tasks? *arXiv*, abs/2403.07718, 2024. URL https://arxiv.org/abs/2403.07718.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, et al. DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents. *arXiv preprint arXiv:2506.11763*, 2025. URL https://arxiv.org/abs/2506.11763.

Yu Fan, Jingwei Ni, Jakob Merane, Etienne Salimbeni, Yang Tian, Yoan Hermstruwer, Yinya Huang, Mubashara Akhtar, Florian Geering, Oliver Dreyer, Daniel Brunner, Markus Leippold, Mrinmaya Sachan, Alexander Stremitzer, Christoph Engel, Elliott Ash, and Joel Niklaus. Lexam: Benchmarking legal reasoning on 340 law exams. *arXiv*, abs/2505.12864, 2025. URL https://arxiv.org/pdf/2505.12864.

Zhiwei Fei, Xiaoyu Shen, D. Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Ji-Dong Ge. Lawbench: Benchmarking legal knowledge of large language models. *arXiv*, abs/2309.16289, 2023. URL https://arxiv.org/abs/2309.16289.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Ronald J Gilson. Value creation by business lawyers: Legal skills and asset pricing. *The Yale Law Journal*, 94(2):239–313, 1984.

Google DeepMind. Gemini 3.1 Pro Model Card, 2026a. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/. Accessed: 2026-09-22.

Google DeepMind. Gemini 3.6 Flash Model Card, 2026b. URL https://deepmind.google/models/model-cards/gemini-3-6-flash/. Accessed: 2026-09-22.

Google DeepMind. Gemini 3.7 Flash Model Card, 2026c. URL https://deepmind.google/models/model-cards/gemini-3-7-flash/. Accessed: 2026-09-22.

Google DeepMind. Gemini 3.8 Flash Model Card, 2026d. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/. Accessed: 2026-09-22.

Niko Grupen, Gabe Pereyra, and Julio Pereyra. Open-Sourcing Harvey's Long Horizon Legal Agent Benchmark. Harvey research blog, 2026a. URL https://www.harvey.ai/blog/introducing-harveys-legal-agent-benchmark.

Niko Grupen, Gabe Pereyra, and Julio Pereyra. Initial Results on Legal Agent Benchmark. Harvey research blog, 2026b. URL https://www.harvey.ai/blog/legal-agent-benchmark-initial-results.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. Legalbench: a collaboratively built benchmark for measuring legal reasoning in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Kathleen Weiss Hanley and Gerard Hoberg. The information content of IPO prospectuses. *The Review of Financial Studies*, 23(7):2821–2864, 2010. doi: 10.1093/rfs/hhq024.

Harvey Team. Introducing BigLaw Bench. Harvey research blog, 2024. URL https://www.harvey.ai/blog/introducing-biglaw-bench.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. Cuad: An expert-annotated nlp dataset for legal contract review. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/6ea9ab1baa0efb9e19094440c317e21b-Paper-round1.pdf.

Abe Bohan Hou, Orion Weller, Guanghui Qin, Eugene Yang, Dawn Lawrie, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. CLERC: A dataset for U. S. legal case retrieval and retrieval-augmented analysis generation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 7913–7928, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.441. URL https://aclanthology.org/2025.findings-naacl.441/.

Xiaomeng Hu, Ying-Er Zhang, Fei Huang, Jianhong Tu, Yang Su, Liang-Hao Deng, Yuxuan Liu, Yan-Tao Liu, Dayiheng Liu, and Tsung-Yi Ho. Occubench: Evaluating ai agents on real-world professional tasks via language environment simulation. *arXiv*, abs/2604.10866, 2026. URL https://arxiv.org/abs/2604.10866.

Zheng Jia, Shengbin Yue, Wei Chen, Siyuan Wang, Yi-Dong Liu, Yun Song, and Zhongyu Wei. Ready jurist one: Benchmarking language agents for legal intelligence in dynamic environments. *arXiv*, abs/2507.04037, 2025. URL https://arxiv.org/abs/2507.04037.

Yi-Xing Jiang, Kameron Collin Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. Medagentbench: A realistic virtual ehr environment to benchmark medical llm agents. *arXiv*, abs/2501.14654, 2025. URL https://arxiv.org/abs/2501.14654.

Kimi Team, Yifan Bai, Yiping Bao, Chandru. M, Jianfeng Cai, Xin-Hao Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y. Charles, H. S. Che, Guanduo Chen, Guang-Yu Chen, Guan-Zheng Chen, Huarong Chen, Jia Chen, Jian-Long Chen, Jun Chen, Kexin Chen, Peng Chen, Ruijue Chen, Wentao Chen, Xinxian Chen, Yang Chen, Yanru Chen, Yifei Chen, Yingjian Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Dazhi Cheng, Yean Cheng, Jialei Cui, Jing Cui, An-Bang Dai, Jiaqi Deng, Haochen Ding, Rui Ding, Shaofeng Ding, Mengfan Dong, Meng xiao Dong, Yuhao Dong, Yuxin Dong, Angang Du, Chenzhuang Du, Dikang Du, Jusen Du, Yulun Du, Yu Fan, Jing Feng, Qiulin Feng, Yichen Feng, Kelin Fu, Qiang Fu, Fu-Juan Gao, Hongcheng Gao, Jingyue Gao, Tong Gao, Wei Gao, Shangyi Geng, Jie Gong, Lin Gong, Sheng Gong, Xiaocheng Gong, Qizheng Gu, Yicheng Gu, Shuhao Guan, Haiqing Guo, Shiqi Guo, Xiang Guo, Zhen Guo, Beixi Hao, Wenxing Hao, Xiaoru Hao, Dailan He, Haotian He, Lehan He, Qi He, Weiran He, Xinran He, Xin-Yi He, Yibo He, Yun He, Chao Hong, Tiange Hong, Hao-Xing Hu, Jiaxi Hu, Rui-Jia Hu, Weiming Hu, Yangyang Hu, Zhenxing Hu, Liang Hua, Jinbin Huang, Ke Huang, Ruiyuan Huang, Siying Huang, Weixiao Huang, Yan Huang, Zhengjie Huang, Zhiqi Huang, Yu-Long Hui, Chao Jia, Yu-Tong Jiang, Zhejun Jiang, Zuoyou Jiang, W.M. Jin, Xinyi Jin, Yu Jing, Huanjun Kong, Guokun Lai, Aidi Li, Cheng Li, Chengyu Li, Cong Li, Fang Li, Guanyu Li, Haoyang Li, Jia Li, Jun-Xiong Li, Lei Li, Letian Li, Lincan Li, Weihong Li, Wentao Li, Xintong Li, Yang Li, Yisheng Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zhaoxin Li, Zheming Li, Zheng-Xiao Li, Zhi-Yuan Li, Jiawei Lin, Xiaohan Lin, Yibo Lin, Zichao Lin, Ziyan Lin, Bill Liu, B. Liu, Chuangye Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yanming Liu, Yi-Bo Liu, Yi-Peng Liu, Zhengying Liu, Zhi-Heng Liu, Enzhe Lu, Haoyu Lu, Lin Lu, Ting Lu, Zhiyuan Lu, Aotian Luo, Gen Luo, Junyu Luo, Yi-Fan Luo, Bohan Lyu, Wenzhou Lyu, Shaoguang Mao, Yuan Mei, Xin Men, Mi Ni, Yixuan Niu, Siyuan Pan, Shujun Peng, Zhangyang Qi, Ruoyu Qin, ZeChao Qin, Zeyu Qin, Haiquan Qiu, Jian Qiu, Jiezhong Qiu, Bowen Qu, Yuhao Qu, Zeyu Shang, Youbo Shao, Han Shen, Jin Shi, Juanfeng Shi, Li-Na Shi, Sheng-Peng Shi, Wingchun Siu, Pengwei Song, Xiao-Xiu Song, Jianlin Su, Yu-Li Su, Zhao-Chen Su, Lin Sui, Jingsong Sun, Junyao Sun, Shaoning Sun, Shuzhen Sun, Tong Sun, Yu-Jun Sun, Yu-Chuan Tai, Chuning Tang, He-Yi Tang, Si-Si Tang, Ze-Cheng Tang, Chao Tian, Rong Tian, Yu Tian, Wei Tu, Chensi Wang, Chuang Wang, Chun jia Wang, Dinglu Wang, Feng Wang, Hailong Wang, Haiming Wang, Hao Wang, Huaqing Wang, Hui Wang, Jiayi Wang, Jing-Lei Wang, Jinhong Wang, Jiuzheng Wang, Li-Nian Wang, Shaobo Wang, Shenzhi Wang, Shuyi Wang, Siyuan Wang, Siyuan Wang, Tianfu Wang, Wenju Wang, Xingran Wang, Xin-Meng Wang, Xinyuan Wang, Xusheng Wang, Yalin Wang, Yangkun Wang, Yao Wang, Yaoyu Wang, Yejie Wang, Yi-Qin Wang, Yucheng Wang, Yuzhi Wang, Zhao-Ji Wang, Zhaowei Wang, Zhengtao Wang, Zhenhao Wang, Zhong-Sheng Wang, Zifan Wang, Chu Wei, Ming Wei, Shouxin Wei, Zichen Wen, Fan Wu, Haoning Wu, Rucong Wu, Wenhao Wu, Xiaoxue Wu, Ying-Nian Wu, Yong Wu, Yuxin Wu, Zijian Wu, Xinglang Xian, Chen Xiang, Yu-Ting Xiang, Bo Xiao, Chenjun Xiao, Xin Xiao, Jin Xie, Xiao-Ming Xie, Yifeng Xie, Zhe Xie, Bowei Xing, Yi-Ming Xiong, Bao-Xin Xu, Boyu Xu, Jiale Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Qin Xu, Shuyao Xu, Suting Xu, Tian-Tian Xu, Tianxiang Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Yezhao Xu, Yue Xu, Ziyao Xu, Hao Xue, Junjie Yan, Yaoyao Yan, Fan Yang, Guang-Fu Yang, Hao Yang, Junwei Yang, Ruo-Tong Yang, Wenjie Yang, Xiaofei Yang, Xin-Yu Yang, Yi Yang, Yiling Yang, Ying Yang, Yuchen Yang, Zhen Yang, Zhilin Yang, Ruihan Yang, Zuhao Yang, Haotian Yao, Dan Ye, Hao-Ran Ye, Wenjie Ye, Zhanbo Ye, Bohong Yin, Haoxiang Yin, Xi Yin, Chengzheng Yu, Haozhe Yu, Long Yu, Sheng Yu, Shuying Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Tong-Tian Yue, Weize Yue, Yang Yue, Dunyuan Zha, Haobing Zhan, B. Zhang, Dehao Zhang, Feichang Zhang, Hao Zhang, Haoyuan Zhang, Huanyu Zhang, Jia-Pei Zhang, Jia-Xing Zhang, Jin Zhang, Kai-Yi Zhang, Miao Zhang, Pu Zhang, Qingle Zhang, Rong Zhang, Rui Zhang, Shaoshuai Zhang, Shiyi Zhang, Xiaobin Zhang, Xiaoyun Zhang, Y. Zhang, Yang-Kun Zhang, Yeqing Zhang, Yichi Zhang, Yikun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Zijing Zhang, Bin Zhao, Chenguang Zhao, Feifan Zhao, Jing-Lin Zhao, Jinxiang Zhao, Shuai Zhao, Wen-Shuo Zhao, Xiangyu Zhao, Xuanle Zhao, Yikai Zhao, Zijia Zhao, Hao Zheng, Huabin Zheng, Ruihan Zheng, Shao Jian Zheng, Tengyang Zheng, Haofeng Zhong, Lei Zhong, Longguang Zhong, M. Zhou, Qian Zhou, Runjie Zhou, Ru Zhou, Xinyu Zhou, Yiqiao Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yangjunfeng Zhu, Yuxuan Zhu, Zhengxin Zhu, Zhuang Chen, Weiyu Zhuang, and Xinxing Zu. Kimi k3: Open frontier intelligence. *arXiv*, abs/2607.24653, 2026. URL https://arxiv.org/abs/2607.24653.

Yuta Koreeda and Christopher Manning. ContractNLI: A dataset for document-level natural language inference for contracts. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1907–1919, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.164. URL https://aclanthology.org/2021.findings-emnlp.164/.

Elaine Lau, Mark Ducker, Ronak A. Chaudhary, Hui Wen Goh, Rosemary Wei, Vaibhav Kumar, Saed Qunbar, Guram Gogia, Yi Liu, Scott Millslagle, Nasim Borazjanizadeh, Ulyana Tkachenko, Samuel Eshun Danquah, Collin Schweiker, Vijay Karumathil, Asrith Devalaraju, Varsha Sandadi, Haemi Nam, P. Naraghi Arani, R. Mark Epps, Abdullah Arif, Sahil Bhaiwala, Curtis G. Northcutt, Skyler Wang, Anish Athalye, Jonas Mueller, and Francisco Guzm'an. Bankertoolbench: Evaluating ai agents in end-to-end investment banking workflows. *arXiv*, abs/2604.11304, 2026. URL https://arxiv.org/abs/2604.11304.

Jacob Lauritzen, Emil Sjölander, Ebba Helfer, and Firas Danil. Introducing the Legora BAR. Legora Research, 2026. URL https://legora.com/bar. Updated July 28, 2026. Accessed September 21, 2026.

Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, Wuyue Wang, Yiqun Liu, and Minlie Huang. LegalAgentBench: Evaluating LLM agents in legal domain. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2322–2344, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.116. URL https://aclanthology.org/2025.acl-long.116/.

Yuetai Li, Yichen Feng, Zhangchen Xu, Zixian Ma, Kaiyuan Zheng, Fengqing Jiang, Xinghua Sun, Rulin Shao, Zichen Chen, Yue Huang, Xinyan Han, Brian Lee, Kayla Y. Xu, Shenglai Zeng, Hang Hua, Xiangliang Zhang, Basel Alomair, Ranjay Krishna, Luke S. Zettlemoyer, Pang Wei Koh, Bhaskar Ramasubramanian, Luyao Niu, Xiang Yue, and Radha Poovendran. Jobbench: Aligning agent work with human will. *arXiv*, abs/2605.26329, 2026. URL https://arxiv.org/abs/2605.26329.

Sida Liu. Client influence and the contingency of professionalism: The work of elite corporate lawyers in china. *Law & Society Review*, 40(4):751–782, 2006. doi: https://doi.org/10.1111/j.1540-5893.2006.00280.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5893.2006.00280.x.

Xing Han Lù, Amirhossein Kazemnejad, Nicholas Meade, Arkil Patel, Dongchan Shin, Alejandra Zambrano, Karolina Stańczak, Peter Shaw, Christopher Pal, and Siva Reddy. Agentrewardbench: Evaluating automatic evaluations of web agent trajectories. *arXiv*, abs/2504.08942, 2025. URL https://arxiv.org/abs/2504.08942.

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. Agentboard: An analytical evaluation board of multi-turn llm agents. *arXiv*, abs/2401.13178, 2024a. URL https://arxiv.org/abs/2401.13178.

Zeyao Ma, Bohan Zhang, Jing Zhang, Jifan Yu, Xiaokang Zhang, Xiaohan Zhang, Sijia Luo, Xi Wang, and Jie Tang. Spreadsheetbench: Towards challenging real world spreadsheet manipulation. *arXiv*, abs/2406.14991, 2024b. URL https://arxiv.org/abs/2406.14991.

MiniMaxAI. MiniMax-M3 Model Card, 2026. URL https://huggingface.co/MiniMaxAI/MiniMax-M3. Accessed: 2026-09-22.

Sebastian Nagl, A. Mayrhofer, Martin Heidebach, Aleyna Koçak, Anne Zettelmeier, Elly Breu, A. Cathleen Greiner, Sofija Milijas, and Matthias Grabmair. Benger: Benchmarking llm systems on subsumption-based legal reasoning in german law. *arXiv*, abs/2605.28183, 2026. URL https://arxiv.org/abs/2605.28183.

OpenAI. GPT-5.5 System Card, 2026a. URL https://deploymentsafety.openai.com/gpt-5-5/gpt-5-5.pdf. Accessed: 2026-09-22.

OpenAI. GPT-5.6 System Card, 2026b. URL https://deploymentsafety.openai.com/gpt-5-6/gpt-5-6.pdf. Accessed: 2026-09-22.

OpenAI. GPT-6 Astra System Card, 2026c. URL https://deploymentsafety.openai.com/gpt-6-astra. Accessed: 2026-09-22.

Tian Pan, Xuan Lin, Wenyan Yang, Qi He, Shi-Song Chen, Li Qi, Wanqing Xu, Hongwei Feng, Bo Xu, and Yanghua Xiao. Rubriceval: A rubric-level meta-evaluation benchmark for llm judges in instruction following. *arXiv*, abs/2603.25133, 2026. URL https://arxiv.org/abs/2603.25133.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Sim'on Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, Natalie S. Kim, Patrick Chao, Samuel Miserendino, Gildas Chabot, David Li, Michael Sharman, Alexandra Barr, Amelia Glaese, and Jerry Tworek. Gdpval: Evaluating ai model performance on real-world economically valuable tasks. *arXiv*, abs/2510.04374, 2025. URL https://arxiv.org/abs/2510.04374.

Julio Pereyra. Extending Legal Agent Bench to M&A Due Diligence. Harvey research blog, 2026. URL https://www.harvey.ai/blog/legal-agent-bench-m-and-a-due-diligence.

Qwen Team. Qwen3.7: The Agent Frontier, 2026a. URL https://qwen.ai/blog?id=qwen3.7. Accessed: 2026-09-22.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork, 2026b. URL https://qwen.ai/blog?id=qwen3.8. Accessed: 2026-09-22.

Daniel L. Schwartz, John D. Bransford, and David Sears. Efficiency and innovation in transfer. In Jose P. Mestre (ed.), *Transfer of Learning from a Modern Multidisciplinary Perspective*, pp. 1–51. Information Age Publishing, 2005. URL https://people.cs.vt.edu/~kafura/CS6604/Papers/Efficiency-Innovation-Transfer.pdf.

Securities and Futures Commission. Code of conduct for persons licensed by or registered with the securities and futures commission, October 2024. URL https://www.sfc.hk/-/media/TC/files/LIC/handbook/Code-of-Conduct-for-Persons-Licensed-by-or-Registered-with-the-Securities-and-Futures-Commission-Eng.pdf.

Securities and Futures Commission. Circular to Licensed Corporations Carrying Out Sponsor Work, 2026. URL https://apps.sfc.hk/edistributionWeb/api/circular/openFile?lang=EN&refNo=26EC4. Circular 26EC4, January 30.

Shanghai Stock Exchange. Stock issuance and listing review rules (april 2024 revision), April 2024. URL https://www.sse.com.cn/ipo/rules/star/c/c_20240430_10787292.shtml. Notice No. 49 of 2024, April 30.

Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, Huy Nghiem, Tahseen Rabbani, Ye Htet, Brian Jang, Sumana Basu, Aishwarya H. Balwani, Denis Peskoff, Marcos Ayestaran, Sean M. Hendryx, Brad Kenstler, and Bing Liu. Researchrubrics: A benchmark of prompts and rubrics for evaluating deep research agents. *arXiv*, abs/2511.07685, 2025. URL https://arxiv.org/abs/2511.07685.

Shenzhen Stock Exchange. Stock issuance and listing review rules (2024 revision), April 2024. URL https://docs.static.szse.cn/www/lawrules/rule/allrules/bussiness/W020240430572636488364.pdf. Notice No. 341 of 2024, April 30.

Yuzhen Shi, Huanghai Liu, Yiran HU, Song Gaojie, Xu Xinran, Yubo Ma, Tianyi Tang, Li Zhang, Qingjing Chen, Feng Di, Wenbo Lv, Weiheng Wu, Kexin Yang, Sen Yang, Wei Wang, Rongyao Shi, Qiu Yuanyang, Yuemeng Qi, Zhang Jingwen, Sui Xiaoyu, Yifan Chen, Zhang Yi, An Yang, Bowen Yu, Dayiheng Liu, Junyang Lin, Weixing Shen, Bing Zhao, Charles L. A. Clarke, and HU Wei. PLAWBENCH: A rubric-based benchmark for evaluating LLMs in real-world legal practice. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), pp. 10067–10116, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.458. URL https://aclanthology.org/2026.acl-long.458/.

Marjorie M. Shultz and Sheldon Zedeck. Predicting lawyer effectiveness: Broadening the basis for law school admission decisions. *Law & Social Inquiry*, 36(3):620–661, 2011. doi: https://doi.org/10.1111/j.1747-4469.2011.01245.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1747-4469.2011.01245.x.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. Paperbench: Evaluating ai's ability to replicate ai research. *arXiv*, abs/2504.01848, 2025. URL https://arxiv.org/abs/2504.01848.

Olly Styles, Sam Miller, Patricio Cerda-Mardini, Tanaya Guha, Victor Sanchez, and Bertie Vidgen. Workbench: a benchmark dataset for agents in a realistic workplace setting. *arXiv*, abs/2405.00823, 2024. URL https://arxiv.org/abs/2405.00823.

Weihang Su, Baoqing Yue, Qingyao Ai, Yiran Hu, Jiaqi Li, Changyue Wang, Kaiyuan Zhang, Yueyue Wu, and Yiqun Liu. Judge: Benchmarking judgment document generation for chinese legal system. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, pp. 3573–3583, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715921. doi: 10.1145/3726302.3730295. URL https://doi.org/10.1145/3726302.3730295.

Zi-Rui Tang, Xuanhe Zhou, Yu-Mou Liu, Linchun Li, Wei-Zheng Wang, Hong-Yi Huang, Jun Zhou, Jiacheng Song, Shao-Lin Yu, Jinqi Wang, Zihan Zhou, Hongyi Zhou, Yu-Ting Lv, Jinyang Li, Jiashuo Liu, Ruoyu Chen, Chun-Wei Liu, Guoliang Li, Jihua Kang, and Fan Wu. Workspace-bench 1.0: Benchmarking ai agents on workspace tasks with large-scale file dependencies. *arXiv*, abs/2605.03596, 2026. URL https://arxiv.org/abs/2605.03596.

Tencent Hy Team. Hy3 Model Card, 2026. URL https://huggingface.co/tencent/Hy3. Accessed: 2026-09-22.

The Stock Exchange of Hong Kong Limited. Main board listing rules, practice note 21: Due diligence by sponsors in respect of initial listing applications, 2026. URL https://www.hkex.com.hk/-/media/HKEX-Market/Listing/Rules-and-Guidance/Listing-Rules/Consolidated-PDFs/Main-Board-Listing-Rules/consol_mb.pdf.

Bertie Vidgen, Arjun Mann, Abby Fennelly, Jonathan Stanly, Lucas Rothman, Marcos Herrera Burstein, Julien Benchek, David Ostrofsky, Anirudh Ravichandran, Debnil Sur, Neel Venugopal, Alannah Hsia, Isaac Robinson, Calix Huang, Olivia Varones, Daniyal Khan, Michael R. Haines, Austin Bridges, Jesse Boyle, Koby Twist, Zach Richards, Chirag Mahapatra, Brendan Foody, and Osvald Nitski. Apex-agents. *arXiv*, abs/2601.14242, 2026. URL https://arxiv.org/abs/2601.14242.

Alex Wang, Georg Meinhardt, Jacob Katz, Joseph H.T. Kim, Pratyush K. Chaudhary, Chase Blagden, and Eric Xu. Bigfinancebench: A workflow-grounded benchmark for financial-research agents. *arXiv*, abs/2606.03829, 2026a. URL https://arxiv.org/abs/2606.03829.

Benye Wang, Kang Zhou, Lifan Guo, Feng Chen, and Chi Zhang. Finprobench: Evaluating financial ai agents with role-grounded rubrics derived from professional deliverables. *arXiv*, abs/2608.04077, 2026b. URL https://arxiv.org/abs/2608.04077.

Steven Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dmitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. MAUD: An expert-annotated legal NLP dataset for merger agreement understanding. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 16369–16382, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.1019. URL https://aclanthology.org/2023.emnlp-main.1019/.

Weixuan Wang, Dongge Han, Daniel Madrigal Diaz, Jin Xu, Victor Rühle, and Saravan Rajmohan. Odysseybench: Evaluating llm agents on long-horizon complex office application workflows. *arXiv*, abs/2508.09124, 2025. URL https://arxiv.org/abs/2508.09124.

xAI. Introducing Grok 4.5, 2026a. URL https://x.ai/news/grok-4-5. Accessed: 2026-09-22.

xAI. Introducing Grok 4.6, 2026b. URL https://x.ai/news/grok-4-6. Accessed: 2026-09-22.

Xiaomi MiMo Team. MiMo-V2.5-Pro Model Card, 2026. URL https://huggingface.co/XiaomiMiMo/MiMo-V2.5-Pro. Accessed: 2026-09-22.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Meng Bao, Zora Zhiruo Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Ming-Hsuan Yang, Hao Lu, Amaad Martin, Zhe Su, Leander Melroy Maben, Raj Mehta, Wayne Chi, Lawrence Keunho Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. Theagentcompany: Benchmarking llm agents on consequential real world tasks. *arXiv*, abs/2412.14161, 2024. URL https://arxiv.org/abs/2412.14161.

Thomson Yen, Julian Poeltl, Harshith Srinivas Gear, Yilin Meng, Joshua Fan, Adam Shen, Yilin Liu, Ali Bauyrzhan, Patrick Shea, Siri Du, Haoyan Liu, Daniel Guetta, and Hongseok Namkoong. Mbabench: Evaluating llm agents on end-to-end spreadsheet tasks in finance. *arXiv*, abs/2605.22664, 2026. URL https://arxiv.org/abs/2605.22664.

GLM-4.5 Team Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Cheng-Xing Xie, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, Jian Jiao, Jiaqi Guo, Jingsen Wang, Jin-Cheng Du, Jinzhu Wu, Kedong Wang, Lei Li, Linxi Jim Fan, Lucen Zhong, Mingdao Liu, Mingming Zhao, Pengfan Du, Qian Dong, Rui Lu, Shuang-li, Shulin Cao, Song Liu, Ting Jiang, Xiaodong Chen, Xiaohan Zhang, Xuancheng Huang, Xue Dong, Yabo Xu, Yao Wei, Yifan An, Yilin Niu, Yitong Zhu, Yuan Wen, Yukuo Cen, Yu Bai, Zhongpei Qiao, Zihan Wang, Zi-Kang Wang, Zilin Zhu, Ziqian Liu, Zixuan Li, Bojie Wang, Bosi Wen, Can Huang, Changpeng Cai, Chao Yu, Chen Li, Chen Huang, Chen Wei-Hua Hu, Chenhui Zhang, Chenzheng Zhu, Congfeng Yin, Daoyan Lin, Da-Wei Yang, Dingkun Wang, Ding Ai, Erle Zhu, Fang-Yu Yi, Feiyu Chen, Guohong Wen, Hai-Long Sun, Hai Zhao, Haiyi Hu, Hanchen Zhang, Han Liu, Hanyu Zhang, Hao Peng, Hao Tai, Haobo Zhang, He Liu, Hongwei Wang, Hong Yan, Hongyu Ge, Huan Liu, Hua Chu, Jiani Zhao, Jiachen Wang, Jiajing Zhao, Jia-Ming Ren, Jiapeng Wang, Jiaxin Zhang, Jia-Yi Gui, Jia yin Zhao, Jijie Li, Jing An, Jing Li, Jing Yuan, Jinhua Du, Jinxin Liu, Junkai Zhi, Junwen Duan, Kai Zhou, Kang Wei, Kedong Wang, Keyu Luo, Laiqiang Zhang, Lei Sha, Liang Xu, Lindong Wu, Li Ding, Lu Chen, Minghao Li, Nianyi Lin, Pankova Ta, Qiang Zou, Rong-Huang Song, Ruiqi Yang, Shangqing Tu, Shangtong Yang, Shaoxiang Wu, Sheng-Yan Zhang, Shijie Li, Shuang Li, Shuying Fan, Wei Qin, Wei Tian, Wei-Ning Zhang, Wenbo Yu, Wenjie Liang, Xiang-Tian Kuang, Xiang Cheng, Xiangyang Li, Xiao Yan, Xiaowei Hu, Xiaoying Ling, Xing Fan, Xi Xia, Xinyuan Zhang, Xin-Ze Zhang, Xirui Pan, Xunkai Zhang, Yan-Dong Wu, Yanfu Li, Yangyi Fang, Yidong Wang, Yi fan Zhu, Yi-Shuai Tan, Yi-Lin Zhou, Yiming Pan, Ying Zhang, Yinpei Su, Yi xin Geng, Yong Yan, Yonglin Tan, Yue Bi, Yuhan Shen, Yu-Hao Yang, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yurong Wu, Yutao Zhang, Yu Duan, Yuxuan Zhang, Ze-Xian Liu, Zhen Jiang, Zhenhe Yan, Zhe-Yu Zhang, Zhixiang Wei, Zhuofu Chen, Zhuoer Feng, Zijun Yao, Ziwei Chai, Ziyuan Wang, Zuzhou Zhang, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yu-Chao Dong, and Jie Tang. Glm-5: from vibe coding to agentic engineering. *arXiv*, abs/2602.15763, 2026. URL https://arxiv.org/abs/2602.15763.

Li Zhang, Yuzhen Shi, Yiran Hu, Jingwen Zhang, Wenbo Lv, Yu-Bo Ma, Wei Wang, Rong Shi, Yu-Ping Qiu, Xinran Xu, Yue Qi, Linlin Miao, Jaromir Savelka, Yun Liu, Kevin D. Ashley, Bing Zhao, Huarong Wei, and Lin Qu. Dlawbench: Evaluating llms through multi-turn legal consultation. *arXiv*, abs/2606.13931, 2026. URL https://arxiv.org/pdf/2606.13931.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Da-Cheng Li, Eric P. Xing, Haotong Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv*, abs/2306.05685, 2023. URL https://arxiv.org/abs/2306.05685.

Songhan Zuo, Shengbin Yue, Chiang Tao, Guanying Li, Yun Song, Xuanjing Huang, and Zhongyu Wei. Legalworld: A life-cycle interactive environment for legal agents. *arXiv*, abs/2606.18728, 2026. URL https://arxiv.org/abs/2606.18728.

APPENDIX CONTENTS

A Theoretical Foundations **21**
A.1 Two Dimensions of Transactional Lawyers' Competence . . . 21
A.2 IPO Requirements and the Five Actions . . . 22

B Dataset Design and Construction **23**
B.1 Expert Construction and Quality Control . . . 23
B.2 Dataset Composition and Coverage . . . 23
B.3 Rubric Design and Action Coverage . . . 25
B.4 Expert Participation . . . 26

C Evaluation Protocol and Metrics **27**
C.1 Tools and Execution Environment . . . 27
C.2 Task Profile . . . 29
C.3 Report Grading and Run Handling . . . 30
C.4 Scoring and Action Contributions . . . 31
C.5 Agreement with Expert Graders . . . 31
C.6 Source-Access Diagnostics . . . 31

D Expert Rubric Authoring Handbook **32**
D.1 Scope and Decision Rules . . . 32
D.2 Weights and Counting . . . 33
D.3 Criterion Record . . . 34

E Supplementary Results and Model Behavior **35**
E.1 Source Access . . . 35
E.2 Performance by Listing Category and Industry . . . 35
E.3 Case Studies and Failure Patterns . . . 38

F Worked Workspace: Company A **40**
F.1 Matter Overview . . . 40
F.2 Document Inventory and Accessible Resources . . . 41
F.3 Source Documents and External Records . . . 45
F.4 Rubric Examples . . . 46
F.5 Investigations across Due-Diligence Sections . . . 47

G Task and Grading Prompts **49**
G.1 Task Instructions . . . 49
G.2 LLM Grading Prompt . . . 51

A THEORETICAL FOUNDATIONS

The framework links two dimensions of transactional legal competence to five actions grounded in IPO practice.

A.1 TWO DIMENSIONS OF TRANSACTIONAL LAWYERS' COMPETENCE

Lawyer-effectiveness research identifies investigation, research, writing, analysis, problem solving, and practical judgement as distinct competencies (Shultz & Zedeck, 2011). Professional inference connects the diagnosis of a problem to an appropriate response (Abbott, 2014); research on Chinese corporate lawyers examines this process in transactional practice (Liu, 2006). Expertise research likewise distinguishes routine proficiency from adaptive problem solving (Bohle Carbonell et al., 2014) and relates efficient execution to innovation (Schwartz et al., 2005). Together, these accounts motivate evaluating both the reliability of legal work and the analysis and judgement developed from it. Table 3 connects these dimensions to the five actions.

Table 3: **Two dimensions of transactional lawyers' competence.** The last two columns describe how we use the source distinctions to operationalize diligence and judgement in IPO work.

Source and evidence type	Source distinction or content	Support for diligence	Support for judgement
Bohle Carbonell et al. (2014). Review of expertise research.	Routine expertise emphasizes proficient procedures; adaptive expertise supports problem solving in unfamiliar situations.	Motivates reliable execution of evidence review, verification, and reporting.	Motivates interpreting case evidence and developing responses to legal problems.
Abbott (2014). Theory of professional work.	Diagnosis, inference, and treatment connect professional knowledge to the handling of individual cases.	Evidence gathered and checked through Review and Verify supports professional diagnosis.	Identify and Assess connect case diagnosis to legal reasoning and appropriate responses.
Shultz & Zedeck (2011). Lawyer-effectiveness research.	Identifies 26 effectiveness factors, including investigation, research, writing, analysis, problem solving, and practical judgment.	Investigation, research, writing, and work organization provide legal-domain components for evidence and report quality.	Analysis, problem solving, and practical judgment provide components for identifying legal issues and determining responses.
Liu (2006). Study of Chinese corporate lawyers.	Examines legal documents, diagnosis, inference, and responses to clients within transactional practice.	Connects evidence and document work to the professional output assessed by Review, Verify, and Deliver.	Connects issue diagnosis and contextual legal reasoning to Identify and Assess.
SFC Code (2024), para. 17. Official sponsor requirements.	Combines evidence and record obligations with material-issue analysis and remedial advice.	Verification, disclosure, and records (17.6, 17.5, 17.10) make reliable execution observable.	Material-issue analysis and remedial advice (17.4(d), 17.3(b)) make judgment and responses observable.

From constructs to observable requirements. We operationalize *diligence* through evidence coverage and the correctness and completeness of the report: Review concerns reading the required sources and citing them as supporting evidence, Verify concerns factual, legal, and consistency accuracy, and Deliver concerns a complete report with substantive conclusions and professional presentation. We operationalize *judgement* through supported legal analysis and actionable judgement: Identify establishes each legal issue with its required facts, legal grounds, and their application, and Assess grades the significance of each risk and proposes remedies suited to the facts. Each action is measured through expert-authored rubrics applied to reports and tool records; Appendix C.4.1 specifies the assignments and scoring.

A.2 IPO Requirements and the Five Actions

In the Chinese mainland, the joint IPO legal-practice rules (China Securities Regulatory Commission et al., 2022) require lawyers to read professional opinions and retain supporting records (Art. 4; Review), verify legal matters and evidence (Arts. 2–4; Verify), and produce accurate, complete legal opinions (Arts. 2–3; Deliver). Article 3 requires evidence analysis and legal-risk identification (Identify). Article 5 and Article 6, paragraph 1(7), require evaluation of evidential limitations, legal defects, remedial measures, and their effect on the offering (Assess). In Hong Kong, paragraph 17 of the SFC Code of Conduct governs sponsor due diligence, including evidence assessment, disclosure, advice on remedies, and judgements of listing suitability (Securities and Futures Commission, 2024). Transaction lawyers support sponsors in both jurisdictions (China Securities Regulatory Commission, 2022; Securities and Futures Commission, 2024); the SFC also emphasizes adequate expertise and resources when engaging legal advisers (Securities and Futures Commission, 2026). Table 4 connects these shared IPO requirements to the LDD actions evaluated here.

Table 4: **Professional requirements supporting the five actions across the five listing categories.** Cells summarize selected requirements and identify articles/paragraphs and duty holders. Shared mainland lawyer duties are described in the text.[1]

Action	The Chinese mainland				Hong Kong
	A-share Main Boards	**STAR Market**	**ChiNext**	**BSE**	**Main Board**
Review	Retain verification records and working papers (SH/SZ 17, 28; S/L).	Retain verification records and working papers (SH 17, 28; S/L).	Retain verification records and working papers (SZ 17, 28; S/L).	Retain verification materials and working papers (BJ 24; L).	Review financial, litigation, and title records (PN21 12(b), 13(i)(l); S).
Verify	Verify matters within professional responsibility (SH/SZ 28; L).	Verify matters within professional responsibility (SH 28; L).	Verify matters within professional responsibility (SZ 28; L).	Check application and disclosure materials (BJ 23; S).	Question reliability; reconcile expert information (PN21 2, 5(b); S).
Deliver	Give supported eligibility opinions; ensure document quality (SH/SZ 20, 23, 28; S/L).	Give supported eligibility opinions; ensure document quality (SH 20, 23, 28; S/L).	Give supported eligibility opinions; ensure document quality (SZ 20, 23, 28; S/L).	Produce accurate, complete professional documents (BJ 24; L).	Document eligibility; complete disclosure (PN21 4; HK 17.4(a)(b); S).
Identify	Understand and analyze issuer risks (SH/SZ 27; S).	Analyze technology attributes and issuer risks (SH 19, 27; S).	Analyze innovation, growth, and issuer risks (SZ 19, 27; S).	Understand risks (BJ 23; S); disclose material risks (21; I).	Investigate litigation, legal conditions, and intellectual property (PN21 13(i)(j)(m); S).
Assess	Assess board fit and listing eligibility (SH/SZ 19, 27; S).	Assess STAR attributes and listing eligibility (SH 19, 27; S).	Assess ChiNext fit and listing eligibility (SZ 19, 27; S).	Assess eligibility (BJ 23; S); make professional determinations (24; L).	Recommend remedies; assess listing suitability (HK 17.3(b)(ii)(iii), 17.4(c)(d); S).

[1] *Sources and scope.* SH: Shanghai Stock Exchange, April 2024 rules (Shanghai Stock Exchange, 2024); SZ: Shenzhen Stock Exchange, April 2024 rules (Shenzhen Stock Exchange, 2024); BJ: Beijing Stock Exchange (BSE), April 2024 rules (Beijing Stock Exchange, 2024). PN21: Main Board due-diligence practice note (The Stock Exchange of Hong Kong Limited, 2026). HK: SFC Code of Conduct, October 2024 edition, paragraph 17 (Securities and Futures Commission, 2024). L: lawyers or securities-service providers within their professional remit; S: sponsors; I: issuers. SH/SZ denotes both exchanges' Main Boards. Sources describe duties; action labels are our synthesis.

B Dataset Design and Construction

This section describes the source materials, workspace composition, expert annotations, and human contributions.

B.1 Expert Construction and Quality Control

The construction team comprises 37 legal experts. The benchmark contains 63 distinct IPO matters selected across five listing categories and twelve industries. Each matter models a complete legal due diligence assignment. Public prospectuses and other filed disclosures provide the company structures, ownership, transactions, assets, and legal issues used to construct its synthetic workspace.

Matter selection and issue extraction. Experts select publicly filed IPO matters across listing categories and industries. They examine prospectuses, legal opinions, supplemental legal opinions, lawyers' work reports, and other public filings, and set the matter's reporting period and report date. From these sources they extract matter-specific investigation issues, classify them by topic and subtopic, and draft the expected factual and legal conclusions. Facts are specified at a level that permits checking individual records, such as each patent implicated in an ownership issue. The matter's task instructions use its report date as the legal cutoff.

Mapping and anonymization. For each matter, an entity fact card, an event timeline, and a document inventory map the expected facts to the workspace. The inventory identifies the documents carrying each issue's evidence; the released document index lists each PDF and its navigation category. Selected identifying details are replaced consistently across files while legally relevant relationships and dates are retained. The drafting checklist groups documents into fourteen practical categories.

Document drafting and review. Experts author the supporting documents from the mapped facts and distribute related evidence across files so that issues require cross-document review. The benchmark workspaces contain 119–311 PDFs per matter, with 12,151 PDFs and 30,243 pages overall; the documents are rendered predominantly as scans. A legal expert checks the fact card, and cross-file review reconciles names, dates, amounts, ownership, and event sequences to make the corporate history traceable. This review establishes a consistent factual baseline before controlled document defects and external-record distractors are introduced and documented. A second legal expert reviews the resulting rubrics. Appendix B.4 consolidates personnel roles, education, and compensation.

B.2 Dataset Composition and Coverage

The matters cover 12 broad industry categories. The most frequent are information technology (16 matters), industrials (10), raw materials (9), consumer staples (5), and healthcare (5). Matters contain between 119 and 311 PDFs, between 311 and 1,493 pages, and between 198 and 367 scoring rubrics. The median counts are 182 PDFs, 457 pages, and 239 scoring rubrics. Investigation requirements range from 3 to 17 items per matter, with a median of 12.

Table 5 summarizes the dataset by listing category.

Table 5: ALawBench by listing category. Database counts use parsable records; positive and negative columns partition the 15,557 scoring rubrics.

Listing category	Matters	PDFs	Pages	DB records	Distractors	Positive	Negative
A-share Main Boards	21	4,136	9,841	2,660	618	2,035	2,900
Beijing Stock Exchange	17	3,194	7,622	2,233	508	1,780	2,468
STAR Market	10	1,935	4,481	1,507	259	1,027	1,431
ChiNext	8	1,501	3,474	1,054	275	834	1,163
Hong Kong Main Board	7	1,385	4,825	1,260	303	791	1,128
Total	**63**	**12,151**	**30,243**	**8,714**	**1,963**	**6,467**	**9,090**

Listing categories and industries. We group matters into five listing categories. The A-share Main Boards category combines the Main Boards of the Shanghai Stock Exchange (SSE) and Shen-

zhen Stock Exchange (SZSE); Beijing Stock Exchange (BSE) is a separate exchange; STAR Market and ChiNext denote the SSE STAR Market and SZSE ChiNext Market, respectively. Hong Kong Main Board denotes the Main Board of The Stock Exchange of Hong Kong Limited (SEHK). All seven Hong Kong matters concern Main Board listings. Tables 6 and 7 report matter counts and mean workloads for each group. Performance by these groups is reported in Appendix E.2.

Table 6: Listing categories and per-matter mean workload counts. Positive and negative rubric counts follow the same scoring inventory as Table 5.

Listing category	Matters	PDFs	Pages	Positive	Negative
A-share Main Boards	21	196.95	468.62	96.90	138.10
Beijing Stock Exchange	17	187.88	448.35	104.71	145.18
STAR Market	10	193.50	448.10	102.70	143.10
ChiNext	8	187.62	434.25	104.25	145.38
Hong Kong Main Board	7	197.86	689.29	113.00	161.14

Table 7: Industry groups and per-matter mean workload counts. Positive and negative rubric counts follow the same scoring inventory as Table 5.

Industry	Matters	PDFs	Pages	Positive	Negative
Information technology	16	184.06	457.62	116.62	153.94
Industrials	10	192.50	578.40	102.80	142.50
Raw materials	9	186.89	430.33	100.89	140.89
Consumer staples	5	189.00	400.80	110.40	142.00
Healthcare	5	175.60	419.20	111.00	144.80
Consumer discretionary	4	226.75	553.00	87.75	144.25
Real estate	3	241.00	552.67	84.33	133.33
Materials	3	173.67	447.33	80.67	138.00
Energy	3	193.67	485.67	85.00	125.33
Communication services	2	190.50	556.50	79.00	154.00
Financials	2	191.00	400.50	80.00	136.50
Utilities	1	281.00	581.00	139.00	152.00

Per-matter document counts. Table 8 retains the source workbook identifiers for the individual matters.

Table 8: Per-matter document counts, with IDs retained from the source workbook.

ID	PDFs	Pages	ID	PDFs	Pages	ID	PDFs	Pages
1	215	1493	**22**	169	498	**43**	174	412
2	285	690	**23**	227	481	**44**	166	412
3	242	686	**24**	263	475	**45**	182	407
4	146	677	**25**	235	475	**46**	172	399
5	169	624	**26**	182	473	**47**	162	399
6	257	622	**27**	184	472	**48**	167	396
7	241	619	**28**	159	470	**49**	156	393
8	281	581	**29**	148	462	**50**	165	392
9	176	558	**30**	218	461	**51**	142	384
10	209	539	**31**	184	461	**52**	145	382
11	192	539	**32**	188	457	**53**	182	380
12	171	534	**33**	153	453	**54**	173	376
13	244	530	**34**	133	453	**55**	168	372
14	286	525	**35**	170	452	**56**	246	371
15	232	525	**36**	135	451	**57**	156	367
16	194	521	**37**	198	450	**58**	154	360
17	237	514	**38**	235	436	**59**	161	355
18	252	511	**39**	204	433	**60**	171	354
19	311	505	**40**	209	425	**61**	133	336
20	160	502	**41**	190	416	**62**	150	321
21	239	499	**42**	184	416	**63**	119	311

B.3 Rubric Design and Action Coverage

The benchmark uses 15,557 independently scored rubrics: 6,467 positive and 9,090 negative. Positive weights are $+3$, $+7$, and $+10$; negative weights are -3, -7, and -10. Table 9 combines their distribution across the five actions with the theoretical scoring intervals. Review, Verify, and Deliver form diligence; Identify and Assess form judgement.

Each rubric belongs to exactly one of the five actions in Table 9. A rubric may require several facts jointly; its decision rule fixes which conditions establish satisfaction. The unit of scoring remains one independently verifiable requirement. Appendix D gives the expert authoring instructions.

Table 9: **Rubric distribution and theoretical scoring intervals by action.** (a) Counts at each signed weight sum to 15,557. (b) Bounds on the mean raw score and mean contribution to the net score rate across all 63 cases; pp denotes percentage points. Bounds apply the rubric counting rules and mutually exclusive penalty tiers. Verify has no finite lower bound because errors can incur repeated penalties.

(a) Rubric counts by action and signed weight

Action	−10	−7	−3	+3	+7	+10	Total
Review	63	63	1,510	0	63	0	1,699
Verify	62	63	5,310	0	0	0	5,435
Deliver	63	63	1,893	309	0	0	2,328
Identify	0	0	0	3,790	399	35	4,224
Assess	0	0	0	1,866	4	1	1,871
Total	**188**	**189**	**8,713**	**5,965**	**466**	**36**	**15,557**

(b) Theoretical intervals for 63-case means

Action	Raw score (points)	Net-rate contribution (pp)
Review	$[-88.90,\ 7.00]$	$[-27.17,\ 2.14]$
Verify	$(-\infty,\ 0.00]$	$(-\infty,\ 0.00]$
Deliver	$[-107.24,\ 14.71]$	$[-32.73,\ 4.49]$
Identify	$[0.00,\ 230.37]$	$[0.00,\ 67.61]$
Assess	$[0.00,\ 89.46]$	$[0.00,\ 25.76]$
Total	$\mathbf{(-\infty,\ 341.54]}$	$\mathbf{(-\infty,\ 100.00]}$

Theoretical scoring intervals. For matter c and action a, let l_{ca} and u_{ca} be its lowest and highest permitted raw scores. The upper bound sums its positive weights. The lower bound applies each penalty's counting rule, retaining only the most severe tier within a mutually exclusive group. Review penalties count once; Deliver penalties count once except for one formatting rubric that permits up to three distinct error types. Verify includes penalties with no cap on error occurrences, giving $l_{c,\text{Verify}} = -\infty$. Identify and Assess contain only positive rubrics and therefore have a lower bound of zero. Panel (b) averages these endpoints over the 63 cases. Its net-rate endpoints average $100l_{ca}/P_c$ and $100u_{ca}/P_c$, where P_c is the case's total available positive points, following Appendix C.4.1. The upper contributions sum to 100 percentage points. These rubric-defined intervals apply to both task profiles.

B.4 Expert Participation

Thirty-seven legal experts with legal-practice experience contributed to the 63 matters. Table 10 describes their roles; construction stages draw on the same expert pool.

B.4.1 Roles and Review Process

Construction and review follow the process in Appendix B.1; the table identifies the contribution at each stage.

Table 10: Human participation throughout construction and evaluation. The construction stages draw on the same 37-person expert pool.

Stage	Personnel allocation	Human contribution
Source selection and fact extraction	Construction expert pool	Select the IPO matters; extract company fact cards, relationships, and event timelines from disclosures.
Workspace materials	Construction expert pool	Anonymize entities and author the supporting materials under the fourteen-category checklist.
Factual and cross-file review	Legal-expert review within the construction process	Check the fact card and reconcile names, dates, amounts, ownership, and other facts across documents before injection.
Defects and external records	Construction expert pool	Construct controlled defects, supported database records, and distractors; document the intended conflicts.
Rubric development	Authoring experts; one additional expert reviews annotations	Specify requirements, evidence, decision boundaries, signed weights, and triggers; check the resulting rubrics against the case materials.
Grader calibration	Three independent legal experts per sampled item	Independently label each of 3,321 rubric decisions; use the two-of-three majority as the human reference.
Automated evaluation	Gemini 3.8 Flash plus deterministic audits	Judge reports against the rubrics and combine decisions with tool-log checks.

The grader calibration spans eleven workspaces, with 9,963 individual expert judgements. Against majority labels, Gemini 3.8 Flash achieves accuracy of 0.932, F1 of 0.761, and Cohen's $\kappa = 0.722$ (Appendix C.5). The automated pipeline scores the full benchmark.

B.4.2 Education Background

The personnel roster records two experts at bachelor's level, thirty-two at master's level, and three at doctoral level. Table 11 reports the aggregate distribution over all 37 construction experts.

Table 11: Aggregate education background of the 37-person construction team, using the highest education level recorded in the personnel roster. All contributors have legal-practice experience.

Education level	Experts	Share (%)
Bachelor's	2	5.41
Master's	32	86.49
Doctoral	3	8.11
Total	**37**	**100.00**

B.4.3 Compensation

Hourly compensation was approximately USD 25–40 for construction and annotation, and USD 40–60 for expert review and grader calibration.

C Evaluation Protocol and Metrics

The protocol specifies the execution environment, submission setting, scoring rules, and human-label comparison.

C.1 Tools and Execution Environment

Figure 8 summarizes report production and independent evaluation. The ALawBench engine provides general-purpose file and planning tools, while three MCP services expose document, corporate-record, and legal-research tools. After an engine iteration ends, the harness collects the submitted report, converts DOCX output to Markdown, and runs the goal checker when the revision loop is enabled.

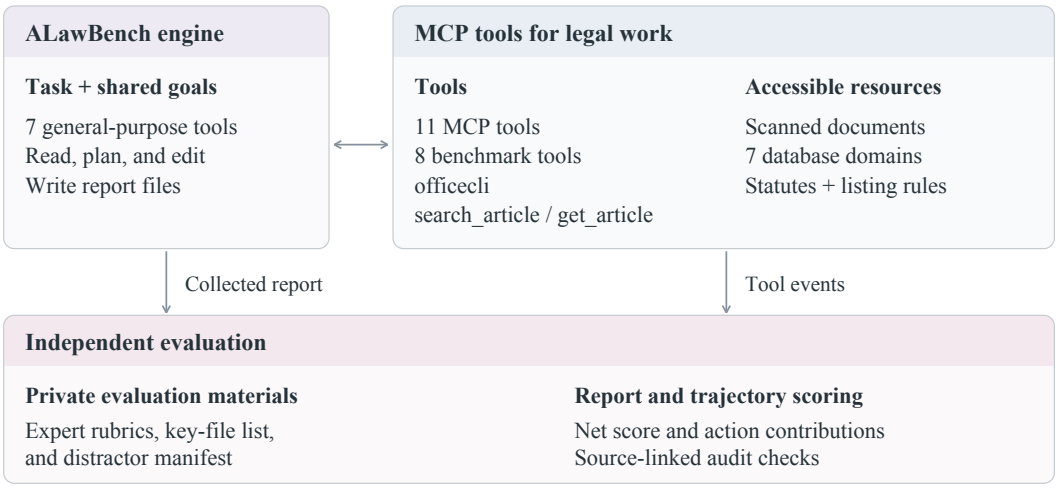


Figure 8: Standard experimental interface and evaluation workflow. Seven general-purpose tools and eleven MCP tools support investigation and report production. Database availability varies by matter. The collected report and tool trajectory are evaluated using private rubrics, key-document lists, and distractor provenance.

Table 12: The standard experimental interface: seven general-purpose tools and eleven MCP tools. Names omit the service prefix; each row is one registered tool.

Tool	Operation
General-purpose tools (7)	
`read_file`	Read files in the permitted workspace.
`edit`	Edit workspace files.
`grep_search`	Search workspace text.
`list_directory`	List workspace files and directories.
`todo_write`	Maintain the investigation task list.
`enter_plan_mode`	Enter the engine's planning mode.
`exit_plan_mode`	Leave planning mode.
ALawBench MCP tools (8)	
`glob_files`	Find workspace documents by path pattern.
`grep_documents`	Search readable document text.
`read_document_tool`	Read supported document formats.
`ocr_document_tool`	OCR selected pages or a complete PDF.
`write_docx_tool`	Write structured content to a DOCX report.
`query_offline_db_tool`	Query the matter's corporate records across seven database domains in the benchmark.
`query_listing_rule_tool`	Search the configured listing-rule corpus.
`review_docx_layout_tool`	Render and visually review a report page.
OfficeCLI MCP tool (1)	
`officecli`	Create, edit, inspect, or render Office documents.
PKULaw MCP tools (2)	
`search_article`	Semantically search legal provisions.
`get_article`	Retrieve a specific legal provision.
Total: 18 tools	**7 general-purpose tools + 11 MCP tools**

Recorded configurations. The retained trajectories record tool configuration and initialization events for workspace investigation, research, and report production.

Evidence services. The seven database domains are corporate registration, industrial intellectual property, copyright, domain names, judicial records, credit information, and leases and procurement. Across the benchmark, these comprise 23 table types, identified by database and table name; availability varies by matter. The lease and procurement database supplies 71 records in one matter. Distractor provenance is retained separately for evaluation. OCR accepts a page range or a complete PDF: 95,341 recorded operations process more than two pages, with a maximum of 38. The traces also record 28,302 in-process OCR cache hits. The task instructions require checking legal authorities against the report date specified for each matter. PKULaw is a professional Chinese legal-information database covering legislation and case law. We use its MCP tools `search_article` for semantic search over legal provisions and `get_article` for retrieving specified provisions during legal verification. The `query_listing_rule_tool` queries the configured listing-rule corpus. OfficeCLI exposes one tool, `officecli`, whose command argument selects document operations; `review_docx_layout_tool` is a separate benchmark tool that renders a page through OfficeCLI and obtains a visual layout review.

C.2 Task Profile

All runs use the ALawBench engine with the matter-specific LDD assignment, shared goal requirements, tool guidance, and report date as the legal cutoff in the opening instructions. The requirements appear under a Goal heading. After each collected report, a separate call to the evaluated model checks the report against these goals and returns unmet requirements for revision, for up to 20 rounds. Like the agent, the checker has no access to rubrics or key-document checklists. The task specifies eight shared investigation requirements, presented as eight paragraphs in Appendix G. Independent rubric scoring applies to the final collected report. We use each model's default reasoning effort.

C.2.1 Evaluated Models

The main benchmark evaluates 21 models on 63 IPO matters. Net score rates summarize matter-level performance. The high-coverage analysis uses the populations defined in Appendix C.6. Table 13 lists their families, names, and checkpoints, together with Gemini 3.8 Flash, the independent grader.

Table 13: Model families, names, and checkpoints for 21 evaluated models and the independent grader, Gemini 3.8 Flash. Citations identify the corresponding model references.

Family	Model	Checkpoint
Kimi	Kimi K3 (Kimi Team et al., 2026)	`moonshotai/Kimi-K3`
GLM	GLM 5.3 (Zeng et al., 2026)	`zai-org/GLM-5.3`
	GLM 5.2 (Zeng et al., 2026)	`zai-org/GLM-5.2`
Qwen	Qwen 3.8 Max (Qwen Team, 2026b)	`qwen3.8-max-0902`
	Qwen 3.7 Max (Qwen Team, 2026a)	`qwen3.7-max-2026-06-08`
GPT	GPT 6 Astra (OpenAI, 2026c)	`gpt-6-astra`
	GPT 5.6 Sol (OpenAI, 2026b)	`gpt-5.6-sol`
	GPT 5.5 (OpenAI, 2026a)	`gpt-5.5-2026-04-23`
Grok	Grok 4.6 (xAI, 2026b)	`grok-4.6`
	Grok 4.5 (xAI, 2026a)	`grok-4.5`
DeepSeek	DeepSeek V4 Pro (DeepSeek-AI et al., 2026)	`deepseek-ai/DeepSeek-V4-Pro-0813`
	DeepSeek V4 Flash (DeepSeek-AI et al., 2026)	`deepseek-ai/DeepSeek-V4-Flash-0731`
Claude	Claude Fable 5 (Anthropic, 2026a)	`claude-fable-5`
	Claude Opus 5 (Anthropic, 2026b)	`claude-opus-5`
MiniMax	MiniMax M3 (MiniMaxAI, 2026)	`MiniMaxAI/MiniMax-M3`
Doubao	Doubao Seed 2.1 Pro (ByteDance Seed Team, 2026)	`doubao-seed-2-1-pro-260628`
Gemini	Gemini 3.8 Flash (Google DeepMind, 2026d)	`gemini-3.8-flash`
	Gemini 3.7 Flash (Google DeepMind, 2026c)	`gemini-3.7-flash`
	Gemini 3.6 Flash (Google DeepMind, 2026b)	`gemini-3.6-flash`
	Gemini 3.1 Pro (Google DeepMind, 2026a)	`gemini-3.1-pro-preview`
HY	HY3 (Tencent Hy Team, 2026)	`tencent/Hy3`
MiMo	MiMo V2.5 Pro (Xiaomi MiMo Team, 2026)	`XiaomiMiMo/MiMo-V2.5-Pro`

C.3 Report Grading and Run Handling

C.3.1 Scoring Procedure

Report grading. We organize report grading by action: Review, Verify, Deliver, Identify, and Assess. For each report, we divide the rubrics within each action into batches of 20 rubrics. Each call to the external grader, Gemini 3.8 Flash (Google DeepMind, 2026d), receives the report and one batch; the final batch contains the remaining rubrics when fewer than 20 remain. Appendix G.2 provides the grading prompt. The grader returns an individual decision for each rubric, supported by evidence from the report. Positive rubrics are judged as satisfied or unsatisfied; negative rubrics are judged as triggered or not triggered, with occurrence counts for cumulative penalties. The grader also determines whether the report adopts distractor records as company facts or risks.

Item scoring. Satisfied positive rubrics receive their full weight; unsatisfied ones receive zero. Negative rubrics deduct their weight according to the stated trigger and counting rules. Mutually exclusive penalty tiers apply only the most severe triggered tier. We associate each report with its run and trajectory, combine the item decisions with the checks below, and aggregate scores by action. Appendix C.4.1 defines action contributions.

Document access. File-access checks contribute to Review. The trajectory records tool calls, arguments, and results; backend execution events record success status, accessed paths, pages, character counts, and database tables. We identify document access from reading, OCR, and document-search events. The full-document-access rubric awards $+7$ when all visible documents have recorded access. Mutually exclusive key-document penalties deduct 3 points when the unread share is at least 1% but below 5%, 7 points when it is at least 5% but below 10%, and 10 points when it is at least 10%.

Legal citations and application. We extract legal citations from the report and retrieve the corresponding laws and provisions from PKULaw using `search_article` and `get_article`. The grader compares each citation with the retrieved text, the case facts, expert-annotated legal bases, and relevant dates to check the law's name, version, article number, provision content, and applicability. Applicability checks consider the listing board, the issue being analyzed, and the relevant time. Each confirmed erroneous citation incurs a seven-point penalty in Verify. Multiple errors at the same citation location count once.

C.3.2 Run Inclusion

The benchmark contains 63 matters. For each model, we average over all 63 matters, assigning zero to missing or screened-out reports. The same matter set and zero-score rule apply to each action. Source-access diagnostics use only reports with matched scoring and trajectory records.

C.4 Scoring and Action Contributions

C.4.1 Action Definitions and Net Contributions

The five actions and their two dimensions are defined in Section 2.2. The following formulas specify their contributions to the report score.

Each rubric has an action-based identifier, such as `REVIEW-N001`. The prefix names its action, P or N specifies a positive or negative rubric, and the three-digit suffix identifies it within the matter, action, and polarity. The released rubric files and indexes use the same identifiers and action order.

For matter c, let P_c be the total available positive points, which ranges from 207 to 556 across matters. Let e_{cma} and q_{cma} be model m's earned positive points and actual penalty magnitude for action a. Its contribution to the net score rate is

$$A_{cma} = 100\frac{e_{cma} - q_{cma}}{P_c}, \qquad A_{ma} = \mathrm{mean}_c\, A_{cma}. \tag{2}$$

All actions share the matter denominator. Their model means add to the rubric-based net score rate: $N_m = D_m + J_m = \sum_a A_{ma}$. Actions and dimensions contribute percentage points to N_m; the net rate has an upper bound of 100.00% and no finite lower bound when penalties accumulate.

For judgement, let P_{cJ} be the available positive points assigned to Identify and Assess. Its mean available contribution is $\bar{U}_J = \mathrm{mean}_c(100P_{cJ}/P_c)$. We report the judgement score as $100J_m/\bar{U}_J$, using unrounded values in calculations.

C.5 Agreement with Expert Graders

We randomly sample 3,321 rubric decisions, each identified by a *matter* × *evaluated model* × *RubricID* tuple, spanning eleven workspaces. Three legal experts independently assign `true` or `false` to each item. The human reference label is `true` when at least two experts vote `true`, and `false` otherwise. We compare Gemini 3.8 Flash's decision with this two-of-three majority label for every sampled item.

Taking `true` as the positive class, TP and TN count matching true and false decisions; FP and FN count false positives and false negatives relative to the human majority. With $n = \mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}$, we compute

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{n}, \qquad \mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \qquad \kappa = \frac{\mathrm{Acc} - p_e}{1 - p_e}, \tag{3}$$

where $p_e = p_H p_G + (1 - p_H)(1 - p_G)$ is chance agreement and p_H and p_G are the proportions of `true` labels from the human majority and Gemini 3.8 Flash, respectively. We estimate 95% confidence intervals (CIs) for F1 and κ using 6,000 bootstrap resamples of individual rubric decisions, keeping the human and grader labels paired. Table 14 reports the results.

C.6 Source-Access Diagnostics

Document access and professional analysis. The diagnostic in Section 4.3 matches scored reports to their trajectory audits by model and matter. Legal experts select the key documents that

Table 14: **Gemini 3.8 Flash agreement with expert majority labels.** Each of the 3,321 items receives three independent expert judgements; the reference uses a two-of-three majority.

Comparison	n	Acc	F1	F1 95% CI	Cohen's κ	κ 95% CI
Gemini 3.8 Flash vs. majority	3,321	0.932	0.761	[0.730, 0.791]	0.722	[0.686, 0.756]

require priority review for each LDD matter. Counting each matter once, the 63 checklists contain 9,377 files; their share of the dataset's 12,151 PDFs is $100 \times 9{,}377/12{,}151 = 77.2\%$. For report r, let K_r be the key-document checklist used for that report and V_r the set of files with recorded access in that run. Repeated accesses to the same file count once. Key-document coverage and the subset selected at threshold τ are

$$C_r = \frac{|K_r \cap V_r|}{|K_r|}, \qquad \mathcal{H}_\tau = \{r : \text{report and score available},\ C_r \geq \tau\}.$$

The main-text analysis uses the subset $\mathcal{H}_{0.95}$. Figure 7 and Table 15 assign every scored report to the coverage interval $[0, 0.5)$, $[0.5, 0.8)$, $[0.8, 0.95)$, or $[0.95, 1]$ and average each matter-level action contribution A_{cma} over the reports in each interval. Confidence intervals use the t distribution across reports within each interval. As a robustness check, we regroup the same reports by their coverage of all workspace documents listed in the trajectory audits. The two coverage measures have a report-level Spearman correlation of 0.97, and the pattern is unchanged. From the lowest to the highest group, Verify changes by $+0.76$ points and Assess by $+3.32$, while Review, Deliver, and Identify improve by 9.86 to 13.10 points.

Table 15: **Mean action contributions by key-document coverage.** Contributions are in percentage points, and Change is the highest group minus the lowest.

Coverage (%)	Runs	Review	Verify	Deliver	Identify	Assess
< 50	543	-13.86	-5.05	-14.67	10.29	6.89
50–80	333	-11.01	-5.10	-10.68	13.45	7.60
80–95	98	-7.76	-5.64	-8.14	18.47	9.11
≥ 95	349	-3.62	-4.56	-4.08	23.51	10.40
Change		$+10.25$	$+0.49$	$+10.59$	$+13.22$	$+3.50$

For rubric group g, let n_{rg} be the number of eligible positive rubrics and s_{rg} the number judged satisfied in report r. Complete factual checks, risk-identification checks, and remediation checks use the factual-identification rubrics in Identify, risk-identification rubrics in Identify, and remediation rubrics in Assess, respectively. Each rubric contributes one binary decision, regardless of its point weight. Risk identification here uses the risk-identification rubrics within Identify. The subset means are

$$\overline{C}_\tau = \frac{100}{|\mathcal{H}_\tau|} \sum_{r \in \mathcal{H}_\tau} C_r, \qquad \overline{Q}_{g,\tau} = \frac{100}{|\mathcal{H}_\tau|} \sum_{r \in \mathcal{H}_\tau} \frac{s_{rg}}{n_{rg}}.$$

Every selected report has equal weight. At the main-text threshold, the report-level risk fractions average to 28.03%, and mean coverage is 99.41%.

D Expert Rubric Authoring Handbook

This handbook describes how experts turn each case's expected legal work into independently decidable criteria. Each case has rubric files for five actions; Appendix C.3 describes how their decisions are scored.

D.1 Scope and Decision Rules

Assign an action and a polarity. For every case, authors consider all five actions: Review for reading and citing required sources, Verify for factual, legal, and consistency errors, Deliver for

report completeness, substantive conclusions, and presentation, Identify for establishing legal issues with their facts, legal grounds, and application, and Assess for risk grading and remedies. Each scored criterion belongs to exactly one action. Positive criteria reward expected or especially strong output; failure to satisfy one earns zero. Negative criteria identify an erroneous or harmful output or a specified diligence failure; absence of the trigger incurs zero penalty. Positive and negative decisions are independent. The `evaluation_group` field describes what is checked within the assigned action.

Write one decidable requirement. State an accurate, unambiguous condition that an evaluator can judge as `true` or `false`. Separate independently meaningful facts, risk categories, remedies, and errors into different criteria so that the same defect is neither rewarded nor penalized twice. Across the case, the criteria should cover the required evaluation points. A criterion may still require several inseparable facts jointly when their conjunction establishes one legal conclusion; list each indispensable fact and say explicitly that all are required. Distinguish these *core requirements* from helpful background that is not a pass condition. Do not turn a long list of optional context into an all-or-nothing requirement. For alternative acceptable grounds or remedies, state whether any one suffices or give the exact threshold (for example, at least two of three listed measures).

Specify what counts as satisfaction. Give the correct case-specific values, legally relevant dates, acceptable alternatives, and the rule for accepting materially equivalent wording in the criterion itself. Business descriptions need not match a fixed phrase when the underlying meaning is correct. Conversely, missing core facts do not receive credit merely because the answer points in the right direction. Require an exact heading, numbering order, or fixed phrase only when the task instructions call for it; put such a form requirement in its own criterion. Distinct errors, such as failure to enroll employees in social insurance, use of an understated contribution base, and payment through a third party, should be evaluated separately. Apply each criterion to its specified report location or evidence scope; a fact elsewhere does not automatically satisfy a section-specific requirement.

Supply the evidence needed for the decision. For report-content criteria, the `statement` should include the expected facts or triggering errors so that a grader can apply the decision rule without opening an answer key or data-room file. Document-access criteria identify the document inventory or key-document list to compare with recorded tool events. Legal-citation criteria specify the errors to check against retrieved provisions and case dates. State the comparison rule explicitly instead of referring the grader to an unspecified answer key.

Second expert review. Check the task instructions and case evidence against every correct value and legal proposition. Test both a satisfying and a nonsatisfying answer; check for overlap and omissions across criteria; and verify action, polarity, weight, and counting metadata. Split overpacked criteria, demote optional facts to context, and make alternative acceptable formulations explicit.

D.2 Weights and Counting

Authors classify each criterion by issue and importance or harm. Table 16 shows the signed weights recorded for the corresponding tiers. Positive importance reflects the consequence of missing the required work for the listing analysis. Negative severity considers harm to the legal conclusion or professional duty, how readily a lawyer can detect the mistake, and the cost of checking and repairing it.

Table 16: Rubric weight tiers and their signed values.

Weight tier	Weight	Authoring interpretation
Critically Important	$+10$	A central finding whose omission defeats a major part of the investigation, such as a material listing risk.
Important	$+7$	A substantial finding or diligence safeguard, including a medium-severity risk or complete document access.
Slightly Important	$+3$	A narrower fact, legal ground, risk grade, remedy, or required report detail.
Slightly Detrimental	-3	A limited error or disclosure, formatting, or diligence defect that is usually straightforward to identify and correct.
Detrimental	-7	A material error requiring closer source or legal verification, such as a wrong legal citation or an intermediate document-coverage failure.
Critically Detrimental	-10	A severe defect that undermines trust in the work, such as an unlawful remedy or high document-coverage failure.

An unsatisfied positive criterion always contributes zero. For a negative criterion, specify whether a trigger is counted once, once per distinct occurrence, or under a bounded rule; separate mutually exclusive severity tiers and apply only the most severe triggered tier. The released `trigger_count` field is null when no counting instruction is recorded; some positive criteria also state that they count once. A binary trigger decision can therefore coexist with an occurrence count for cumulative penalties.

D.3 Criterion Record

The released JSON has one file per case and action. Each file records its schema version, case and action identifiers, dimension and side keys, English and Chinese labels, criterion count, and a `rubrics` array. Table 17 follows the fields stored in each criterion. The author drafts the decision rule; the record stores its scoring and evaluation metadata.

Table 17: Fields in the released rubric records.

Field	Meaning and authoring check
`id`, `statement`	Stable action-prefixed identifier and the self-contained positive pass condition or negative trigger, including required facts, alternatives, and evidence scope.
`action`, `dimension`, `side`	Action and routing keys; `dimension` repeats the action key, while `side` groups the three diligence actions and two judgement actions.
`polarity`, `score`, `priority`	Positive or negative direction, signed weight, and the corresponding descriptive weight tier stored in the release.
`scoring_status`, `rubric_type`	Inclusion status and the criterion's issue type; the released scoring criteria have status `included`.
`criterion_type`, `criterion_necessity`	`Objective` or `Subjective` decision type; `Explicit` or `Implicit` requirement source.
`evaluation_group`	Evaluation topic, such as risk identification, factual accuracy, or document coverage.
`trigger_count`	Null or a recorded counting instruction; check occurrence units, caps, and mutually exclusive tiers where applicable.
`eval_method`, `agent_scored`	Evaluation routing and whether the separate agent-based check is used.
`rule_key`, `inspection_component`	Optional keys used for designated scoring rules or inspection components.

E SUPPLEMENTARY RESULTS AND MODEL BEHAVIOR

Results cover source access, differences across matter groups, and observed model failures.

E.1 SOURCE ACCESS

E.1.1 CONTENTS OF THE KEY-DOCUMENT SUBSET

The key documents are the matter-specific source materials that legal experts designate for priority review. They include corporate registration records and articles of association; shareholder registers, financing agreements, and board resolutions; audit reports, tax filings, and payment records; property certificates, leases, and mortgages; patent certificates and research-and-development agreements; customer and supplier contracts; operating licenses and environmental approvals; employment contracts and social-insurance records; litigation and arbitration documents; and approvals for projects financed with offering proceeds. Together, these materials supply evidence for checking corporate history, ownership, operations, assets, compliance, and the legal risks addressed in the report.

Across the 63 matters, the expert-designated checklists identify 9,377 of the 12,151 workspace PDFs, or **77.2%** of all files. Each matter contains 43–259 key documents (median: 144). Selection is matter-specific: the checklist names the individual documents to review, including related records that must be compared across files. For example, Company A's checklist includes financing agreements, jointly owned patent certificates, and the corresponding collaborative research-and-development agreement, connecting ownership and financing terms to intellectual-property rights.

E.2 PERFORMANCE BY LISTING CATEGORY AND INDUSTRY

We disaggregate the Figure 1 Net score by the five listing categories in Table 5 and the twelve industry labels in the source workbook. Each entry averages matter-level Net (positive credit minus penalties, divided by the matter's available positive score) over all assigned matters in that group, with missing or screened-out reports contributing zero. Weighting the subgroup means by the group matter counts recovers overall Net.

Listing categories. Table 18 reports model Net by listing category; the corresponding workloads are in Table 6. GPT 6 Astra leads in all five listing categories, from 34.51 on ChiNext to 44.47 on STAR Market. Averaged across the 21 models, category-level Net ranges from -4.07 on Beijing Stock Exchange matters to 2.48 on Hong Kong Main Board matters.

Table 18: Model Net (%) by listing category. Net is averaged within each group. Model order follows the overall ranking; bold marks column leaders among individual models.

Model	A-share Main Boards	Beijing Stock Exchange	STAR Market	ChiNext	Hong Kong Main Board
GPT 6 Astra	**38.89**	**42.31**	**44.47**	**34.51**	**41.06**
Qwen 3.8 Max	28.83	24.74	28.59	28.37	25.99
GLM 5.3	30.46	24.22	24.99	27.76	26.35
Kimi K3	32.04	23.25	27.88	21.94	24.53
Claude Opus 5	18.67	13.41	25.54	19.04	18.31
Grok 4.6	11.14	6.48	7.95	10.19	16.00
GPT 5.6 Sol	8.04	2.19	6.56	10.27	9.00
Claude Fable 5	2.30	-0.05	-3.50	-0.04	3.50
DeepSeek V4 Flash	-2.07	-4.54	-1.60	14.72	6.00
GPT 5.5	-4.11	-1.75	-3.68	4.39	9.78
Grok 4.5	-6.82	-5.79	-5.86	-5.70	6.68
MiniMax M3	-5.00	-10.84	-5.72	-8.28	2.73
GLM 5.2	-9.19	-13.30	-16.37	-5.66	-5.28
Gemini 3.6 Flash	-12.17	-14.51	-7.27	-15.72	-2.69
Doubao Seed 2.1 Pro	-14.66	-14.67	-22.21	-24.88	-13.32
DeepSeek V4 Pro	-18.50	-20.71	-18.38	-20.41	-16.94
Qwen 3.7 Max	-16.25	-24.87	-21.89	-21.22	-19.41
Gemini 3.7 Flash	-20.02	-30.46	-4.26	-47.30	-3.06
HY3	-24.71	-22.51	-15.97	-18.64	-27.50
Gemini 3.1 Pro	-23.65	-21.57	-25.50	-25.41	-19.65
MiMo V2.5 Pro	-33.87	-36.52	-34.56	-35.79	-29.95
Mean across models	**-0.98**	**-4.07**	**-0.99**	**-2.75**	**2.48**

Industries. Table 19 reports model-level Net by industry, using the groups characterized in Table 7. GPT 6 Astra leads in ten of twelve industries, including the three largest groups: information technology (42.17), industrials (45.19), and raw materials (35.51). Qwen 3.8 Max leads in consumer staples (36.42), and Kimi K3 leads in financials (34.76). The corresponding cross-model means for the three largest groups are 4.37, -5.26, and -8.40. The source workbook distinguishes *raw materials* and *materials*.

Table 19: Model Net (%) across all twelve industries. Net is averaged within each group. Model order follows the overall ranking; bold marks column leaders among individual models.

Model	Information technology	Industrials	Raw materials	Consumer staples	Healthcare	Consumer discretionary	Real estate	Materials	Energy	Communication services	Financials	Utilities
GPT 6 Astra	**42.17**	**45.19**	**35.51**	34.75	**51.46**	**34.21**	**48.11**	**39.88**	**45.94**	**18.79**	19.68	**51.43**
Qwen 3.8 Max	31.72	20.72	19.02	**36.42**	36.74	24.87	29.74	26.22	43.45	-13.17	27.74	42.38
GLM 5.3	29.60	26.52	14.15	29.02	34.06	34.13	21.70	37.87	27.46	11.71	32.47	40.40
Kimi K3	32.81	22.40	15.69	28.63	32.95	20.25	39.98	39.11	25.42	0.68	**34.76**	30.68
Claude Opus 5	24.96	-1.99	16.19	26.53	25.63	13.29	21.55	36.84	17.72	17.19	10.97	32.23
Grok 4.6	10.07	6.94	3.12	13.32	22.36	9.41	4.83	12.74	18.36	4.02	1.23	24.06
GPT 5.6 Sol	10.13	3.29	0.38	7.42	13.37	17.19	19.04	13.18	-13.85	-10.55	-8.61	28.92
Claude Fable 5	4.42	-9.03	-5.52	10.75	18.35	1.29	16.69	-1.97	-16.60	-6.66	-11.92	-1.77
DeepSeek V4 Flash	7.54	-4.19	-5.94	4.87	0.13	-7.66	2.16	6.19	13.23	-34.43	-14.25	35.98
GPT 5.5	10.93	0.50	-11.17	9.51	-6.94	-10.00	-1.35	-5.54	-11.05	-24.40	-3.36	7.95
Grok 4.5	-0.81	-7.20	-14.66	3.69	-1.33	-18.00	-6.24	7.81	-6.81	-3.71	-10.25	21.63
MiniMax M3	1.58	-8.77	-11.72	5.18	-1.23	-12.02	-16.55	1.81	-29.64	-36.32	-4.44	17.22
GLM 5.2	-0.82	-18.13	-15.74	-4.34	-13.67	-8.80	-14.28	-2.33	-33.58	-21.29	-4.11	-2.21
Gemini 3.6 Flash	-9.35	-8.98	-7.27	-15.77	-15.61	-20.98	-22.33	-5.90	-7.57	-15.17	-15.13	-5.74
Doubao Seed 2.1 Pro	-12.14	-16.53	-24.21	-3.04	-17.08	-14.97	-29.49	-11.32	-31.37	-35.30	-12.97	-20.53
DeepSeek V4 Pro	-14.38	-24.49	-26.18	-11.95	-19.40	-17.67	-13.02	-6.25	-31.29	-17.72	-32.85	-15.45
Qwen 3.7 Max	-9.93	-22.66	-28.87	-8.60	-20.39	-34.50	-26.09	-16.84	-34.89	-31.85	-28.60	-5.96
Gemini 3.7 Flash	-16.89	-18.76	-38.72	-18.37	-7.41	-15.19	-3.84	-37.38	-18.56	-58.43	-25.89	-36.87
HY3	-14.96	-30.27	-21.31	0.61	-13.04	-30.94	-33.90	-29.65	-49.96	-52.79	-23.02	9.05
Gemini 3.1 Pro	-9.78	-30.08	-31.01	-11.43	-22.65	-36.78	-41.43	-11.16	-30.37	-29.73	-36.66	-23.62
MiMo V2.5 Pro	-25.03	-34.93	-38.21	-18.56	-37.02	-45.73	-47.87	-20.37	-47.43	-55.30	-61.62	-38.41
Mean across models	**4.37**	**-5.26**	**-8.40**	**5.65**	**2.82**	**-5.65**	**-2.50**	**3.47**	**-8.16**	**-18.78**	**-7.94**	**9.11**

E.3 Case Studies and Failure Patterns

E.3.1 Failure Patterns across the Five Actions

Table 20 organizes observed trajectory and report failures by the five evaluated actions. Review failures concern cited supporting evidence, Verify failures concern factual and legal errors, and Deliver failures concern report completeness. Identify failures leave the legal issue incompletely established, while Assess failures propose responses that do not address it. Table 21 provides rubric-level examples for all five actions. Frequent reading and correct document formatting can coexist with failures in substantive conclusions, risk analysis, and remedies.

Table 20: **Observed failure patterns across the five actions.** The rows summarize failures in execution trajectories and submitted reports.

Action	Failure pattern	Observed behavior
Review	Untraceable supporting evidence	The report refers generically to certificates or records without identifying the supporting file.
Verify	Nonexistent citations; distractor adoption	Cited files are absent from the inventory, or misleading records enter the report as facts. High tool use can coexist with incorrect entity attribution.
Deliver	Missing or incomplete reports and conclusions	Reading produces no report; abbreviated reports omit required sections; repeated revisions retain omissions. A submitted report may also stop at identifier matching without a substantive legal conclusion.
Identify	Incomplete issue identification	The risk analysis omits a required supporting fact, even when that fact appears elsewhere in the report.
Assess	Remedies that address the wrong issue	The report proposes releasing patent pledges when the identified issue concerns co-ownership and consent restrictions.

E.3.2 Company A: Intellectual-Property Case Study

Table 21 attributes the examples in Figure 5 to the evaluated models and records their rubric-level scores. The excerpts are translated and condensed from Company A's IP sections under the revision-loop setting. The first three rows show recorded negative-rubric contributions; zero denotes that no penalty was recorded. Gemini 3.1 Pro receives zero on the source-naming rubric despite the unnamed-source excerpt, so the qualitative contrast and the recorded grading decision differ. The last two rows show points earned out of the three available for each rubric. Appendix F.4 provides the scoring conditions.

Table 21: **Model examples from Company A's IP section.** Each row compares outputs under the same rubric. Scores are the recorded rubric-level contributions: negative values are penalties; positive-rubric results are shown as earned points out of three.

Action and rubric	Weaker excerpt: model, output, and score	Stronger excerpt: model, output, and score
Review REVIEW-N014	**Gemini 3.1 Pro** (0)**.** Cites IP certificates and database records without naming a specific supporting file in the IP section.	**MiMo V2.5 Pro (0).** Names the *Joint R&D Agreement.pdf* and patent and copyright database tables, providing a traceable supporting source.
Verify VERIFY-N077	**Kimi K3** (-3)**.** Lists ZL2022XXX456.7 as issuer-owned while marking its validity as unverified. Adopts the distractor in both the patent list and risk analysis.	**Grok 4.6 (0).** Excludes the record from the issuer's patent portfolio. The selected distractor-adoption penalty is not triggered.
Deliver DELIVER-N014	**HY3** (-3)**.** Reports only that document and database identifiers match, without a substantive IP conclusion.	**DeepSeek V4 Flash (0).** Concludes that eighteen patents are valid, four are co-owned, and IP protection covers core LLM technologies.
Identify IDENTIFY-P009	**Claude Fable 5 (0/3).** Identifies four patents co-owned with University X and consent restrictions, but omits the portfolio total of eighteen from the risk analysis.	**Gemini 3.7 Flash (3/3).** Combines the eighteen-patent total, four co-owned patents, University X's identity, and restrictions on transfers and exclusive licensing in its risk analysis.
Assess ASSESS-P019	**HY3 (0/3).** Recommends repaying debt and releasing patent pledges, addressing a different issue from co-ownership.	**Claude Fable 5 (3/3).** Proposes confirming use and revenue-sharing arrangements with University X and obtaining an exclusive licence if needed.

From evidence to conclusions. Gemini 3.1 Pro's generic source references contrast with MiMo V2.5 Pro's named agreement. Kimi K3 expresses uncertainty about validity while still attributing the distractor to the issuer in two report sections; Grok 4.6 excludes it. A qualification about validity does not resolve incorrect ownership attribution. HY3's identifier matching likewise stops short of the substantive IP conclusion supplied by DeepSeek V4 Flash.

Complete risk identification. Claude Fable 5 states the eighteen-patent total elsewhere in its report but omits it from the risk analysis. The rubric requires the risk and all four supporting facts together, so that analysis receives zero of three points. Gemini 3.7 Flash brings the facts together and earns all three. The difference is complete issue construction from available facts.

Remedies tied to the actual issue. HY3's proposed release of pledges leaves co-ownership unaddressed. Claude Fable 5 supplies a targeted remedy despite missing a required fact in its risk analysis, showing that these two requirements can diverge within one report. Figure 9 traces a separate Company B distractor into a model's legal analysis.

E.3.3 Company B: Distractor Adoption

Figure 9 contrasts the expected verification with an observed DeepSeek V4 Flash error in Company B's litigation section. The litigation summary in the workspace and arbitration award establish a capital-increase dispute involving Subsidiary B1 and Person X and others, with RMB 90 million at stake; the claim was upheld and enforcement is ongoing. A judicial-database search returns a constructed public-information record with the same parties and amount but a different dispute type (equity transfer) and status (arbitration application withdrawn). The agent should compare these fields, check the supporting documents, and flag the conflict; an additional case requires corroboration before inclusion. The archived tool response confirms retrieval of the altered record. The final report retains the correct case but calls the two records "two different cases" and adopts the

withdrawn arbitration as supplementary litigation information. The report therefore introduces an unsupported additional case; its recorded VERIFY-N054 score is zero.

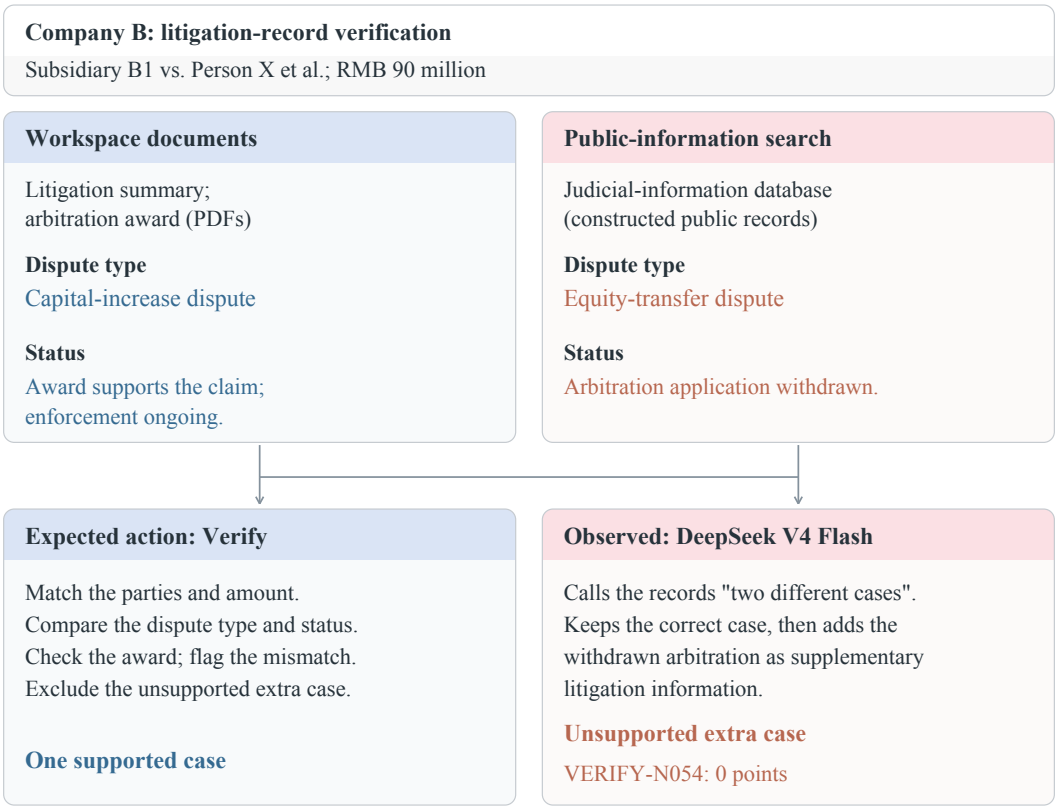


Figure 9: **A conflicting public record becomes an unsupported additional case.** The upper panels distinguish workspace documents from constructed public-information search results; the lower panels contrast expected verification with DeepSeek V4 Flash's observed report. The altered record is a benchmark distractor. Source and report text is translated and condensed.

F Worked Workspace: Company A

This example uses the Company A matter from Figure 5. We first describe the assignment and accessible materials, then list the full workspace, show selected document and database contents, present the corresponding rubrics, and trace connections across due-diligence sections. Names use anonymous English labels throughout; patent identifiers have their middle digits replaced by XXX consistently across documents, records, and rubrics.

F.1 Matter Overview

Company A develops large language models and is preparing for a Hong Kong Main Board IPO. Its workspace contains 192 PDFs totalling 539 pages. The model can also query six corporate databases containing fourteen tables and 258 records, including constructed distractors, and retrieve laws through the legal-search tools. The IP investigation concerns eighteen registered patents, four jointly owned with University X, alongside copyrights, trademarks, domains, and protective agreements.

The task is to investigate the requested modules and produce an LDD report under the instructions in Appendix G. The model must reconcile certificates, agreements, and external records, distinguish the issuer from other entities, and explain identified risks and appropriate responses. Rubrics and distractor annotations are evaluation materials; they are not part of the model's workspace.

F.2 Document Inventory and Accessible Resources

All PDFs sit in one flat document directory in the workspace. The tree lists every document as D001–D192, followed by all external database tables and the legal-retrieval interfaces. The document titles are anonymous English descriptions. Separate certificates, versions, and supplements remain separate files; the external branches denote resources queried through tools.

D017 is a co-owned patent certificate, D033 is the joint R&D agreement, and D054 is an employee IP agreement template. Their selected contents appear in Appendix F.3.

Company A: complete accessible workspace

Documents **Corporate records** **Legal materials**

```
Company A: agent-accessible materials
|-- task.md: LDD assignment and shared goal requirements
|-- dataroom/ [flat directory; 192 PDFs; 539 pages]
|   |-- D001 Business licence: registered-address check (variant
|       13).pdf
|   |-- D002 AI ethics guidelines.pdf
|   |-- D003 AI safety assessment.pdf
|   |-- D004 Board resolution on A-share listing.pdf
|   |-- D005 Series A investment agreement (2022-05).pdf
|   |-- D006 Series B1 investment agreement (2023-02).pdf
|   |-- D007 Series B2 investment agreement (2023-08).pdf
|   |-- D008 Series B3 investment agreement (2024-01).pdf
|   |-- D009 Series B4 investment agreement (2024-08).pdf
|   |-- D010 Series B5 investment agreement (2024-11).pdf
|   |-- D011 Series B6 investment agreement (2024-12 to 2025-05).pdf
|   |-- D012 ESG and strategy committee rules.pdf
|   |-- D013 ESG policy.pdf
|   |-- D014 ISO certificates: consolidated set.pdf
|   |-- D015 Patent certificate: low-resource model fine-tuning.pdf
|   |-- D016 Concert-party agreement (2023-03-01).pdf
|   |-- D017 Patent certificate: co-owned patent 1.pdf
|   |-- D018 Patent certificate: standard patent 2.pdf
|   |-- D019 CSRC filing notice (2025-12-15).pdf
|   |-- D020 Major customer contract: Customer A.pdf
|   |-- D021 Major customer contract: Customer B.pdf
|   |-- D022 Major customer contract: Customer C.pdf
|   |-- D023 Major customer contract: Customer D.pdf
|   |-- D024 Major customer contract: Customer D (enhanced
|       variant).pdf
|   |-- D025 Corporate income-tax return (2024).pdf
|   |-- D026 Subsidiary A12 valuation report (2023-09).pdf
|   |-- D027 Supplier contract: computing capacity.pdf
|   |-- D028 Articles of association: IPO amendment (2025-06-28).pdf
|   |-- D029 Articles of association: final pre-conversion version.pdf
|   |-- D030 Articles of association: pre-conversion (enhanced
|       variant).pdf
|   |-- D031 Articles of association: incorporation (2019-06-11).pdf
|   |-- D032 Housing provident-fund contributions (2025-06-30).pdf
|   |-- D033 Joint R&D agreement for co-owned patents.pdf
|   |-- D034 Related-party list (2025-12-22).pdf
|   |-- D035 Content moderation policy.pdf
|   |-- D036 Export-control compliance policy (2025-01).pdf
|   |-- D037 Trademark certificate: Brand A large model.pdf
|   |-- D038 Patent certificate: improving LLM instruction
|       following.pdf
|   |-- D039 Patent certificate: reinforcement-learning dialogue
|       optimization.pdf
|   |-- D040 Patent certificate: multimodal visual question
|       answering.pdf
|   |-- D041 Patent certificate: web-navigation AI-agent training.pdf
```

Company A: accessible workspace (continued)

```
|   |-- D042 Patent certificate: automatic code generation and
|       completion.pdf
|   |-- D043 Patent certificate: Transformer-based Chinese text
|       classification.pdf
|   |-- D044 Patent certificate: vector-retrieval question
|       answering.pdf
|   |-- D045 Patent certificate: attention-based text
|       summarization.pdf
|   |-- D046 Patent certificate: efficient mixture-of-experts
|       inference.pdf
|   |-- D047 Patent certificate: distributed LLM training
|       acceleration.pdf
|   |-- D048 Patent certificate: LLM calling-interface system.pdf
|   |-- D049 Trademark certificate: Brand A, Class 42.pdf
|   |-- D050 Trademark certificate: Brand A software, Class 9.pdf
|   |-- D051 Trademark certificate: Company A, Class 42.pdf
|   |-- D052 Shareholder S1 partnership agreement.pdf
|   |-- D053 Employee incentive plan (2025-06-05).pdf
|   |-- D054 Employee intellectual-property ownership agreement.pdf
|   |-- D055 Employee roster (2025-06-30).pdf
|   |-- D056 Patent certificate: pretrained-model generative
|       information extraction.pdf
|   |-- D057 Overseas subsidiary: Subsidiary A1 registration
|       certificate.pdf
|   |-- D058 Overseas subsidiary: Subsidiary A2 registration
|       (2023-11-23).pdf
|   |-- D059 Overseas subsidiary: Subsidiary A3 registration
|       (2022-01-06).pdf
|   |-- D060 Overseas subsidiary: Subsidiary A4 registration
|       certificate.pdf
|   |-- D061 Overseas subsidiary: Subsidiary A5 registration
|       (2023-09-21).pdf
|   |-- D062 Overseas-investment filings.pdf
|   |-- D063 Value-added telecom licence: Company A.pdf
|   |-- D064 Value-added telecom licence: Subsidiary A7.pdf
|   |-- D065 Angel investment agreement (2022-01).pdf
|   |-- D066 Subsidiary business licence: Subsidiary A6.pdf
|   |-- D067 Subsidiary business licence: Subsidiary A7.pdf
|   |-- D068 Subsidiary business licence: Subsidiary A8.pdf
|   |-- D069 Subsidiary business licence: Subsidiary A9.pdf
|   |-- D070 Subsidiary business licence: Subsidiary A10.pdf
|   |-- D071 Subsidiary business licence: Subsidiary A11.pdf
|   |-- D072 Subsidiary business licence: Subsidiary A12.pdf
|   |-- D073 Audit committee rules.pdf
|   |-- D074 Conversion-date audit report (2025-01-31).pdf
|   |-- D075 Customer affiliation statement.pdf
|   |-- D076 Registration change: Series A (2022-05).pdf
|   |-- D077 Registration change: Series B1 (2023-02).pdf
|   |-- D078 Registration change: Series B2 (2023-08).pdf
|   |-- D079 Registration change: Series B3 (2024-01).pdf
|   |-- D080 Registration change: Series B4 (2024-08).pdf
|   |-- D081 Registration change: Series B5 (2024-11).pdf
|   |-- D082 Registration change: Series B6 (2025-05).pdf
|   |-- D083 Registration change: angel round (2022-01).pdf
|   |-- D084 Registration change: Subsidiary A12 acquisition
|       (2023-09).pdf
|   |-- D085 Registration change: equity transfer (2021-08).pdf
|   |-- D086 Company C contract.pdf
|   |-- D087 Nomination committee rules.pdf
|   |-- D088 Revenue breakdown by customer.pdf
|   |-- D089 Corporate conversion plan (2025-03-26).pdf
|   |-- D090 Government grants: consolidated documents.pdf
```

Company A: accessible workspace (continued)

```
|   |-- D091 Third-party supplier data-processing agreement.pdf
|   |-- D092 Data-security policy.pdf
|   |-- D093 Patent certificate: knowledge-graph entity and relation
|       extraction.pdf
|   |-- D094 Patent certificate: knowledge-enhanced language-model
|       pretraining.pdf
|   |-- D095 Patent certificate: multi-turn dialogue context
|       understanding.pdf
|   |-- D096 Patent certificate: mathematical-formula extraction from
|       text.pdf
|   |-- D097 Patent certificate: language-model safety alignment.pdf
|   |-- D098 Trademark certificate: Company A, Class 9.pdf
|   |-- D099 Domain registration certificate: Domain A1.pdf
|   |-- D100 Domain registration certificate: Domain A2.pdf
|   |-- D101 Domain registration certificate: Domain A3.pdf
|   |-- D102 Domain registration certificate: Domain A4.pdf
|   |-- D103 Domain registration certificate: Domain A5.pdf
|   |-- D104 Office/R&D lease: Site L01.pdf
|   |-- D105 Office/R&D lease: Site L02.pdf
|   |-- D106 Office/R&D lease: Site L03.pdf
|   |-- D107 Office/R&D lease: Site L04.pdf
|   |-- D108 Office/R&D lease: Site L05.pdf
|   |-- D109 Office/R&D lease: Site L06.pdf
|   |-- D110 Office/R&D lease: Site L07.pdf
|   |-- D111 Office/R&D lease: Site L08.pdf
|   |-- D112 Office/R&D lease: Site L09.pdf
|   |-- D113 Office/R&D lease: Site L10.pdf
|   |-- D114 Office/R&D lease: Site L11.pdf
|   |-- D115 Office/R&D lease: Site L12.pdf
|   |-- D116 Office/R&D lease: Site L13.pdf
|   |-- D117 Office/R&D lease: Site L14.pdf
|   |-- D118 Office/R&D lease: Site L15.pdf
|   |-- D119 Office/R&D lease: Site L16.pdf
|   |-- D120 Office/R&D lease: Site L17.pdf
|   |-- D121 Shareholder S2 partnership agreement (2021-06-23).pdf
|   |-- D122 Equity transfer: Shareholder S1 to Shareholder S2
|       (2021-08).pdf
|   |-- D123 Equity transfer: Shareholder S1 to Shareholder S3
|       (2021-08).pdf
|   |-- D124 Deep-synthesis management policy.pdf
|   |-- D125 Termination of special rights (2025-06).pdf
|   |-- D126 Independence statement.pdf
|   |-- D127 Independent non-executive directors' independence
|       confirmations.pdf
|   |-- D128 Cash-flow statement: management analysis.pdf
|   |-- D129 Generative-AI filings: consolidated certificates.pdf
|   |-- D130 User service agreement.pdf
|   |-- D131 User service agreement (enhanced variant).pdf
|   |-- D132 Shareholder S3 partner register (2025-12-30).pdf
|   |-- D133 Shareholder S3 partnership agreement (2021-06-23).pdf
|   |-- D134 R&D expense super-deduction filing.pdf
|   |-- D135 Social-insurance contributions (2025-06-30).pdf
|   |-- D136 Lease: property with title defect 1.pdf
|   |-- D137 Lease: property with title defect 2.pdf
|   |-- D138 Lease: property with title defect 3.pdf
|   |-- D139 Lease: property with title defect 4.pdf
|   |-- D140 Lease: property with title defect 5.pdf
|   |-- D141 Lease: standard property 2.pdf
|   |-- D142 Lease: standard property 3.pdf
|   |-- D143 Lease: standard property 4.pdf
|   |-- D144 Lease: standard property 5.pdf
|   |-- D145 Lease: standard property 6.pdf
```

Company A: accessible workspace (continued)

```
|   |-- D146 Lease: standard property with title certificate.pdf
|   |-- D147 Competition and conflict-of-interest statement.pdf
|   |-- D148 Executive Order 14105 compliance assessment (2025-02).pdf
|   |-- D149 Cybersecurity classified-protection filing.pdf
|   |-- D150 Alternative computing-capacity procurement plan
|       (2025-02).pdf
|   |-- D151 Algorithm filings: consolidated certificates.pdf
|   |-- D152 Amended concert-party agreement (2023-04-05).pdf
|   |-- D153 Audited financial statements (2022-2024).pdf
|   |-- D154 Shareholder resolution: corporate conversion
|       (2025-03-26).pdf
|   |-- D155 Shareholder resolution: Subsidiary A12 acquisition
|       (2023-09).pdf
|   |-- D156 Shareholder resolution: equity transfer (2021-08).pdf
|   |-- D157 Shareholder resolution: Hong Kong listing.pdf
|   |-- D158 Shareholders' agreement: special rights (2022-01).pdf
|   |-- D159 Shareholder register after Series A (2022-05).pdf
|   |-- D160 Shareholder register after Series B1 (2023-02).pdf
|   |-- D161 Shareholder register after angel round (2022-01).pdf
|   |-- D162 Shareholder register at incorporation (2019-06-11).pdf
|   |-- D163 General meeting resolution: share subdivision
|       (2025-06-28).pdf
|   |-- D164 Share-capital verification report.pdf
|   |-- D165 Share-capital history (2019-06-11 to 2025-12-22).pdf
|   |-- D166 Subsidiary A12 share-acquisition agreement
|       (2023-09-07).pdf
|   |-- D167 Business licence after conversion (2025-03-26).pdf
|   |-- D168 Business licence: registered-address check.pdf
|   |-- D169 Business licence: business scope.pdf
|   |-- D170 Business licence after share subdivision (2025-06-28).pdf
|   |-- D171 Business licence at incorporation (2019-06-11).pdf
|   |-- D172 Director-appointment resolutions: consolidated set.pdf
|   |-- D173 Board resolution: export-control compliance (2025-02).pdf
|   |-- D174 Board resolution: special rights (2022-01).pdf
|   |-- D175 Board resolution: termination of special rights
|       (2025-06).pdf
|   |-- D176 Board resolution: committee establishment
|       (2025-06-05).pdf
|   |-- D177 Directors' biographies and qualification statements.pdf
|   |-- D178 Remuneration committee rules.pdf
|   |-- D179 Incorporation registration documents (2019-06-11).pdf
|   |-- D180 Subsidiary list (2025-12-22).pdf
|   |-- D181 Privacy policy.pdf
|   |-- D182 Capital verification: Series A (2022-05).pdf
|   |-- D183 Capital verification: Series B1 (2023-02).pdf
|   |-- D184 Capital verification: Series B2 (2023-08).pdf
|   |-- D185 Capital verification: Series B3 (2024-01).pdf
|   |-- D186 Capital verification: Series B4 (2024-08).pdf
|   |-- D187 Capital verification: Series B5 (2024-11).pdf
|   |-- D188 Capital verification: Series B6 (2025-05).pdf
|   |-- D189 Capital verification: angel round (2022-01).pdf
|   |-- D190 Capital verification: corporate conversion
|       (2025-03-26).pdf
|   |-- D191 Capital verification: incorporation (2019-06-11).pdf
|   `-- D192 High-tech enterprise certificate.pdf
|-- External corporate records [queried through tools; 258 records]
|   |-- DB1: Corporate registration
|   |   |-- Principal personnel [32 records]
|   |   |-- Basic company information [41 records]
|   |   |-- Registration changes [25 records]
|   |   |-- Shareholders and contributions [43 records]
|   |   `-- Administrative licences [37 records]
```

Company A: accessible workspace (continued)

```
|   |-- DB2: Patents and trademarks
|   |   |-- Patents [27 records]
|   |   `-- Trademarks [12 records]
|   |-- DB3: Copyright registrations
|   |   |-- Copyright in works [3 records]
|   |   `-- Software copyright [10 records]
|   |-- DB4: Domain registrations
|   |   `-- Domain registrations [9 records]
|   |-- DB5: Judicial records
|   |   |-- Enforcement records [5 records]
|   |   `-- Judicial decisions [6 records]
|   `-- DB6: Credit information
|       |-- Abnormal business operations [3 records]
|       `-- Administrative penalties [5 records]
`-- External legal materials [retrieved through law tools]
    |-- Statute search: laws applicable at the matter cutoff
    `-- Article retrieval: selected law, version, and article
```

F.3 Source Documents and External Records

The following boxes show translated selections from actual files and database rows in the Company A workspace. Document IDs refer to the complete tree above; the filenames use the same anonymous descriptions. Each example includes the source content and the specific fact that it can establish.

Workspace file: D017 — co-owned patent certificate

File. `D017 Patent certificate: co-owned patent 1.pdf`
Selected fields (page 1 of 2).

- Patent: knowledge-graph-based joint entity and relation extraction.
- Patent number: `ZL2022XXX789.3`.
- Rights holders: Company A; University X.
- Application date: 20 April 2022; grant date: 15 November 2023.
- Term stated in the certificate: twenty years from the application date.

What the file establishes. The certificate names both rights holders. The model must retain the joint ownership when matching this patent to DB2 and discussing ownership restrictions.

Workspace file: D033 — joint R&D agreement

File. `D033 Joint R&D agreement for co-owned patents.pdf`
Selected clauses (page 2 of 5).

- Clause 4.1 states a general rule of Company A ownership of jointly developed IP and complete commercial-use rights.
- Clause 4.2 permits joint ownership under a separate agreement when the collaboration mainly supports University X's research objectives.
- Clause 4.3 records four jointly owned patents at the agreement date.
- Clause 4.4 permits University X to use the joint patents and related results for academic publication and teaching, subject to prior written notice and confidentiality; commercial use is excluded from that permission.

What the file establishes. The general ownership clause has an explicit exception. The four joint patents cannot be described as solely owned by Company A by quoting Clause 4.1 alone.

Workspace file: D054 — employee IP agreement template

File. `D054 Employee intellectual-property ownership agreement.pdf`
Selected content (pages 1–2 of 4). The agreement covers confidentiality, non-competition, and IP ownership. The employee name and identification fields are blank. Clause 3.1 assigns to Company A the IP

created using its material or technical resources or while performing assigned work; Clause 3.3 requires the employee to assist with applications, registration, and transfer procedures.
What the file establishes. This is the content of the supplied agreement template. Blank employee fields do not establish that a particular employee signed it. The report should distinguish the stated IP policy from evidence of individual execution.

External records and constructed distractors. The external patent table contains both supported entries and deliberately misleading records. The examples below pair selected source fields with evaluator annotations explaining the conflicts. They illustrate three distinct checks: entity and identifier matching, ownership-history verification, and deduplication. The conflicts are established by cross-document evidence against the benchmark's constructed factual baseline.

External record: DB2 patent table — supported entry

Patent number: `ZL2022XXX789.3`
Name: Knowledge-graph-based joint entity and relation extraction.
Current rights holders: Company A; University X.
Acquisition: Original acquisition through joint R&D.
Status: Valid; jointly owned.
Cross-check: The patent number, holders, and dates agree with D017. D033 supplies the surrounding joint-ownership terms.

External distractor 1: near-matching number and company name

Selected DB2 fields. Patent `ZL2024XXX461.4`; title: a method for improving large-language-model instruction following; holder: Company A*.
Constructed conflict. The supported patent has the same title but number `ZL2024XXX460.7` and holder Company A. Company A* is an independent entity with a similar name and a different corporate registration identifier.
Required discrimination. Match the full patent number and the legal entity. A matching title or similar company name is insufficient to add a patent to the issuer's portfolio. Adopting this distractor is assessed by VERIFY-N070.

External distractor 2: invented transfer of a real patent

Selected DB2 fields. Patent `ZL2022XXX789.3`; current holder: University X's laboratory alone; acquisition: transfer from Company A.
Constructed conflict. D017 and the supported DB2 entry name Company A and University X as joint holders. The benchmark's factual baseline contains no such transfer.
Required discrimination. A correct patent number does not establish a correct ownership history. Reconcile the asserted transfer with the certificate and agreement before treating it as a change in title. Stating that Company A ceased to be a holder triggers VERIFY-N071.

External distractor 3: duplicate record inflates the portfolio

Selected DB2 fields. A second record repeats patent `ZL2024XXX460.7` and its instruction-following title.
Constructed conflict. The extra row represents the same patent. The supported portfolio contains eighteen registered patents, not nineteen.
Required discrimination. Deduplicate records by the patent identifier while resolving holder names. Counting both rows as independent patents, or increasing the portfolio to nineteen on that basis, triggers VERIFY-N078.

F.4 Rubric Examples

The eight rubrics below are selected from this matter's rubric set. They are grouped by Review, Verify, Deliver, Identify, and Assess. Five correspond to the actions illustrated in Figure 5; three additional Verify rubrics assess the external-record errors shown above. English renderings retain the rubric IDs, signed weights, and scoring conditions. Triggered negative rubrics deduct their stated weight; satisfied positive rubrics earn their stated weight. Otherwise, the contribution is zero.

Rubric examples: intellectual-property investigation

1. **REVIEW-N014 — Review (−3).** The IP verification-basis discussion must name concrete supporting files, such as patent certificates, D033, D054, or corporate database records. Deduct three points if the discussion is absent, is only a heading, or names no specific file. A generic reference to "IP documents" is insufficient; naming a specific supporting file avoids this particular penalty.

2. **VERIFY-N070 — Verify (−3).** Deduct three points when the report adopts `ZL2024XXX461.4` as an issuer patent, or assigns that number and Company A* as the holder of the instruction-following patent. The supported number is `ZL2024XXX460.7`, held by Company A.

3. **VERIFY-N071 — Verify (−3).** Deduct three points when the report states that `ZL2022XXX789.3` was transferred to University X or its laboratory as sole holder and that Company A is no longer a holder. The supported fact is continuing joint ownership, with no such transfer.

4. **VERIFY-N077 — Verify (−3 per occurrence).** Do not list patent `ZL2022XXX456.7`, a purported simple text-classification utility model, as Company A's patent, or infer that Company A holds utility-model patents from it. The rubric specifies eighteen registered invention patents and excludes this number. Each adoption of the distractor triggers the deduction.

5. **VERIFY-N078 — Verify (−3).** Deduct three points when the report counts `ZL2024XXX460.7` twice as independent patents, or uses the duplicate to increase Company A's portfolio to nineteen patents. The two records refer to one patent; the total remains eighteen.

6. **DELIVER-N014 — Deliver (−3).** Provide a substantive IP verification conclusion addressing the company's IP system or protective measures. The rubric describes eighteen registered patents, five copyrights, five trademarks, five domains, and protective arrangements. Deduct three points if the conclusion is absent, consists only of a heading, or makes no substantive judgment. Merely saying that identifiers match is insufficient.

7. **IDENTIFY-P009 — Identify (+3).** Identify the joint-ownership risk and state all four required facts: Company A has eighteen registered patents in China; four are jointly owned; the co-owner is University X through its knowledge-engineering laboratory; and use, licensing, or transfer requires all co-owners' consent unless otherwise agreed, creating uncertainty or restrictions on use. Award three points only if both the risk and all four facts are present.

8. **ASSESS-P019 — Assess (+3).** Recommend at least one substantive measure addressing joint ownership: inspect the four patents' R&D agreements to confirm title and use rights; determine whether they cover core technology or products; negotiate exclusive use or acquire the co-owner's share if they do; or improve academic-collaboration agreements to clarify ownership and prioritize exclusive company rights. Any one qualifying measure, including an equivalent targeted remedy, earns three points.

F.5 Investigations across Due-Diligence Sections

Company A's workspace connects due-diligence sections through shared entities, personnel, assets, and contractual conditions. These connections require repeated movement between sections: a customer relationship leads back to financing and director affiliations; a later agreement changes the interpretation of an earlier investor right; and a licensing finding carries forward into the review of the proposed share structure. Agents must retain these links and revisit earlier findings when another document changes their significance. Table 22 traces five such investigations through specific documents and pages in the workspace tree. Each row separates the facts established by the materials from the further checks and report conclusions that those facts inform.

Table 22: **Connections across Company A's due-diligence sections.** Shared entities and contractual conditions link separate parts of the investigation. Document IDs refer to the complete workspace tree in Appendix F.2; page numbers refer to the source PDFs.

Connected sections	Documents and established facts	Revisiting and connecting findings
Financing, customers, and related-party transactions	**D006: Series B1 investment agreement** (pp. 1–2) records a RMB 100 million contribution by an institution. **D075: customer-relationship statement** (p. 1) identifies the same institution as a model-service customer and the employer of a non-executive director. **D034: related-party list** (pp. 1–2) records these affiliations. **D088: customer revenue breakdown** (p. 1) places the institution among the top five customers. D075 also describes pricing review, director recusal, and disclosure procedures.	Start with the customer list, return to the financing records and director affiliations to establish the counterparty's overlapping roles, then revisit the sales relationship as a related-party transaction. The report must connect the commercial relationship to the applicable approval, conflict-management, and disclosure procedures. A customer-only summary would omit why these procedures matter.
Shareholding, personnel, and IP	**D132: shareholding-platform partner register** (pp. 1–2) lists 25 limited partners, including 16 full-time interns designated as consultants, with roles in model research and development. **D152: amended concert-party agreement** (pp. 1–2) includes the platform among the parties coordinating voting and other shareholder decisions. **D054: employee IP agreement** (pp. 1–2) is a template with blank individual fields; its assignment clauses address work produced during employment or using company resources.	Trace the platform from the shareholding arrangements to its individual contributors, then examine their work status and the agreements covering their technical output. The consultant designations make the scope and execution of the employee template a further question to investigate. Carry the resulting findings into both the platform-compliance and IP discussions, identifying any individual coverage that remains to be verified.
Investor rights, IPO progress, and liquidity	**D158: special-rights shareholders' agreement** (pp. 3–4) provides for redemption if an IPO is not completed by February 2028, with a price based on invested principal and 12% annual simple interest, less cash dividends. **D125: rights-termination agreement** (p. 2) terminates redemption rights at the initial listing application but restores them in specified circumstances, including withdrawal or rejection. **D128: cash-flow analysis** (pp. 1–2) reports monthly cash burn and estimates 27 months of runway assuming an unchanged burn rate.	Read the termination agreement back against the original redemption terms, then check the IPO stage to establish which provisions apply. Bring any resulting redemption exposure into the liquidity assessment alongside management's cash-burn assumptions. The title "rights-termination agreement" cannot settle the issue: its restoration conditions determine whether the earlier obligation can become relevant again.

Continued on the next page

Table 22: **Connections across Company A's due-diligence sections (continued).**

Connected sections	Documents and established facts	Revisiting and connecting findings
Business licensing, shareholding, and governance	**D063: telecommunications business licence** (p. 1) identifies the licensed activity as information services within the second category of value-added telecommunications services, limited to internet information services. **D028: IPO articles of association** (pp. 3–4) describes the domestic-share and H-share structure. Article 18 caps aggregate foreign holdings, including foreign holdings through H shares, at 50% and requires measures to maintain compliance with that ceiling.	Carry the licensed-business finding into the review of the proposed investor mix and share structure. Return to the licence and charter when assessing how ownership changes will be monitored and controlled. The report should connect business-qualification findings with shareholding and governance arrangements, rather than record the licence and the charter clause as unrelated items.
IP, R&D cooperation, and related-party transactions	**D033: joint R&D agreement** (pp. 1–2) establishes cooperation with University X. It provides for company ownership as the general rule, written agreements for joint-ownership exceptions, four jointly owned patents, and academic-use conditions for the university. **D034: related-party list** (p. 2) also identifies the laboratory as a related party and records co-founders' university appointments. The research counterparty therefore appears in both the IP and related-party materials.	Follow the shared patents back to the collaboration terms, then cross-check the collaborator against the related-party list and personnel affiliations. Return to the IP analysis to distinguish general ownership terms from the documented exceptions, while also examining governance and disclosure of the research relationship. The same collaboration must be understood consistently in both sections of the report.

G Task and Grading Prompts

This appendix collects the task instructions and the LLM grading prompt. Both were written and used in Chinese, and the versions below are English translations prepared for this paper. The task instructions contain the complete Company A assignment in English, including its embedded illustrations, all seventeen investigation modules, all fourteen file categories, and the eight shared investigation requirements. Names and identifiers use presentation aliases. The shared requirements appear under a Goal heading and are also supplied to the same-model goal checker. Appendix C.2 specifies the round limit, completion condition, and retained runtime settings.

G.1 Task Instructions

Full task instructions

Background. Company A (the "Company") proposes an initial public offering of overseas-listed shares (H shares) and listing on the Main Board of The Stock Exchange of Hong Kong Limited. As a legal professional at Firm X, you are commissioned by the Company to conduct preliminary legal due diligence on the Company and its related enterprises and issue an LDD report. The report will inform the Company and Firm X's assessment and planning of the IPO project.
Report date: 30 December 2025. This report date is the legal cutoff for determining which laws, versions, and provisions apply.
1. Investigation assignment. Review all supporting files supplied by the Company in the data room and search the corporate-information and legal databases. Comprehensively investigate: (1) legal status and corporate history; (2) share capital and its changes; (3) principal shareholders and actual controllers; (4) independence; (5) principal business and operating qualifications; (6) related-party transactions and competition; (7) principal assets—intellectual property; (8) principal assets—leased premises; (9) corporate governance and standardized operations; (10) directors, supervisors, and senior management; (11) tax

compliance; (12) employment, social insurance, and housing provident funds; (13) changes in principal customers and customer concentration; (14) overseas structure and cross-border compliance; (15) material acquisitions; (16) ESG governance and sustainability disclosures; and (17) attempted A-share listing and the subsequent move to a Hong Kong listing. Propose remediation for any legal-compliance risks or defects identified and deliver a complete LDD report.

2. Required report contents. The report must contain the following three parts.

Part I: Legal issues. Enumerate every legal issue identified. For each issue include:

(a) **Issue overview:** summarize the issue in one sentence.

(b) **Detailed description:** explain the issue and its potential effects on Company A's listing, cite specific facts and legal grounds, and provide legal analysis and a conclusion where legal questions arise.

(c) **Impact on the listing process:** classify the impact as major, moderate, or minor, with reasons and supporting grounds.
Illustration: Moderate. (1) The tax dispute concerns approximately RMB 4 million in additional tax and late-payment charges, about 1.5% of the latest audited net assets of RMB 269,440,267.01; the ratio is below 10% and the absolute amount below RMB 5 million. (2) It arose from an industry-wide Indian tax reform rather than an isolated violation by the illustrative Indian subsidiary. (3) Central excise duty ceased following the Indian GST reform on 1 July 2017, so the risk is not continuing. (4) However, the dispute remains unresolved and the arbitration committee has not issued its review decision, creating uncertainty.

(d) **Relevant facts and/or legal grounds:** cite the specific information in the corresponding data-room files for factual grounds; cite the particular statutory provisions and their content for legal grounds.
Illustration: Article 11 of the *Measures for the Administration of Individual Income Tax on Income from Equity Transfers (Trial)* (State Taxation Administration Announcement No. 67 of 2014) provides that the competent tax authority may assess equity-transfer income where, among other specified circumstances, the declared income is manifestly low without a justified reason.

(e) **Remediation:** propose clear, executable measures, such as amending the articles of association or obtaining a supplementary written confirmation from Company A.

Part II: Company investigation. For each requested investigation module record:

(a) **Verified facts:** detail the specific facts confirmed.
Illustration: Company A's legal status. Company A is a joint-stock company converted from its limited-liability predecessor and registered with the Chaoyang District market-regulation authority in Beijing on 26 March 2025. It holds a business licence with Registration ID A. Eight promoters, including Shareholder S1, Person P1, and Person P2, established the company. Registered capital is RMB 40,281,069, comprising 402,810,690 shares with a par value of RMB 0.10 each. Person P1 is its legal representative. Its business scope includes technology development, promotion, transfer, consulting and services; software development; information-system integration; data processing and storage support; and AI basic-software development.

(b) **Verification process:** detail the review performed, the main content and findings in the documents, and comparisons and discrepancies between supporting files and database searches.
Illustration: The legal team reviewed the complete corporate-registration archive since incorporation, inspected originals of the current business licence and other government-issued documents relevant to legal status, and examined the articles of association and related shareholder resolutions in writing.

(c) **Verification basis:** list the supporting materials for the entire module, including filenames and identifiers corresponding to the data room.

(d) **Verification conclusion:** analyze the module as a whole and state a clear conclusion; list individual problems or remediation needs separately.
Illustrations: (i) The legal team concludes that Company A is a duly established and validly existing joint-stock company with no circumstance requiring termination under applicable laws, regulations, normative documents, or its articles. (ii) A stated item in the business scope recorded on the licence differs from actual operations; describe the discrepancy and recommend a specific amendment to the relevant article, followed by submission to the shareholders' meeting for approval.

Part III: Supporting-file inventory. Classify every supporting file supplied in the data room by filename and substantive content; where they differ, use the content. Output an inventory containing the first-level category and filenames. Within each category, name and number the files according to their substantive content. Use these fourteen first-level categories:

1. Company basics: business licence, articles, organization chart, and related files.
2. Ownership structure and shareholders: shareholder register and licences, share-transfer agreements, and shareholder resolutions concerning ownership changes.

3. Corporate governance: shareholder, board, and supervisory-board meeting materials and internal management rules.
4. Financing rounds: contribution and capital-increase agreements and supporting payment records.
5. Business and operating qualifications: operating licences, contracts, and related materials.
6. Related-party transactions: related-party lists, contracts, policies, and resolutions.
7. Intellectual property: trademarks, patents, software copyrights, and other IP materials.
8. Finance and taxation: audit reports, capital-verification reports, tax-payment records, and related files.
9. Principal assets: property-title certificates, leases, and other tangible-asset materials.
10. Employment: employee rosters, social-insurance records, employment contracts, and related files.
11. Environmental protection and production safety: discharge permits, environmental expenditure details, and environmental and safety policies.
12. Corporate history: resolutions, contracts, valuation reports, and other materials concerning changes since incorporation.
13. Special compliance matters: industry- or background-specific compliance materials, such as valuation filings involving state-owned assets.
14. Other: files not falling within categories 1–13.

3. Formatting requirements.

(a) Use a clear, consistent heading hierarchy, such as Part I/Part II—I, II, III—1.1, 1.2, or an equivalent hierarchy, with headings emphasized in bold.

(b) Use consistent numbering throughout; number risk issues and investigation modules continuously and uniformly.

(c) Present more than five items of the same kind in a list or table; connect tables appropriately to the narrative and align their columns.

(d) Include necessary professional-risk limitations, such as restricting the report's purpose and stating that it is based only on materials obtained by the report date.

Goal: shared investigation requirements

1. Review the supporting files for every requested investigation module. Describe detailed facts, the verification process, and specific contents from the data room and corporate databases. For patents, state names, numbers, grant dates, and origins; for each equity transfer, state its date, quantity, amount, transferor, and transferee.
2. Report database findings on corporate registration, intellectual property, administrative penalties, litigation, and arbitration. Compare them item by item with supporting documents, label the sources, and retain relevant information from both sources.
3. State the full names of intellectual-property rights holders to distinguish the issuer from entities with identical or similar names.
4. State validity periods for patents and trademarks. Give event dates and explain why penalties, litigation, and arbitration are included, such as occurrence within the preceding three years or an older material matter involving at least RMB 500,000.
5. Explain all legal compliance risks and defects in the legal-issues section.
6. Retrieve cited laws to verify their names, versions, and provisions. State the law's name and version, article number, and the relevant provision or its summary.
7. Organize the report into legal issues, company investigation, and a supporting-file appendix. Embed the file list in the report rather than submitting it separately.
8. Submit only the final report, without intermediate drafts.

The goal checker receives the shared Goal requirements and the collected report in a separate call to the evaluated model. It supplies itemwise completion decisions and remaining gaps. The eight paragraphs specify the shared investigation requirements. The task requires verification of cited law names, versions, and provisions against each matter's report date, using the tools in Table 12.

G.2 LLM Grading Prompt

The following is an English translation of the rubric-grading prompt. Each call evaluates 20 rubrics from one action against the report; the final batch contains fewer than 20 when necessary (Appendix C.3). Braced fields are filled with the current action, reference material, batch identifiers,

rubric statements, report, and JSON output template. The common instructions and batch input are shown separately below.

Rubric grading: common instructions

You are the judge model for an IPO legal due diligence evaluation project. Your task is to evaluate the legal due diligence report produced by the model under evaluation, rubric by rubric, strictly according to the supplied rubric criteria.

Core judging principles

1. **Follow each rubric's criterion (`statement`) strictly.** Each rubric is self-contained: all correct values and lists of cases needed for the decision are included in its `statement`.
2. **Boolean decisions.** For a positive rubric, decide whether it is satisfied (`satisfied=true`) or not (`false`); for a negative rubric, decide whether it is triggered (`triggered=true`) or not (`false`).
3. **Trigger counts.** When a negative rubric is triggered, provide `trigger_count` according to that rubric's counting rule. For a rubric that counts only once, enter 1 when triggered.
4. **Positive and negative decisions are independent.** Failure to satisfy a positive rubric does not imply that a negative rubric is triggered. Evaluate the two independently.
5. **No discretionary credit.** Return only `true` or `false`; do not award partial satisfaction or make discretionary compromises.
6. **Evidence first.** Base each decision on the report text and quote relevant passages in `reason`. If no corresponding content is found, mark a positive rubric as unsatisfied and a negative rubric as not triggered, unless that negative rubric specifically targets missing or undisclosed content.
7. **Evaluate only the supplied IDs.** Do not evaluate IDs absent from the current rubric list, and do not omit any rubric in the list.

Output requirements

Return the grading results strictly as JSON. Output only JSON and no other content.

Rubric grading: batch input

Current evaluation action: {action ID}-{action name}
{action-specific instructions}
{additional reference material}
{current batch and rubric IDs to evaluate}
Rubrics to evaluate ({N} in total)
{criteria for the rubrics in this batch}
Report produced by the model under evaluation
{report text}
Output format
{JSON output template}
Evaluate every rubric above individually, {N} in total, without omitting any.
Requirements for `reason`: Include the complete reasoning process and quote evidence from the report. For negative rubrics, explain the basis for counting triggers.
Requirements for positive-rubric `evidence_quote`: When `satisfied=true`, provide a verbatim passage from the report. Do not paraphrase or join separate excerpts with ellipses.