FINRISKATLAS: A DECISION-ALIGNED BENCHMARK FOR RISK REVIEW AND EVIDENCE SEEKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Financial risk review requires more than knowledge of financial rules: a model must produce supported outputs from available evidence and decide when to seek additional information. We introduce FinRiskAtlas, an expert-reviewed Chinese-language benchmark for evaluating these complementary capabilities. Its static component contains 9,742 instances across 42 knowledge families and eleven review operations. Each operation is specified before instance construction by its visible evidence, decision object, required output, and scoring rule. FinRisk-Ask complements these tasks with offline evaluation of 680 pre-action states reconstructed from 104 professional review trajectories. Models receive only the record and history available before the reviewer acted, choose Ask or Proceed, and generate one focused evidence request when asking. Evaluation separates agreement with recorded reviewer actions from request alignment with expert-verified evidence needs. Later observations support target verification but remain hidden during inference. Across 33 model configurations, the knowledge leader attains the highest reported score on only one of the eleven review operations; selecting it for information extraction leaves an 18.01-point gap to the operation leader. On FinRisk-Ask, well-targeted requests can coexist with weak agreement on recorded decisions to proceed. FinRiskAtlas provides operation-specific profiles and evidence-seeking diagnostics for comparing models across review tasks.

1 INTRODUCTION

Financial risk and compliance review involves several distinct uses of evidence. Given a transaction record, a reviewer may need to extract contract terms, identify applicable provisions, or draft a supported review opinion. These operations can share the same record while requiring different outputs and success criteria. The merchant example in Figure 1 illustrates a further choice: when ownership remains unverified, the next step may be to request documentation before advancing the review. Evaluating a large language model (LLM) for this work therefore requires assessing both its performance on the specified operation and its choice of whether and what additional evidence to request. We use decision-aligned to describe evaluation tied to a specific review operation or evidence-seeking action and the information available at that point in the case record.

Financial benchmarks evaluate numerical reasoning, broad financial capabilities, and Chinese financial knowledge (Chen et al., 2021; Xie et al., 2024; Nie et al., 2025). Recent work further examines open-domain financial search, verifiable reasoning, and professional deliverables (Hu et al., 2026; Xie et al., 2026; Zhang et al., 2026). For financial risk review, a complementary question is whether a configuration selected for strong domain knowledge also performs best on the required operation. Answering it requires preserving each operation's output requirements and scoring criteria.

Evidence acquisition introduces a second evaluation requirement. Performance on a fixed record does not directly test a model's choice to seek additional evidence. In financial risk review, a reviewer must decide whether to advance the current step and which evidence to request when information is missing. Prior work examines information seeking under incomplete inputs (Li et al., 2024; 2025; Shao et al., 2026) and learns proactive questioning policies from offline expert dialogues (Wei et al., 2026). Evaluation in risk review must connect these decisions to the information available before the reviewer acted. Action and request content require separate assessment because an Ask decision can still miss the needed evidence.

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

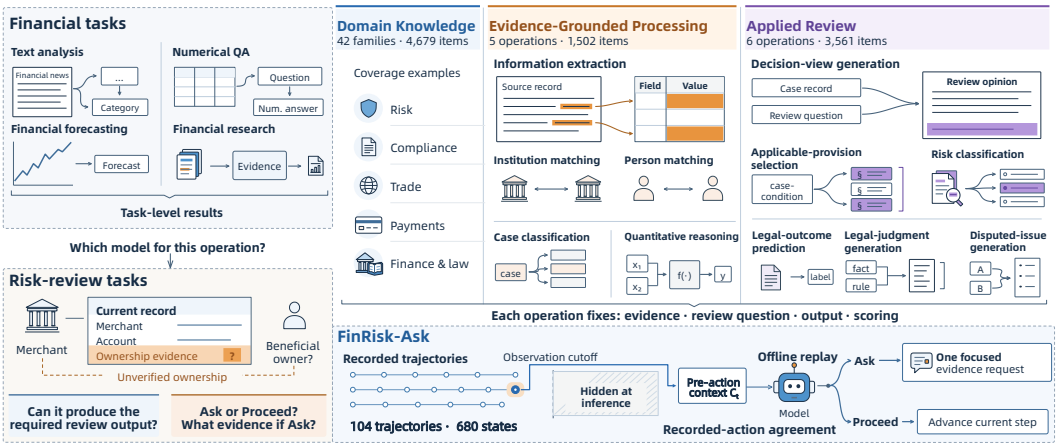

Figure 1: Motivation and overview of FinRiskAtlas. Left: connecting financial-task results to the required review output and the choice to seek evidence. Upper right: knowledge coverage and eleven fixed-evidence review operations. Lower right: FinRisk-Ask replays pre-action states; later information supports evidence-target verification while remaining hidden throughout inference.

We introduce FinRiskAtlas, an expert-reviewed Chinese-language benchmark that addresses these two requirements. Its static component takes the professional review operation as the unit of downstream evaluation. An operation contract specifies the visible evidence, decision object, required output, and scoring rule before instance construction. These contracts guide source selection, item construction, and scoring, tying each result to a defined reviewer output. The static benchmark contains 9,742 instances across 42 knowledge families and eleven review operations, organized into Domain Knowledge, Evidence-Grounded Processing, and Applied Review (Figure 1). The layers distinguish supporting knowledge, evidence processing, and case-level review, while each reported score remains tied to the scoring rule defined for its family. Its scope includes financial risk and compliance review and the legal analysis of financial disputes. Six domain experts defined the taxonomy and reviewed the source-grounded items; independent audits assess reference validity and evaluation reliability for the items released with the benchmark (Appendix D).

FinRisk-Ask complements these fixed-evidence tasks with offline replay of 680 pre-action states reconstructed from 104 de-identified enterprise review trajectories. At each state, the model receives only the record and history available before the reviewer acted. It chooses Ask to request additional evidence or Proceed to advance the current review step using the available record. An Ask response includes one focused evidence request. Actions are compared with the immediately recorded reviewer transition. Requests are assessed against expert-verified needs for information unavailable at that state, allowing alternative formulations that address the same need. Later observations support target verification but remain hidden during inference. This replay compares configurations at a common evidence boundary while separating action agreement from request targeting.

Evaluation of 33 model configurations shows why these distinctions matter. The configuration with the highest Domain Knowledge score attains the highest reported score on only one of the eleven review operations. Selecting it for information extraction leaves an 18.01-point gap to the strongest configuration for that operation. On FinRisk-Ask, models can generate well-targeted requests while frequently disagreeing with recorded decisions to proceed. These results support three contributions.

- **Operation-centric benchmark construction.** We introduce FinRiskAtlas, an expert-reviewed benchmark for financial knowledge and review operations. Contracts specify the visible evidence, decision object, required output, and scoring rule before instance construction.
- **Trajectory-grounded evidence-seeking evaluation.** We construct FinRisk-Ask for offline evaluation of actions and focused evidence requests at reconstructed pre-action states, using recorded reviewer actions and expert-verified evidence needs as separate references.
- **Empirical characterization of review capabilities.** We quantify operation-specific gaps from knowledge-based screening and show how action agreement and request alignment reveal different configuration strengths on the same review states and records under shared evidence boundaries.

2 Related Work

2.1 Financial LLM Evaluation

Financial LLM evaluation covers numerical reasoning, domain knowledge, and professional applications. FinQA, ConvFinQA, and TAT-QA evaluate reasoning over financial text, tables, and conversational context (Chen et al., 2021; 2022; Zhu et al., 2021). PIXIU and FinBen broaden coverage to financial language, prediction, and decision-making tasks (Xie et al., 2023; 2024), while CFinBench and FinEval provide evaluations tailored to the Chinese financial context (Nie et al., 2025; Guo et al., 2025). CNFinBench covers financial capability, compliance, and safety; FinGuard-Bench targets regulatory violations in model interactions (Ding et al., 2026; Dou et al., 2026).

Recent work also evaluates evidence retrieval and professional outputs. Finance Agent Benchmark and FinSearchComp assess tool-supported financial research and search (Bigeard et al., 2025; Hu et al., 2026). FinChain evaluates intermediate reasoning alongside final answers (Xie et al., 2026); BigFinanceBench scores auditable financial derivations, and FinProBench derives role-grounded rubrics from practitioner deliverables (Wang et al., 2026; Zhang et al., 2026). FinRiskAtlas complements these settings by evaluating distinct risk-review outputs under fixed evidence, with scoring rules specific to each review operation, and assessing evidence-seeking actions at reconstructed pre-action states, with later evidence withheld during inference.

2.2 Diagnostic Evaluation for Professional Tasks

In professional domains, LegalBench organizes tasks around forms of legal reasoning, LawBench distinguishes legal knowledge memorization, understanding, and application, and EHRStruct derives structured-record tasks from clinical needs (Guha et al., 2023; Fei et al., 2024; Yang et al., 2026). These approaches assess capabilities at a finer granularity than a domain-wide score. In general-purpose evaluation, CheckList uses capability-based behavioral tests, while HELM compares models across scenarios and metrics (Ribeiro et al., 2020; Liang et al., 2023). FinRiskAtlas applies this diagnostic principle through operation contracts specified before instance construction, keeping the evaluation unit consistent across source formats. We test whether knowledge-based screening retains the strongest configuration for each operation in the reported pool.

2.3 Information Acquisition under Incomplete Evidence

Clarification, selective answering, and deferral study how systems respond to ambiguous inputs or uncertainty about an answer (Zhang et al., 2024; Cole et al., 2023; Mozannar & Sontag, 2020). Information-seeking evaluations additionally examine which missing information to obtain. MediQ studies adaptive clinical questioning, QuestBench evaluates question selection for underspecified reasoning tasks, and RealFin examines omitted premises in financial questions (Li et al., 2024; 2025; Dai et al., 2026). SeekBench evaluates grounding, recovery, and evidence sufficiency along model-generated search traces and intermediate steps using retrieved evidence (Shao et al., 2026).

Learn-to-Ask trains and evaluates proactive dialogue policies for questioning and stopping using later observations in offline expert conversations (Wei et al., 2026). Inspired by this work, FinRisk-Ask evaluates whether to request evidence or advance the current financial review step, and what evidence to request. Action references come from the immediately recorded reviewer transitions. Experts verify that each request target addresses an unresolved need and concerns information absent from the pre-action context, with later observations supporting verification. The benchmark measures action agreement and request alignment separately across configurations evaluated at the same reconstructed review states.

3 FinRiskAtlas: Operation-Centric Benchmark Design

FinRiskAtlas evaluates Chinese-language financial risk and compliance review in two complementary settings: producing a required output from a fixed record and selecting an evidence-seeking action at a reconstructed review state. Both distinguish model-visible information from evaluation references (Figure 2). This section describes the static benchmark's evaluation units, coverage, and construction; Section 4 presents the trajectory-based FinRisk-Ask replay protocol.

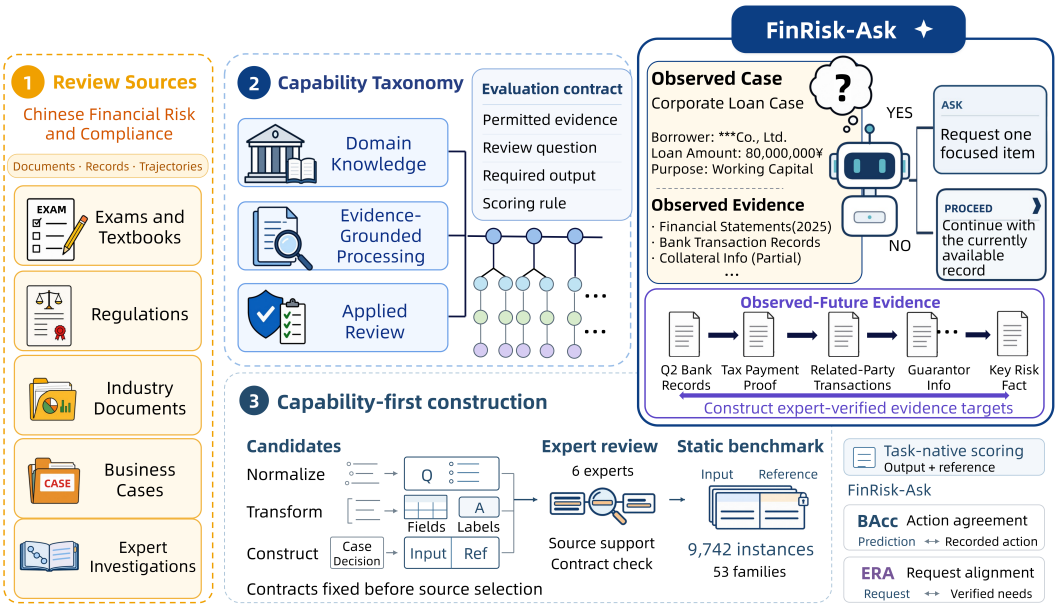

Figure 2: Construction and evaluation in FinRiskAtlas. Operation contracts guide source selection, instance construction, and scoring for static tasks. FinRisk-Ask uses pre-action contexts for inference and separate references for action and request evaluation. Later observations support evidence-target verification but remain hidden from candidate models throughout inference.

3.1 Evaluation Units and Operation Contracts

We define a review operation by the professional question it addresses and the output required from the permitted evidence. In the static benchmark, a family denotes either a knowledge area or one fixed-evidence review operation, and an instance is an item evaluated under that family's contract. The benchmark contains 42 knowledge families and eleven operation families. Capability layers group these families by review role; FinRisk-Ask states are counted separately.

Before instance construction, each family f is specified by an evaluation contract

$$\Gamma_f = (c_f, \mathcal{I}_f, d_f, \mathcal{Y}_f, s_f), \tag{1}$$

where c_f is the assessed capability, $\mathcal{I}_f$ specifies the sources, fields, and temporal boundaries visible to the model, d_f is the question or decision to be resolved, $\mathcal{Y}_f$ is the required output schema, and s_f is the parsing and scoring protocol. These specifications guide the construction of model inputs and references and the scoring of responses within each family under a fixed evidence boundary.

For information extraction, the contract fixes the requested fields and their order, with credit assigned at each field position. Applicable-provision selection requires a provision set and uses set-based scoring. Legal-outcome prediction uses the recorded disposition as its reference label while withholding it from the input. Different source formats can instantiate the same operation when they satisfy its contract. Tasks requiring different decisions or outputs remain separate even when they draw on the same record. Full operation contracts appear in Appendix B, together with each operation's native scoring rule and the task-specific output schema used to parse model responses.

3.2 Capability Taxonomy and Coverage

Domain experts defined three capability layers before candidate-model evaluation. The layers group capabilities by their role in review, from supporting knowledge through evidence processing to case-level outputs under the corresponding operation contracts.

Domain Knowledge contains 4,679 items in 42 families across sixteen taxonomy groups, covering financial risk, fraud, money laundering, compliance, trade and international finance, technology and

Table 1: Fixed-evidence review operations and scoring. For legal-judgment, disputed-issue, and decision-view generation, semantic rubrics jointly score coverage of the reference and factual correctness.

Operation	Required output	Scoring
Evidence-Grounded Processing		
Case classification	Routing label	Exact label agreement
Information extraction	Ordered field sequence	Field-level agreement
Institution matching	Institution identity decision	Exact label agreement
Person matching	Person identity decision	Exact label agreement
Quantitative reasoning	Normalized answer or option	Item-specific answer score
Applied Review		
Risk classification	Risk category	Exact label agreement
Applicable-provision selection	Provision set	Set-based partial credit
Legal-outcome prediction	Disposition label	Exact label agreement
Legal-judgment generation	Reasoned legal analysis	Coverage and correctness
Disputed-issue generation	Issue list	Coverage and correctness
Decision-view generation	Supported review opinion	Coverage and correctness

security, commerce and payments, and relevant law. These families assess the concepts and rules used to interpret supplied evidence across financial risk and compliance review settings.

Evidence-Grounded Processing contains five operations and 1,502 instances covering classification, extraction, identity matching, and computation over supplied records. Their outputs are structured labels, field values, identity decisions, or normalized quantitative answers used for scoring.

Applied Review contains six operations and 3,561 instances assessing case-level risk and legal conclusions and the generation of professional review outputs. Its scope includes financial and commercial disputes. Legal-judgment generation requires a reasoned analysis of a legal disposition, whereas decision-view generation requires a supported position on the stated review question for the particular case, based on the evidence available in the supplied case record.

The static benchmark totals 9,742 instances in 53 families. Table 1 summarizes the operation outputs and scoring rules. Extraction gives equal weight to each of the 1,005 requested fields across 249 instances, so instances with more fields contribute more to the aggregate score. For multiple-answer knowledge items and provision selection, an exact match to the reference set earns credit 1, a nonempty proper subset earns 0.5, and all other sets earn 0 for the evaluated item.

Domain Knowledge uses an unweighted macro-average over its 42 families. The five- and six-operation layer means provide descriptive overviews, while individual operation scores guide configuration comparison. Full inventories and scoring details appear in Appendices A and B, including the family inventories and the counting rules used here for score aggregation.

3.3 Source-Grounded Construction and Validation

Six domain experts defined the taxonomy, reviewed items, and adjudicated disagreements. The team comprised four experts in financial risk and compliance and two in the legal review of financial and commercial disputes. Construction proceeded from family specification to source selection, candidate preparation, and subsequent expert review for each family in the benchmark.

Sources include examinations and textbooks, regulatory documents and industry guidance, financial and legal records, and completed risk-review trajectories. Candidates follow three construction routes. Direct normalization standardizes source-supported questions and answer formats. Structured transformation converts records into classification, extraction, matching, or quantitative tasks. Source-conditioned construction pairs case evidence with review questions and references supported by documented decisions. In each route, the family contract fixes model-visible information and the material used to construct the operation-specific reference for scoring each constructed instance.

GPT-5.5 Extra High drafted Domain Knowledge candidates from the corresponding sources. Experts set the reference keys and checked each retained item against its cited material. Unsupported

keys, unresolved ambiguity, and unstable scoring criteria led to revision or removal. For record-grounded operations, references came from source documents or trajectories and were expert-verified; model assistance for these operations was limited to formatting and candidate-span identification. None of the evaluated configurations was used to construct benchmark items or reference labels. Family-level review checked consistency with the specified capability, information regime, decision object, output, and scoring protocol. Instance-level review checked source support, ambiguity, duplication, and provenance. Input checks removed answer-bearing metadata and information outside the permitted temporal boundary. Parsers were tested against expected schemas and formatting variants; reference answers are introduced only after each response is parsed.

Independent second-pass audits used prespecified pools and annotators blinded to the initial judgments. Agreement on contract instantiation was 93.7% ($\kappa = 0.81$) across 795 candidates. Source checks confirmed support for 95.8% of 400 sampled knowledge items, with failures corrected or removed before release. Reference-validity agreement ranged from 90.8% to 97.1% ($\kappa = 0.78$–0.89) across the audited structured-operation pools. Appendix D gives the audit populations, adjudication procedures, trajectory-reference results, and semantic-evaluator validation.

4 FinRisk-Ask: Evaluating Evidence-Seeking Decisions

FinRisk-Ask complements fixed-evidence operations by evaluating whether a model requests additional evidence or advances the current review step. Completed professional trajectories provide shared pre-action contexts and separate references for actions and requests. The evaluation replays each decision point from the information available immediately before the recorded action.

4.1 State Reconstruction and Reference Construction

For a decision point t, C_t contains the case record and history available immediately before the reviewer acted. Given C_t, the model chooses Ask to request additional evidence or Proceed to advance the current review step using the available record. The immediately recorded reviewer transition supplies the reference action a_t^*. Only C_t is shown; the reference action, later evidence, subsequent actions, and outcomes remain hidden during inference at every state.

A candidate state is retained when its pre-action context can be reconstructed, its next action maps unambiguously to Ask or Proceed, and no post-action information appears in the input. Each retained Ask state additionally requires a nonempty target set $\mathcal{E}_t^+$. Experts use later observations to verify that targets concern information absent at t, address unresolved needs, and constitute meaningful acquisition objectives. Targets represent evidence needs rather than request wording.

The evaluation set contains 680 states from 104 completed, de-identified enterprise review trajectories: 583 Ask and 97 Proceed. Institutional review, approval, and de-identification preceded inclusion in the evaluation set. All retained states support action evaluation, and all recorded Ask states support request evaluation. Appendix C details the reconstruction procedure, retained-state criteria, and eligibility checks applied to every candidate state before model evaluation begins.

4.2 Action and Request Evaluation

Models output either Proceed or Ask with one focused evidence request. We score the action choice and request content separately. Ask recall (AskR) and Proceed recall (ProceedR) are the percentages of matching predictions within the two recorded-action branches. Balanced Recorded-Action Agreement (BAcc) averages these recalls, giving the branches equal weight despite their different sizes. Invalid actions receive zero agreement credit. A parsed Ask action remains an Ask prediction even when the accompanying request is invalid and receives zero request credit.

On a recorded Ask state, the fixed semantic evaluator assigns credit 1 to a valid request that directly addresses at least one target in $\mathcal{E}_t^+$, 0.5 to relevant but incomplete alignment, and 0 otherwise. Proceed predictions, invalid actions, and invalid requests receive zero request credit. Evidence-Request Alignment (ERA) averages these credits over all recorded Ask states and scales the result to 0–100. It measures how consistently a configuration both selects Ask and requests relevant evidence. Ask predictions on recorded Proceed states affect action agreement but are not included in ERA.

Conditional Request Alignment (CRA) averages the same credits over recorded Ask states on which the model predicts Ask, including zero-credit invalid requests. It measures alignment within the subset selected by that configuration, while AskR measures coverage of the full recorded Ask branch. Within a repetition where CRA is defined, $\mathrm{ERA} = \mathrm{AskR} \times \mathrm{CRA}/100$. Reporting CRA alongside AskR distinguishes conditional request alignment from coverage when interpreting ERA. CRA is undefined when its conditional subset is empty because the model never predicts Ask.

All scores use a 0–100 scale and are computed within each repetition before averaging. Appendix C provides the full scoring and aggregation rules, and Appendix D reports validation of the fixed semantic evaluator against the adjudicated expert labels across all reported evaluation settings.

4.3 Reference Validation

Two experts who had not encoded the audited states judged the model-visible context C_t without access to later evidence or recorded actions. The audit covered all 97 Proceed states and 100 trajectory-stratified Ask states. After adjudication, the recorded action was accepted on 93.4% of the 197 states, with both actions acceptable on 8.1%. On this audited subset, restricting scoring to accepted states or crediting either action on dual-acceptable states preserves the BAcc and ERA leaders, with AskR remaining higher than ProceedR for every configuration. Appendix D details the audit and the subset-based sensitivity analyses reported for the reconstructed states.

5 Experiments

5.1 Experimental Setup

We evaluate 33 open-weight and API-served model configurations using zero-shot prompts that request the contract-specified output, without demonstrations or requested chain-of-thought. A configuration includes the model identifier, serving setup, and inference settings. Each operation uses its native parser and scorer, and all items remain in the scoring denominators, including empty or unresolved responses. Reported scores average the available repetition-level results. Appendix B describes the protocol and run metadata; Appendix E reports the full score matrix.

DeepSeek-V4-Flash is the fixed semantic evaluator for the three generation operations and FinRisk-Ask request alignment. Its own outputs are excluded from comparisons that use these judgments. Blinded validation against adjudicated expert labels yields linearly weighted κ_w of 0.73–0.80 and Spearman ρ of 0.82–0.87 across the four settings (Appendix D). Each structured operation admits all 33 configurations, while each generation operation admits 32. Analyses jointly involving all eleven operations or the FinRisk-Ask metrics use the same 32 non-evaluator configurations. Appendix Table 9 gives the complete eligibility rules for each comparison reported in Section 5.

Scores are reported on a 0–100 scale under task-specific criteria. Rankings, correlations, layer summaries, and shortlist gaps use the reported matrix at two-decimal precision. All leaders and ranks refer to point estimates within the eligible configuration pool. We examine variation across review operations (RQ1), performance excluded by knowledge-based screening (RQ2), and the relationship between action selection and request targeting (RQ3), using each comparison's eligible pool.

5.2 RQ1: How Does Performance Vary across Review Operations?

Different configurations attain the highest reported scores in the three layers: Gemini-3-Flash-preview scores 80.69 on Domain Knowledge, GPT-5.6-Sol scores 85.22 on Evidence-Grounded Processing, and Gemini-2.5-Flash scores 63.99 on Applied Review (Figure 3(a)). These means summarize performance under each layer's constituent scoring rules. Across the eleven operations, ten configurations attain the eligible score maxima, and Appendix Figure 4(a) marks these maxima in each column for direct operation-wise comparison across configurations.

Nearly identical Knowledge scores can also accompany different operation profiles. Gemini-3-Flash-preview and Kimi-K3 score 80.69 and 80.68 on Knowledge, respectively. Gemini-3-Flash-preview scores 14.10 points higher on applicable-provision selection, while Kimi-K3 scores 8.47 points higher on legal-outcome prediction (Appendix Figure 4(b)). Choosing between them therefore depends on the review operation and output required in the intended deployment setting.

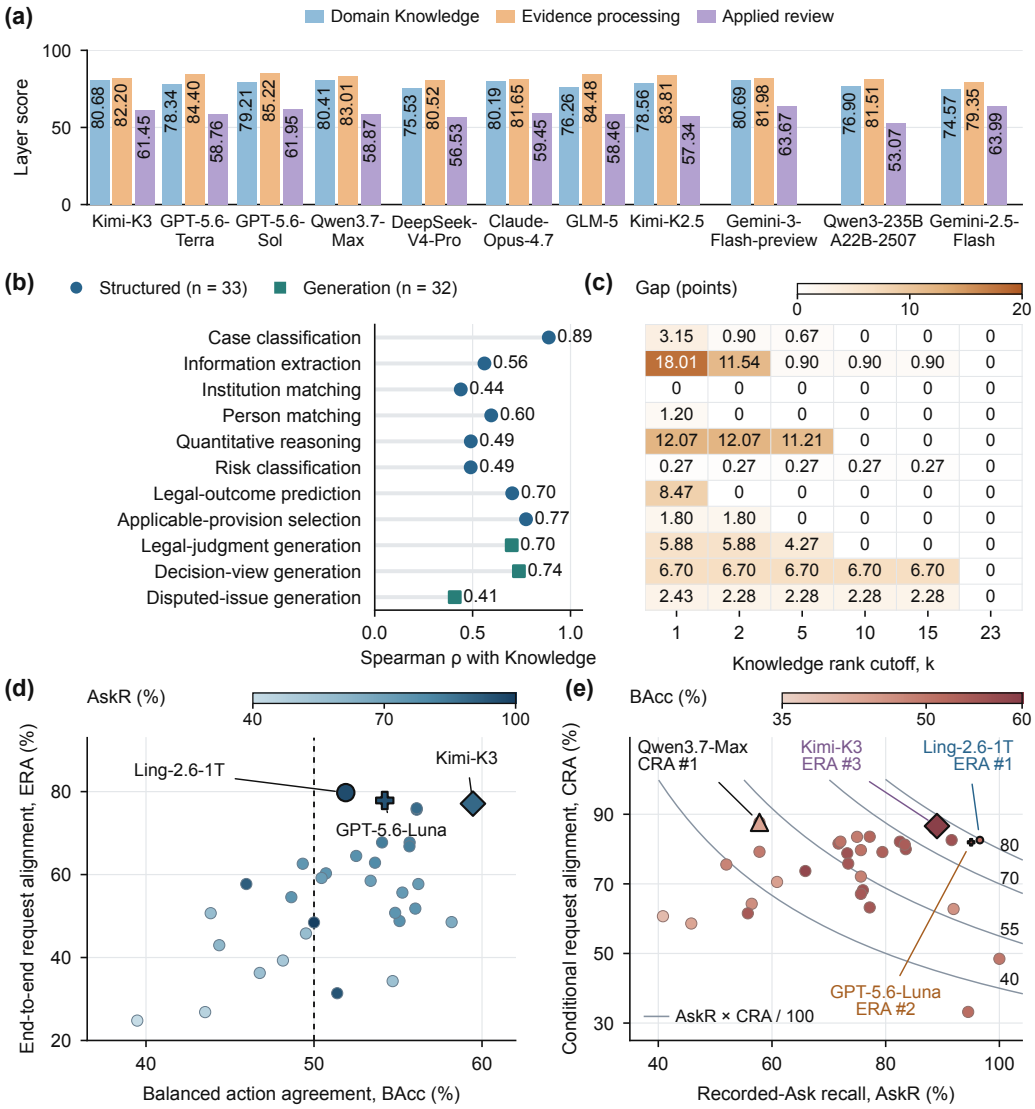

Figure 3: Operation performance and evidence-seeking behavior. (a) Layer summaries for eleven selected configurations. (b) Knowledge–operation Spearman correlations. (c) Score gaps at Knowledge-rank cutoffs with ties retained. (d) BAcc versus ERA, with BAcc 50 marking constant-action policies. (e) AskR versus CRA, with curves showing $\mathrm{AskR} \times \mathrm{CRA}/100$ reference values. Panels (b,c) use 33 configurations for structured operations and 32 for generation; panels (d,e) use 32 non-evaluator configurations under the shared inference, eligibility, and scoring settings.

Each operation retains its native scoring rule, so the layer means in Figure 3(a) summarize profiles within a layer. Leader counts and rank correlations compare configurations within the same operation, where the scoring rule and eligible pool are fixed. This supports operation-level comparison and preserves the task-specific interpretation of each reported score in the analysis.

Knowledge correlates positively with every operation in the evaluated pool, from 0.41 for disputed-issue generation to 0.89 for case classification (Figure 3(b)). Among the 32 complete non-evaluator profiles, the mean pairwise correlation of operation rankings is 0.42; 35 of 55 pairs are below 0.5. Appendix Figure 6(a,b) shows the knowledge-near profiles and rank spectrum. The spread reflects different output structures and error criteria across operation contracts. Domain Knowledge offers useful ordering information but does not identify the strongest configuration for each evaluated operation across the structured and generation tasks in the reported model pool.

5.3 RQ2: What Does Knowledge-Based Screening Exclude?

We measure the target-operation performance retained by a Knowledge-based shortlist. For operation o, let $\mathcal{M}_o$ be the eligible configuration pool, $S_{m,o}$ the reported operation score, and $r^K_{m,o}$ the descending competition rank by Knowledge. Tied Knowledge scores share a rank. The shortlist at cutoff k is $\mathcal{M}_o(k) = \{m \in \mathcal{M}_o : r^K_{m,o} \le k\}$, and its score gap is defined as follows:

$$\mathrm{Reg}_o(k) = \max_{m \in \mathcal{M}_o} S_{m,o} - \max_{m \in \mathcal{M}_o(k)} S_{m,o}. \tag{2}$$

The gap measures the target-operation performance excluded by screening, in the operation's native score units, even when the strongest retained configuration is selected from the shortlist.

For information extraction, the Knowledge leader, Gemini-3-Flash-preview, scores 67.96, compared with 85.97 for GPT-5.6-Luna. The gap is 18.01 points at $k = 1$ and decreases to 11.54 when Kimi-K3 is included at $k = 2$. Paired item-bootstrap 95% intervals are $[13.96, 21.79]$ and $[7.76, 15.71]$, respectively. Appendix B.6 describes paired bootstrap resampling over complete extraction items.

The legal-outcome gap decreases from 8.47 at $k = 1$ to zero at $k = 2$, while the quantitative-reasoning gap is 12.07 at both $k = 1$ and $k = 2$. Quantitative reasoning contains 58 items, and its paired-bootstrap 95% interval at $k = 1$ is $[5.17, 22.41]$ under paired resampling of items.

At wider cutoffs, extraction is within 0.90 points of its maximum by $k = 5$, whereas decision-view generation retains a 6.70-point gap at $k = 15$ (Figure 3(c)). Appendix F gives operation-leader ranks and recovered-maxima counts. Screening can therefore retain near-best performance while excluding the exact operation leader, so shortlists should remain operation-specific.

5.4 RQ3: How Do Action Selection and Request Targeting Differ?

Ling-2.6-1T leads ERA at 79.76 (BAcc 51.89), whereas Kimi-K3 leads BAcc at 59.46 (ERA 77.10; Figure 3(d)). Across 32 non-evaluator configurations, the scores correlate at 0.52 (Spearman). The different leaders show that request alignment on recorded Ask states and balanced action agreement capture distinct behavior on the shared replay states and should be reported separately.

Figure 3(e) separates Ask-state coverage from conditional request alignment. Qwen3.7-Max has the highest CRA, 87.67, but AskR of 57.80 and ERA of 50.69. Ling-2.6-1T combines CRA of 82.58 with AskR of 96.57 and attains the highest ERA. Qwen3.7-Max achieves higher conditional alignment on its selected subset of recorded Ask states, whereas Ling-2.6-1T requests evidence on a larger share of that branch. Reporting both CRA and AskR distinguishes these two aspects of evidence-request performance on Ask states at the fixed pre-action evidence boundary.

AskR exceeds ProceedR for all 32 configurations, with a median difference of 46.78 points. ProceedR is 7.22 for Ling-2.6-1T and 29.90 for Kimi-K3. The highest BAcc, 59.46, is 9.46 points above the 50.00 attained by either constant-action policy. Agreement with recorded continuation decisions remains limited even among configurations with high request alignment. Appendix Figure 5 gives branch recalls and request outcomes for all 32 non-evaluator configurations. This branch gap calls for considering ProceedR together with ERA when comparing action-selection configurations.

6 Conclusion

We introduce FinRiskAtlas, a Chinese-language benchmark for financial risk review and evidence seeking. Operation contracts connect available evidence to required outputs and scoring criteria, while FinRisk-Ask evaluates action choices and focused evidence requests at reconstructed pre-action states. Across the evaluated configurations, knowledge-based screening can exclude higher-scoring alternatives for a target operation, and well-targeted requests can coexist with weak agreement on recorded decisions to proceed. These findings support operation-specific configuration selection and separate assessment of action agreement and request targeting. Future work can extend offline replay to interactive evaluation, testing whether model-generated evidence requests improve subsequent review decisions once the requested evidence becomes available to the reviewer.

Reproducibility Statement

Evaluation protocols and experimental details appear in the appendix. The anonymous supplementary package reproduces the reported rankings, correlations, layer summaries, and shortlist gaps from the published score matrix. Appendix B states which further checks use the included scoring records.

Ethics Statement

FinRisk-Ask uses de-identified enterprise review trajectories under institutional approval. The benchmark is intended to evaluate financial review models in research settings.

AI Use Statement

AI tools assisted with writing, literature discovery, and synthetic data generation. The authors take full responsibility for the content, analyses, results, and claims in this paper.

References

Antoine Bigeard, Langston Nashold, Rayan Krishnan, and Shirley Wu. Finance agent benchmark: Benchmarking LLMs on real-world financial research tasks. arXiv preprint arXiv:2508.00828, 2025.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of EMNLP*, pp. 3697–3711, 2021.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of EMNLP*, pp. 6279–6292, 2022.

Jacob Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968.

Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. Selectively answering ambiguous questions. In *Proceedings of EMNLP*, pp. 530–543, 2023.

Yuyang Dai, Yan Lin, Zhuohan Xie, and Yuxia Wang. RealFin: How well do LLMs reason about finance when users leave things unsaid? In *Findings of ACL*, pp. 25050–25080, 2026.

Jinru Ding, Chao Ding, Yidong Jiang, Wenrao Pang, Boyi Xiao, Zhiqiang Liu, Jiayuan Chen, Yun Zhong, Tiantian Yuan, Junming Guan, Dawei Cheng, and Jie Xu. Beyond knowledge to agency: Evaluating expertise, autonomy, and integrity in finance with CNFinBench. In *Proceedings of ACM SIGKDD, Volume 2*, pp. 8765–8776, 2026.

Huaixia Dou, Jie Zhu, Minghao Wu, Shuo Jiang, Junhui Li, Lifan Guo, Feng Chen, and Chi Zhang. FinGuard: Detecting financial regulatory non-compliance in LLM interactions. arXiv preprint arXiv:2605.29427, 2026.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of EMNLP*, pp. 7933–7962, 2024.

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 *Advances in Neural Information Processing Systems*, volume 36, pp. 44123–44279, 2023.

Xin Guo, Haotian Xia, Zhaowei Liu, Hanyang Cao, Zhi Yang, Zhiqiang Liu, Sizhe Wang, Jinyi Niu, Chuqi Wang, Yanhui Wang, Xiaolong Liang, Xiaoming Huang, Bing Zhu, Zhongyu Wei, Yun Chen, Weining Shen, and Liwen Zhang. FinEval: A Chinese financial domain knowledge evaluation benchmark for large language models. In *Proceedings of NAACL-HLT*, pp. 6258–6292, 2025.

Liang Hu, Jianpeng Jiao, Jiashuo Liu, Dongyuan Mutu, Yanle Ren, Zhoufutu Wen, Kaiyuan Zhang, Xuanliang Zhang, Xiang Gao, Tianci He, et al. FinSearchComp: Towards a realistic, expert-level evaluation of financial search and reasoning. In *International Conference on Learning Representations*, pp. 46402–46428, 2026.

Belinda Z. Li, Been Kim, and Zi Wang. QuestBench: Can LLMs ask the right question to acquire information in reasoning tasks? In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. MediQ: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 28858–28888, 2024.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of EMNLP*, pp. 2511–2522, 2023.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of ICML*, pp. 7076–7087, 2020.

Ying Nie, Binwei Yan, Tianyu Guo, Hao Liu, Haoyu Wang, Wei He, Binfan Zheng, Weihao Wang, Qiang Li, Weijian Sun, Yunhe Wang, and Dacheng Tao. CFinBench: A comprehensive Chinese financial benchmark for large language models. In *Proceedings of NAACL-HLT*, pp. 876–891, 2025.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of ACL*, pp. 4902–4912, 2020.

Jiaqi Shao, Yuxiang Lin, Munish Prasad Lohani, Yufeng Miao, and Bing Luo. Do LLM agents know how to ground, recover, and assess? evaluating epistemic competence in information-seeking agents. In *International Conference on Learning Representations*, 2026.

Alex Wang, Georg Meinhardt, Jacob Katz, Joseph H. Kim, Pratyush K. Chaudhary, Chase Blagden, and Eric Xu. BigFinanceBench: A workflow-grounded benchmark for financial-research agents. arXiv preprint arXiv:2606.03829, 2026.

Fei Wei, Daoyuan Chen, Ce Wang, Yilun Huang, Yushuo Chen, Xuchen Pan, Yaliang Li, and Bolin Ding. Grounded in reality: Learning and deploying proactive LLM from offline logs. In *Proceedings of ICML*, 2026.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. PIXIU: A comprehensive benchmark, instruction dataset and large language model for finance. In *Advances in Neural Information Processing Systems*, volume 36, pp. 33469–33484, 2023.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, Zhiyang Deng, Yuechen Jiang, Zhiyuan Yao, Haohang Li, Yangyang Yu, Gang Hu, Jiajia Huang, Xiao-Yang Liu, Alejandro Lopez-Lira, Benyou Wang, Yanzhao Lai, Hao Wang, Min Peng, Sophia Ananiadou, and Jimin Huang. FinBen: A holistic financial benchmark for large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95716–95743, 2024.

Zhuohan Xie, Daniil Orel, Rushil Thareja, Dhruv Sahnan, Hachem Madmoun, Fan Zhang, Debopriyo Banerjee, Georgi Nenkov Georgiev, Xueqing Peng, Lingfei Qian, Jimin Huang, Jinyan Su, Aaryamonvikram Singh, Rui Xing, Rania Elbadry, Chen Xu, Haonan Li, Fajri Koto, Ivan Koychev, Tanmoy Chakraborty, Yuxia Wang, Salem Lahlou, Veselin Stoyanov, Sophia Ananiadou, and Preslav Nakov. FinChain: A symbolic benchmark for verifiable chain-of-thought financial reasoning. In *Proceedings of ACL*, pp. 14529–14553, 2026.

Xiao Yang, Xuejiao Zhao, and Zhiqi Shen. EHRStruct: A comprehensive benchmark framework for evaluating large language models on structured electronic health record tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(40):34340–34348, 2026.

Chi Zhang, Lifan Guo, Feng Chen, Kang Zhou, and Ben Wang. FinProBench: Evaluating financial AI agents with role-grounded rubrics derived from professional deliverables. arXiv preprint arXiv:2608.04077, 2026.

Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. CLAMBER: A benchmark of identifying and clarifying ambiguous information needs in large language models. In *Proceedings of ACL*, pp. 10746–10766, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of ACL-IJCNLP*, pp. 3277–3287, 2021.

A BENCHMARK CONSTRUCTION AND INVENTORY

This section expands the construction process in Section 3, documenting source preparation, reference verification, quality checks, and inventory. Appendix B specifies parsing and scoring, Appendix C details trajectory reconstruction, and Appendix D reports the independent audits.

A.1 CONTRACT-GUIDED INSTANCE CONSTRUCTION

Construction starts with the family contract defined in Equation 1. For a knowledge family, the contract identifies the concepts or rules to be assessed and the answer format. For a downstream operation, it specifies the review question, permitted evidence, required output, and scoring protocol. Experts fixed these specifications before instance construction. Family assignment follows the contract: heterogeneous records can instantiate the same operation, while one record can support several operations with different outputs or information boundaries.

Static candidates are prepared through direct normalization, structured transformation, or source-conditioned construction.

Direct normalization standardizes the stems, options, and answer formats of source-supported questions, including model-assisted knowledge candidates. Experts check that the normalized question preserves the source meaning and that its reference remains supported by the cited material.

Structured transformation converts financial, entity, and case records into classification, extraction, matching, or quantitative tasks. The transformation preserves the relationship between the visible record and the required answer. For extraction, the requested fields and their order determine the output positions. Matching tasks compare the supplied institution or person descriptions, while classification tasks specify the candidate categories. Quantitative tasks retain the supporting information and answer format required by the question.

Source-conditioned construction uses case materials and documented decisions to create risk and legal review tasks. Applicable-provision selection pairs case facts with candidate provisions and a verified reference set. Legal-outcome prediction uses a pre-disposition record as input and the recorded disposition as its label. Generation tasks pair the permitted case evidence and review question with a source-supported reference for a legal analysis, issue list, or review opinion. These references support semantic evaluation of the required professional outputs.

Across the three routes, each retained instance must satisfy its family contract and pass source checks. Source information links the constructed input and reference to their supporting material.

A.2 SOURCE GROUNDING AND EXPERT ROLES

The source material comprises examinations and textbooks, regulatory documents and industry guidance, financial and legal records, and completed enterprise review trajectories. Six domain experts defined the taxonomy, reviewed items, and adjudicated disagreements. The team comprised four experts in financial risk and compliance and two in the legal review of financial and commercial disputes.

The document source pool draws primarily on collections assembled for continued pretraining (CPT), including the Trade Document Collection and the Risk Document Collection. Their filenames carry the internal identifier 20251018. Externally retrieved materials supplement these collections. Legal records also include publicly available judicial materials obtained from resources of China's Supreme People's Court and selected by topic.

Topic selection determines the coverage of the legal materials, while the operation contract specifies the question and required output for each record. Representative subjects in case classification include lending, insurance, securities, shareholder relationships, payments, and trade or transportation disputes. Within this coverage, the benchmark distinguishes case classification, provision selection, disputed-issue identification, and case-level legal responses. Each case-classification item supplies candidate categories and requires an option label. Option letters identify positions within that item's candidate list; the category descriptions determine their meaning.

Table 2: Distinct professional outputs constructed from one case record. Existing Chinese questions and references are summarized in English, with parties identified by role.

Target output	Requested operation	Condensed reference content
Provision set	Select applicable provisions from the supplied candidates.	Provisions concerning the purchaser's payment obligation and civil procedure.
Issue list	Identify the contested questions requiring determination.	Whether the conditions for payment up to 95% have been met, and the amount payable.
Reasoned legal opinion	Analyze the payment obligations using the case record.	Interpretation of contractual payment milestones using settlement approvals and performance history.
Disposition label	Predict the recorded disposition.	Payment of outstanding purchase price and late-payment damages, together with the treatment of remaining claims.

For Domain Knowledge, GPT-5.5 Extra High drafted stems and options from the corresponding source materials. Experts set the reference keys and checked each retained item against its cited source. Unsupported keys, unresolved ambiguity, and unstable scoring criteria led to revision or removal.

References for Evidence-Grounded Processing and Applied Review came from source records or documented decisions and were expert-verified. Model assistance, where used, was limited to formatting and candidate-span identification. Recorded legal dispositions serve as outcome references, while generation references are assessed under the output requirements of their respective operations. No evaluated configuration was used to construct benchmark items or reference labels.

Table 2 illustrates how one commercial-dispute record supports distinct professional outputs. A supplier seeks payment under a staged concrete-supply contract, while the purchaser disputes whether the payment milestones have been reached. The record includes contract terms, delivery and payment history, party submissions, and evidentiary material. Each constructed entry pairs the permitted case evidence with an operation-specific question and reference, and is counted under that operation.

FinRisk-Ask uses completed, de-identified enterprise review trajectories. The pre-action record supplies the model-visible context, the immediately recorded transition supplies the action reference, and later observations support evidence targets for Ask states. Experts verify that each target concerns information unavailable at the reconstructed state and relevant to an unresolved need. Targets describe evidence needs rather than required request wording. Appendix C specifies state retention, action encoding, and target construction.

A.3 Quality Control and Information Boundaries

Family-level review checks whether candidates share the specified capability, information regime, decision object, output, and scoring protocol. Instance-level review checks source support, ambiguity, and provenance. References are verified against the relevant source documents, record transformations, or adjudicated expert decisions. Candidates with unresolved problems are revised or excluded. Every retained item is associated with source information and construction records.

Normalized inputs, source identifiers, and templates are examined for accidental duplication, taking account of the questions, outputs, and information boundaries when a source supports distinct operations. Input checks distinguish legitimate task evidence from information that reveals a reference label. Requested field values belong in an extraction record, whereas legal-outcome inputs exclude the later disposition and answer-bearing metadata. In FinRisk-Ask, the next recorded action, later evidence, and subsequent outcomes are withheld from the pre-action context. Pretraining overlap is discussed in Appendix H.

Deterministic parsers are tested with benchmark examples and formatting variants under the required output schemas. Parsing uses the response and task schema without consulting the reference answer; the reference enters only when the parsed output is scored. All retained items receive initial expert review. Independent second-pass audits use prespecified pools and annotators blinded to the initial

Table 3: Benchmark inventory by evaluation unit. FinRisk-Ask states come from 104 trajectories and are counted separately from static instances.

Component	Evaluation unit	Families	Count
Domain Knowledge	Static instance	42	4,679
Evidence-Grounded Processing	Static instance	5	1,502
Applied Review	Static instance	6	3,561
Static total	Static instance	53	9,742
FinRisk-Ask	Pre-action state	-	680

Table 4: Domain Knowledge inventory across sixteen taxonomy groups. Each row can aggregate multiple evaluation families.

Taxonomy group	Coverage	Families	Items
Financial risk	Risk management, markets, products, valuation, credit, operational, liquidity, and fixed-income risk	8	892
Fraud risk	Card, insurance, telecom, contract, invoice, and AI-enabled fraud	6	624
Money-laundering risk	Anti-money-laundering obligations and typologies	2	244
Compliance risk	Regulatory obligations and compliance controls	4	442
Trade theory and policy	International trade theory and policy	4	457
FDI and multinationals	Foreign direct investment and multinational operations	2	208
International finance	Cross-border finance and settlement	4	499
International business	Business environment and operations	4	438
FinTech	Financial technology	1	103
AI security	AI-related security risk	1	106
Technical security	Security controls	1	103
Security engineering	Security engineering practice	1	104
E-commerce	E-commerce risk and operations	1	139
Payments	Payment systems and controls	1	109
Finance law	Finance-related legal knowledge	1	110
Law	General legal knowledge for review	1	101
Total		42	4,679

judgments. Appendix D reports contract and reference audits, recorded-action acceptability checks, and semantic-evaluator validation with their respective sampling units and agreement statistics.

A.4 Inventory and Counting Conventions

Table 3 reports 9,742 static instances in 53 families and 680 FinRisk-Ask states from 104 trajectories. A source record can support multiple task-specific instances, and a trajectory can contribute multiple pre-action states. Static instances and trajectory states are counted separately. Family sizes reflect the available source material and the items retained for each capability.

Table 4 groups the 42 knowledge families into sixteen taxonomy groups. Each family receives equal weight in Equation 9, so a group's contribution to the Knowledge macro is determined by its family count rather than its item count.

Table 5 reports the downstream inventory. Paired rows combine counts for institution and person matching, legal-outcome prediction and legal-judgment generation, and disputed-issue and decision-view generation. Each operation retains its own contract and score. Information extraction comprises 249 instances with 1,005 requested fields, which are the scoring units. Quantitative reasoning contains 58 instances, the smallest static family. Full output and scoring contracts appear in Appendix B.

FinRisk-Ask contains 583 recorded Ask and 97 recorded Proceed states. These labels refer to the immediately recorded transition at each state. Multiple states from a trajectory share their source history but have separately reconstructed information boundaries. Their metric denominators and state weights are defined in Appendix C.

Table 5: Downstream inventory. Paired operations have joint item counts here but retain separate contracts and scores.

Operation or operation pair	Families	Items
Evidence-Grounded Processing		
Case classification	1	445
Information extraction	1	249
Institution and person matching	2	750
Quantitative reasoning	1	58
Layer total	5	1,502
Applied Review		
Risk classification	1	366
Applicable-provision selection	1	1,000
Legal-outcome prediction and legal-judgment generation	2	1,445
Disputed-issue generation and decision-view generation	2	750
Layer total	6	3,561
Downstream total	11	5,063

Table 6: Selected evaluation settings and their reference definitions. Entries summarize the aspects relevant to FinRiskAtlas.

Study	Evaluated output or behavior	Reference basis
FinChain (Xie et al., 2026)	Intermediate reasoning and final financial answers	Executable symbolic templates
FinSearchComp (Hu et al., 2026)	Financial search and reasoning	Expert-authored queries and task-specific answer criteria
Finance Agent (Bigeard et al., 2025)	Tool-supported financial research	Expert-authored questions and validated answers
BigFinanceBench (Wang et al., 2026)	Financial derivations	Reference answers and weighted step-level rubrics
FinProBench (Zhang et al., 2026)	Professional financial deliverables	Role-grounded rubrics derived from practitioner deliverables
EHRStruct (Yang et al., 2026)	Structured clinical-record tasks	Clinical task definitions and task-specific targets
MediQ (Li et al., 2024)	Interactive clinical questioning	Patient-record-based interactions and diagnostic answers
QuestBench (Li et al., 2025)	Clarification-question selection	Missing variable assignments in underspecified reasoning tasks
SeekBench (Shao et al., 2026)	Evidence use in model-generated search traces	Annotations of evidence and intermediate behavior
Learn-to-Ask (Wei et al., 2026)	Proactive questioning and dialogue termination	Future observations in offline expert dialogues
FinRiskAtlas static component	Financial knowledge and operation-specific outputs	Family contracts and source-supported references
FinRisk-Ask	Ask-or-Proceed actions and focused requests	Recorded reviewer transitions and verified evidence needs

A.5 Comparison of Evaluation Settings

Table 6 compares evaluation units and reference definitions for selected related settings. It covers the evaluation of completed outputs, behavior during information acquisition, and decisions at individual states. Section 2 discusses the broader literature.

FinRiskAtlas evaluates each static operation under its output contract. FinRisk-Ask compares configurations at shared pre-action contexts, using recorded local actions and verified evidence needs as separate references.

Table 7: Information and output contracts for the eleven fixed-evidence operations. Scoring rules are specified in Table 8.

Operation	Visible information	Decision object	Required output
Case classification	Case record and candidate categories	Case category	Routing label
Information extraction	Source record and ordered requested fields	Field values	Ordered field sequence
Institution matching	Two institution descriptions	Institution identity	Identity decision
Person matching	Two person records	Person identity	Identity decision
Quantitative reasoning	Question, supporting information, and options where applicable	Requested quantity or option	Normalized answer
Risk classification	Case evidence and candidate risk categories	Case risk category	Risk category
Applicable-provision selection	Case facts and candidate provisions	Governing provisions	Provision set
Legal-outcome prediction	Facts and procedural record, with the later disposition withheld	Recorded disposition	Disposition label
Legal-judgment generation	Facts, evidence, and procedural record	Reasoned disposition analysis	Legal analysis
Disputed-issue generation	Record with overlapping legal and contractual relationships	Contested issues	Issue list
Decision-view generation	Case record and review question	Review position	Review opinion

B Evaluation Protocol and Reproducibility

This section specifies how model responses become the scores used in Section 5. The protocol separates inference, reference-independent parsing, task-specific scoring, and aggregation. Appendix C defines the state-level action and request metrics, while Appendix D reports annotation reliability and semantic-evaluator validation.

B.1 Evaluation Contracts and Reference Interpretation

Table 7 specifies the visible information, decision object, and required output for each downstream operation. Together with the capability definitions in Section 3 and the scoring rules in Table 8, these specifications instantiate the family contract $\Gamma_f = (c_f, \mathcal{I}_f, d_f, \mathcal{Y}_f, s_f)$.

References are interpreted within these contracts. Legal-outcome prediction measures agreement with a recorded disposition, and risk classification measures agreement with the supplied category reference. Generation tasks assess a professional output against a source-supported reference under a semantic rubric.

B.2 Inference and Reference-Independent Parsing

A configuration comprises the model identifier, serving setup, and inference settings. Models receive a fixed task instruction and the contract-specified information, without demonstrations or requested chain-of-thought. The run manifest records the provider or checkpoint revision, decoding settings, output limit, timestamp, and retry status. Let $R_{m,f}$ denote the available repetition count for configuration m and family f. Scores are computed within each repetition and then averaged.

Transient API failures are retried. All items remain in the denominator of each scored repetition, including empty or unresolved responses. Empty answers and unresolved labels receive zero credit. Extraction instead applies credit independently at each requested field position, so a missing field does not discard correctly extracted fields. In FinRisk-Ask, a parsed Ask action remains an Ask prediction even when its request is invalid; the request then receives zero alignment credit. Appendix C specifies this separation.

Each family defines a prompt, output schema, deterministic parser, and scorer. The same parsing rules apply across configurations and use only the response and task schema. Extraction isolates the

Table 8: Native scoring and aggregation units. FinRisk-Ask denominators and metrics are defined in Appendix C.

Task	Scoring rule	Aggregation unit
Knowledge: single answer	Exact normalized label or option	Item
Knowledge: multiple answers	Exact set: 1; nonempty proper subset: 0.5; otherwise: 0	Item
Classification and matching	Exact label agreement	Item
Information extraction	Agreement at each requested field position	Field
Quantitative reasoning	Item-specific normalized answer score	Item
Applicable-provision selection	Exact set: 1; nonempty proper subset: 0.5; otherwise: 0	Item
Review generation	Harmonic mean of mean coverage and mean correctness	Family component means
FinRisk-Ask action	RAA, AskR, ProceedR, and BAcc	State or reference branch
FinRisk-Ask request	Set-level alignment credits; ERA, CRA, DEH, and REH	Eligible state subset

final answer and reads the prescribed field positions; risk classification isolates the answer-bearing text and resolves its explicit category. Reference answers enter only after parsing, when the resulting output is scored.

B.3 Task-Native Scoring

Let $\hat{y}_{m,i}^{(r)}$ be the response of configuration m to item i in repetition r, y_i the reference, and p_f the family parser. For families that average item credits $g_f \in [0, 1]$, the repetition-level score is

$$S_{m,f}^{(r)} = \frac{100}{N_f} \sum_{i=1}^{N_f} g_f \left(p_f(\hat{y}_{m,i}^{(r)}), y_i \right), \tag{3}$$

where N_f is the number of items in family f. The reported score averages the available repetition-level scores:

$$S_{m,f} = \frac{1}{R_{m,f}} \sum_{r=1}^{R_{m,f}} S_{m,f}^{(r)}. \tag{4}$$

All reported scores use a 0–100 scale. Extraction uses field-level aggregation instead of Equation 3; generation combines family-level coverage and correctness. Table 8 summarizes these distinctions.

Single-choice and true/false knowledge items receive exact-match credit after normalization. Case classification, institution matching, person matching, risk classification, and legal-outcome prediction likewise use exact label agreement. For knowledge multiple-answer items and applicable-provision selection, let $\hat{Y}$ be the parsed option set and Y the reference set. The credit is

$$g_{\text{set}}(\hat{Y}, Y) = \begin{cases} 1, & \hat{Y} = Y, \\ 0.5, & \varnothing \neq \hat{Y} \subsetneq Y, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Invalid or unresolved outputs receive zero credit. All nonempty proper subsets receive the same partial credit, irrespective of their size. A set containing an incorrect additional option receives zero, even when it also contains correct options.

Quantitative reasoning contains 58 items, each scored against its question-specific answer key. The scorer compares the final normalized answer or selected option with that key. Intermediate reasoning is not separately scored, and unresolved answers remain in the item mean.

For information extraction, the parser isolates the final answer and splits it at the literal # delimiter. Each position is compared with its corresponding requested field after punctuation and spacing normalization. Missing positions receive zero credit; values beyond the requested field count are

ignored. Let n_i be the number of requested fields in item i, $\hat{x}^{(r)}_{m,i,j}$ the parsed value at position j, and h the normalized field-agreement indicator. Then

$$S^{(r)}_{m,\text{ext}} = \frac{100}{\sum_{i=1}^{N_{\text{ext}}} n_i} \sum_{i=1}^{N_{\text{ext}}} \sum_{j=1}^{n_i} h\Big(\hat{x}^{(r)}_{m,i,j}, y_{i,j}\Big), \tag{6}$$

with $N_{\text{ext}} = 249$ and $\sum_i n_i = 1{,}005$. Fields receive equal weight, so instances with more requested fields contribute more to the score.

For risk classification, the parser isolates the final answer-bearing text, excluding reasoning blocks, and resolves an explicit answer span to one of six prompt categories: export control, fraud, money laundering, sanctions, prohibited or restricted sales, and no risk. A category mentioned only in a negated, hypothetical, conflicting, or unresolved expression is not treated as a selected answer. Resolution uses the response and prompt vocabulary, not the reference label. All 366 cases remain in the denominator.

For each response in an open-ended review operation, the fixed evaluator assigns coverage $u^{(r)}_{m,i} \in \{0, 0.5, 1\}$ and correctness $v^{(r)}_{m,i} \in \{0, 0.5, 1\}$ under the family rubric. Their repetition-level family means are

$$\bar{u}^{(r)}_{m,f} = \frac{1}{N_f} \sum_{i=1}^{N_f} u^{(r)}_{m,i}, \qquad \bar{v}^{(r)}_{m,f} = \frac{1}{N_f} \sum_{i=1}^{N_f} v^{(r)}_{m,i}. \tag{7}$$

The generation score is

$$S^{(r)}_{m,f} = \begin{cases} 100 \dfrac{2\bar{u}^{(r)}_{m,f} \bar{v}^{(r)}_{m,f}}{\bar{u}^{(r)}_{m,f} + \bar{v}^{(r)}_{m,f}}, & \bar{u}^{(r)}_{m,f} + \bar{v}^{(r)}_{m,f} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Coverage and correctness are averaged over the family before taking their harmonic mean. Equation 4 then averages these repetition-level scores.

B.4 Semantic Evaluation

DeepSeek-V4-Flash is the fixed evaluator for legal-judgment, decision-view, and disputed-issue generation, following rubric-guided LLM evaluation (Liu et al., 2023; Zheng et al., 2023). It receives the task input, source-supported reference, candidate response, and family rubric, and returns the coverage and correctness labels used above. Candidate responses are presented without model identifiers. Evaluator prompts, settings, and output parsing are fixed independently of candidate generation and applied uniformly. The evaluator's own candidate outputs remain in the full result tables but are excluded from comparisons that depend on its judgments.

FinRisk-Ask uses the same evaluator to assign one request-alignment label in $\{0, 0.5, 1\}$ against the complete evidence-target set. It receives the valid request, targets, and evaluator context view specified in Appendix C. Invalid requests receive zero credit without changing the parsed action.

Blinded validation covers the three generation operations and FinRisk-Ask request alignment. Agreement with adjudicated expert labels ranges from 0.73 to 0.80 in linearly weighted κ_w and from 0.82 to 0.87 in Spearman ρ. Appendix D provides the sampling procedure, human agreement, and task-specific validation results.

B.5 Aggregation and Configuration Eligibility

Let $\mathcal{F}_K$ contain the 42 Domain Knowledge families. The Knowledge macro is

$$K_m = \frac{1}{|\mathcal{F}_K|} \sum_{f \in \mathcal{F}_K} S_{m,f}, \qquad |\mathcal{F}_K| = 42. \tag{9}$$

Each family has equal weight regardless of its item count. Evidence-Grounded Processing and Applied Review use descriptive means of their five and six native operation scores, respectively

Table 9: Configuration eligibility for the reported analyses. Complete result tables retain the evaluator's rows for disclosure.

Comparison	Configurations	Inclusion rule
Domain Knowledge	33	Available 42-family macro
Each of eight structured operations	33	Available native operation score
Each generation operation	32	Available score; fixed evaluator excluded
Complete eleven-operation profiles	32	All operation scores available; evaluator excluded
Joint FinRisk-Ask analyses	32	Action and request metrics available; evaluator excluded
Static operation with FinRisk-Ask	32	Operation and Ask metrics available; evaluator excluded
Evidence-Grounded Processing mean	33	All five constituent scores available
Applied Review mean	32	All six constituent scores available; evaluator excluded

Table 10: Paired item-bootstrap intervals for fixed configuration contrasts. Knowledge cutoffs refer to Section 5.3.

Contrast	Gap	95% interval
Extraction, $k = 1$: Luna vs. Gemini-3	18.01	$[13.96, 21.79]$
Extraction, $k = 2$: Luna vs. Kimi-K3	11.54	$[7.76, 15.71]$
Quantitative reasoning, $k = 1$: Ling-1T vs. Gemini-3	12.07	$[5.17, 22.41]$

(Appendix F.4). Individual operation scores retain their native scoring rules and remain the primary downstream comparisons.

Table 9 specifies the pool used for each analysis. All eight structured operations admit 33 configurations. Generation comparisons exclude the fixed evaluator, leaving 32 configurations. The same 32 configurations are used for complete eleven-operation profiles and joint FinRisk-Ask analyses, including their action metrics.

Rankings, correlations, layer summaries, and shortlist gaps use the reported score matrix at two-decimal precision. An operation leader attains the highest reported score within its eligible pool. Knowledge-based screening uses descending competition ranks, with equal scores sharing a rank and all ties at the cutoff retained. Spearman correlations use average ranks for tied scores. The shortlist gap in Section 5.3 compares the eligible operation maximum with the best operation score retained at the Knowledge cutoff. Score gaps and maximum-recovery counts summarize performance within these eligible pools.

RAA weights all 680 states equally. AskR and ProceedR average over their respective recorded-action branches, and BAcc gives those two recalls equal weight. ERA, DEH, and REH average over all recorded Ask states. CRA uses each repetition's subset of recorded Ask states on which the configuration predicts Ask, including zero-credit invalid requests. Trajectories contribute according to their number of eligible states. Repetition-level identities and aggregation are specified in Appendix C.

B.6 Paired Item Resampling

Table 10 reports percentile 95% intervals for three fixed-configuration contrasts reconstructed from saved responses. We use 1,000 paired item-bootstrap draws with seed 20260915. Each draw uses the same sampled item indices for both configurations in a contrast.

For extraction, a draw samples 249 items with replacement and retains all requested fields within each sampled item. The score divides the total correct fields by the total requested fields in that draw, preserving the field-weighted definition in Equation 6. For quantitative reasoning, a draw samples the 58 items with replacement and averages their native item credits. Interval endpoints are the 2.5th and 97.5th percentiles of the bootstrap score differences.

Luna, Gemini-3, and Ling-1T denote GPT-5.6-Luna, Gemini-3-Flash-preview, and Ling-2.6-1T, respectively. Each gap subtracts the second configuration's score from the first. Both configurations remain fixed across bootstrap draws, so the intervals quantify paired item variation in the saved responses for these contrasts.

B.7 Reproducibility Materials and Data Access

Appendix E reports the full score matrix, and Appendix F defines the supplementary analyses. The anonymous package includes that matrix, the analysis definitions, and the scoring records used for the reported paired contrasts and the two FinRisk-Ask configurations whose state scores are included.

Rankings, correlations, layer summaries, and shortlist gaps are reconstructed from this matrix and the eligibility rules in Table 9, without new model calls. The extraction gaps 18.01 and 11.54 and the quantitative-reasoning gap 12.07 are the differences in that matrix. Their paired intervals are those reported in Table 10.

Institutional review, approval, and de-identification preceded the inclusion of enterprise trajectories. Direct identifiers were removed or replaced, and account numbers and contact fields were dropped. Temporal order and entity co-reference required for evaluation were preserved. The planned data release uses non-commercial research terms, and the evaluation code uses a permissive license. Research-use limitations are discussed in Appendix H.

C FinRisk-Ask Protocol

This section specifies state reconstruction, response processing, and scoring for the offline evaluation in Section 4. At each state, all configurations receive the same pre-action context. Recorded reviewer transitions provide action references, while expert-verified evidence needs provide request targets.

C.1 Trajectory Reconstruction and Retention

For a decision point t in a completed, de-identified enterprise review trajectory, C_t contains the case record and history available immediately before the reviewer acted. The next recorded action, later evidence, subsequent actions, and outcomes are withheld from the model. The information boundary is reconstructed separately at each decision point.

A state is retained when its pre-action context can be reconstructed, its immediately recorded transition maps unambiguously to Ask or Proceed, and its input contains no post-action information. Each retained Ask state additionally requires at least one expert-verified evidence target. Retention precedes candidate-model evaluation.

The evaluation set contains 680 states from 104 trajectories: 583 recorded Ask and 97 recorded Proceed states. Construction audits are reported in Appendix D, and data handling is described in Appendix B.7.

C.2 Recorded Actions and Evidence Targets

The reference action is the immediately recorded reviewer transition at the current review step:

$$a_t^* = \begin{cases} \text{Ask}, & \text{new evidence is requested for the current step,} \\ \text{Proceed}, & \text{the current step advances without a new request.} \end{cases} \tag{10}$$

Experts reviewed the transition mappings and adjudicated disagreements.

For each retained Ask state, later-observed materials support the target set

$$\mathcal{E}_t^+ = \{e_{t,1}, e_{t,2}, \ldots, e_{t,n_t}\}, \qquad n_t \geq 1. \tag{11}$$

Experts verify that each target concerns information unavailable at t, addresses an unresolved need, and constitutes a meaningful acquisition objective. Targets identify evidence needs rather than historical request wording, allowing alternative formulations to address the same need. The complete target set is reserved for evaluation.

C.3 Output Parsing and Request Validity

Given C_t, the model either requests additional evidence or advances the current review step. The output contract is

$$\hat{y}_t = \begin{cases} \mathrm{Ask}(\hat{q}_t), & \text{request one focused item of evidence for this step,} \\ \mathrm{Proceed}, & \text{advance this step without another evidence request.} \end{cases} \tag{12}$$

The action parser returns $\hat{a}_t \in \{\mathrm{Ask}, \mathrm{Proceed}, \mathrm{Invalid}\}$. For Ask, the request parser extracts $\hat{q}_t$ and checks the one-request contract. Parsing uses the response and output schema without consulting the action reference or evidence targets.

A parsed Ask action remains an Ask prediction even when its request is empty, malformed, multi-item, or otherwise invalid. Such requests receive zero request credit. Valid requests proceed to semantic scoring.

C.4 Action and Request Metrics

For one configuration and repetition, suppress their indices and let $\mathcal{T} = \{1, \ldots, N\}$, with $N = 680$. The reference branches are

$$\mathcal{T}_{\mathrm{Ask}} = \{t \in \mathcal{T} : a_t^* = \mathrm{Ask}\}, \qquad \mathcal{T}_{\mathrm{Proceed}} = \{t \in \mathcal{T} : a_t^* = \mathrm{Proceed}\}, \tag{13}$$

with sizes 583 and 97, respectively. All aggregate metrics use a 0–100 scale.

Recorded-Action Agreement (RAA) averages exact action agreement over all states:

$$\mathrm{RAA} = \frac{100}{N} \sum_{t \in \mathcal{T}} \mathbb{I}[\hat{a}_t = a_t^*]. \tag{14}$$

Ask recall and Proceed recall use their respective reference branches:

$$\mathrm{AskR} = \frac{100}{|\mathcal{T}_{\mathrm{Ask}}|} \sum_{t \in \mathcal{T}_{\mathrm{Ask}}} \mathbb{I}[\hat{a}_t = \mathrm{Ask}], \tag{15}$$

$$\mathrm{ProceedR} = \frac{100}{|\mathcal{T}_{\mathrm{Proceed}}|} \sum_{t \in \mathcal{T}_{\mathrm{Proceed}}} \mathbb{I}[\hat{a}_t = \mathrm{Proceed}]. \tag{16}$$

Balanced Recorded-Action Agreement (BAcc) gives the branches equal weight:

$$\mathrm{BAcc} = \frac{\mathrm{AskR} + \mathrm{ProceedR}}{2}. \tag{17}$$

Overall agreement weights the branch recalls by their reference proportions:

$$\mathrm{RAA} = \frac{583}{680}\mathrm{AskR} + \frac{97}{680}\mathrm{ProceedR}. \tag{18}$$

Invalid actions receive no agreement credit and remain in the denominators.

For request scoring on recorded Ask states, the fixed semantic evaluator receives the valid request $\hat{q}_t$, the complete target set $\mathcal{E}_t^+$, and a context view $\widetilde{C}_t$. Candidate models receive C_t. The evaluator makes one judgment against the complete target set:

$$\ell(\hat{q}_t, \widetilde{C}_t, \mathcal{E}_t^+) = \begin{cases} 1, & \text{direct alignment with at least one retained need,} \\ 0.5, & \text{relevant but incomplete alignment, without a direct match,} \\ 0, & \text{an unrelated request.} \end{cases} \tag{19}$$

A focused request earns full credit by addressing any one retained need. For $t \in \mathcal{T}_{\mathrm{Ask}}$, its credit is

$$z_t = \begin{cases} \ell(\hat{q}_t, \widetilde{C}_t, \mathcal{E}_t^+), & \hat{a}_t = \mathrm{Ask} \text{ and } \hat{q}_t \text{ is valid,} \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

Requests on recorded Proceed states affect action agreement but are not scored for request alignment. Evaluator settings and human validation are reported in Appendices B and D.

Evidence-Request Alignment (ERA) averages request credit over all recorded Ask states:

$$\mathrm{ERA} = \frac{100}{|\mathcal{T}_{\mathrm{Ask}}|} \sum_{t \in \mathcal{T}_{\mathrm{Ask}}} z_t. \tag{21}$$

Define the conditional subset $\mathcal{A} = \{t \in \mathcal{T}_{\mathrm{Ask}} : \hat{a}_t = \mathrm{Ask}\}$. When $|\mathcal{A}| > 0$, Conditional Request Alignment (CRA) is

$$\mathrm{CRA} = \frac{100}{|\mathcal{A}|} \sum_{t \in \mathcal{A}} z_t, \qquad \mathrm{ERA} = \frac{\mathrm{AskR} \times \mathrm{CRA}}{100}. \tag{22}$$

ERA captures both Ask-state coverage and request alignment. CRA measures alignment within the configuration's selected subset of recorded Ask states, including zero-credit invalid requests. When $\mathcal{A}$ is empty, ERA is zero and CRA is undefined.

Direct Evidence Hit (DEH) and Relevant Evidence Hit (REH) use the same denominator as ERA:

$$\mathrm{DEH} = \frac{100}{|\mathcal{T}_{\mathrm{Ask}}|} \sum_{t \in \mathcal{T}_{\mathrm{Ask}}} \mathbb{I}[z_t = 1], \tag{23}$$

$$\mathrm{REH} = \frac{100}{|\mathcal{T}_{\mathrm{Ask}}|} \sum_{t \in \mathcal{T}_{\mathrm{Ask}}} \mathbb{I}[z_t \geq 0.5]. \tag{24}$$

Since $z_t \in \{0, 0.5, 1\}$,

$$\mathrm{ERA} = \frac{\mathrm{DEH} + \mathrm{REH}}{2}. \tag{25}$$

The four shares in Figure 5(b) are DEH for direct alignment, $\mathrm{REH} - \mathrm{DEH}$ for partial alignment, $\mathrm{AskR} - \mathrm{REH}$ for zero-credit Ask predictions, and $100 - \mathrm{AskR}$ for no Ask prediction. These shares sum to 100; the last includes Proceed and Invalid actions.

C.5 Repetition Aggregation and Baselines

Each reported metric averages its available repetition-level values. Within a repetition, states receive equal weight in each reference or conditional subset, and BAcc gives equal weight to the two branch recalls. Trajectories contribute according to their number of eligible states. Mean CRA is reported only when its conditional subset is nonempty in every repetition; otherwise it is marked as unavailable.

The additive identities for BAcc, RAA, and ERA remain valid after averaging, up to reporting precision. Ling-2.6-1T illustrates the rounding: the displayed AskR and ProceedR average to 51.90, while the reported BAcc is 51.89; the displayed AskR and CRA multiply to 79.75, while the reported ERA is 79.76. When CRA is defined in every repetition, the request-pathway identity gives

$$\overline{\mathrm{ERA}}_m = \frac{1}{R_m} \sum_{r=1}^{R_m} \frac{\mathrm{AskR}_m^{(r)} \mathrm{CRA}_m^{(r)}}{100}, \tag{26}$$

where R_m is the number of available repetitions for configuration m. The product of separately averaged AskR and CRA need not equal mean ERA. Figure 3(e) therefore uses product curves as reference values.

Always-Ask and Always-Proceed both have BAcc of 50.00, with RAA of 85.74 and 14.26, respectively. Always-Ask specifies no request-generation policy, so its request metrics are not reported. Always-Proceed has zero ERA, DEH, and REH, and undefined CRA. The complete result table in Appendix E includes both references alongside the evaluated configurations.

Together, the action metrics and request diagnostics characterize agreement with recorded reviewer transitions, coverage of recorded acquisition states, and alignment with verified evidence needs. Appendix D reports the corresponding reference and evaluator audits.

D Human Annotation Reliability and Evaluator Validation

This section reports construction reliability, context-only assessment of trajectory references, and validation of the semantic evaluator. Construction audits use source material, action and request

Table 11: Construction audits. N denotes the audit pool size. Rates are exact label agreement except for source confirmation, which measures verified source support.

Audit	Sampling unit and pool	N	Rate (%)	κ
Contract instantiation	15 candidates per family	795	93.7	0.81
Item retention	400 retained and 400 excluded candidates	800	92.1	0.84
Knowledge source confirmation	Sampled knowledge items	400	95.8	-
Structured reference validity	Census of six structured operations	1,868	97.1	0.89
Provision reference validity	Stratified provision sample	250	90.8	0.78
Legal-outcome reference validity	Stratified outcome sample	250	92.4	0.81
State retention	Reconstructed trajectory candidates	742	94.3	0.72
Recorded-transition encoding	Action-labeled states	690	95.8	0.83

audits use model-visible context, and evaluator validation applies the task rubrics. Each analysis is reported with its sampling units and agreement measures.

D.1 Independent Construction Audit

All retained benchmark items received initial expert review. Independent second-pass audits used prespecified sampling units and strata. The second annotator had not constructed the audited item and was blinded to the first judgment. Agreement is computed from pre-adjudication labels, excluding unresolved cases. Audit pools can overlap across construction stages.

The contract-instantiation audit included ten assigned candidates and five neighboring-family controls for each of the 53 families, totaling 795 candidates. The separate retention audit included 400 retained and 400 excluded candidates to assess the application of retention criteria.

Source support was confirmed for 95.8% of the 400 sampled Domain Knowledge items; failures were corrected or removed before release. The 1,868-item structured-reference census covered case classification, information extraction, institution matching, person matching, quantitative reasoning, and risk classification. Provision and legal-outcome references were assessed in separate stratified samples of 250 items each (Table 11).

For the 742 reconstructed candidates, both annotators retained 638 states and excluded 62; each annotator alone retained 21. These counts yield pre-adjudication agreement of 94.3% and $\kappa = 0.72$. Recorded-transition encoding was audited on 690 action-labeled states. These pools cover different construction stages, with 680 states retained for evaluation. Every target set for the 583 released Ask states was reviewed for temporal unavailability and relevance to an unresolved need.

D.2 Context-Only Audit of Recorded Actions

Two experts who had not encoded the audited states independently assessed action acceptability from the same context C_t shown to candidate models. Later observations and recorded actions were withheld. The audit included all 97 recorded Proceed states and 100 trajectory-stratified Ask states. Pre-adjudication agreement was 0.91, with $\kappa = 0.71$.

After adjudication, the recorded action was accepted on 94 of 100 Ask states (94.0%) and 90 of 97 Proceed states (92.8%), totaling 184 of 197 audited states (93.4%). Both actions were acceptable on 8.1% of the audited states; these states are included in the accepted count. The pooled acceptance rate reflects the audit sample's branch proportions.

The accepted-state analysis restricts scoring to the 184 accepted states, retaining their reference labels and request targets and recomputing metric denominators. The dual-action analysis retains all 197 audited states and credits either valid action on dual-acceptable states for action agreement. In

Table 12: Blinded semantic-evaluator validation. H–H compares independent human annotations; J–H compares evaluator judgments with adjudicated human labels.

Setting	N_{resp}	N_{label}	H–H κ_w	J–H κ_w	J–H ρ
Legal-judgment generation	128	256	0.79	0.75	0.82
Decision-view generation	128	256	0.82	0.78	0.86
Disputed-issue generation	128	256	0.77	0.73	0.82
FinRisk-Ask alignment	200	200	0.84	0.80	0.87

both analyses, BAcc averages agreement within the two recorded-action branches, and ERA scores requests on recorded Ask states, assigning zero when no valid Ask request is produced.

Under both variants on the audited subset, Kimi-K3 remains the BAcc leader and Ling-2.6-1T remains the ERA leader among the 32 non-evaluator configurations. AskR remains higher than ProceedR for every configuration.

D.3 Blinded Validation of the Semantic Evaluator

For generation validation, saved responses were available for 27 of the 32 non-evaluator configurations. Qwen3-1.7B, Qwen3-4B, Gemini-2.5-Flash, Gemini-3-Flash-preview, and Claude-Opus-4.7 lacked local response archives and were excluded from the sampling frame. We randomly selected 16 configurations using seed 20260921 and used the same configuration set for all three generation operations. Within each configuration and operation, eight responses were selected across the evaluator's score range, yielding 128 responses per operation. Request validation used 200 FinRisk-Ask requests sampled across configurations and alignment levels. Two experts independently applied the fixed component or request rubrics while blinded to model identity, provider, evaluator scores, and each other's judgments. Rubrics were fixed before annotation, and disagreements were adjudicated after the independent labels were recorded.

We report linearly weighted Cohen's κ (Cohen, 1968) for labels in $\{0, 0.5, 1\}$, using agreement weights $w(a, b) = 1 - |a - b|$. H–H compares the two independent human annotations before adjudication. J–H compares the fixed evaluator with the adjudicated human labels. Spearman ρ measures the association between evaluator and adjudicated human labels in the validation sample.

Each generation response contributes one coverage label and one correctness label, pooled within the task to give 256 paired label observations from 128 responses. Each request contributes one alignment label. Agreement is computed on these component and request labels, separately from the family-level score aggregation in Appendix B.

Table 12 reports independent human agreement of $\kappa_w = 0.77$–0.84. Evaluator agreement with the adjudicated human labels ranges from $\kappa_w = 0.73$ to 0.80, with Spearman $\rho = 0.82$–0.87. On these validation samples, evaluator labels agree with the adjudicated expert labels at the reported level.

E Complete Configuration Results

Tables 13–16 report the score matrix used in Section 5. Scores are repetition means on a 0–100 scale, displayed to two decimal places under the protocols in Appendices B and C. Configurations appear in the same order throughout. The fixed evaluator is marked † and excluded from generation and joint FinRisk-Ask comparisons, while remaining eligible for Knowledge and structured operations. Light gray shading identifies constant-action references and evaluator rows excluded from the corresponding comparison.

E.1 Domain Knowledge

Table 13 presents five display groups alongside the Knowledge macro. Risk and Compliance, Trade, FinTech and Security, Commerce and Payment, and Finance and Law contain 20, 14, 4, 2, and 2 knowledge families, respectively. Overall follows the 42-family macro in Equation 9, rather than an unweighted average of the five group scores. The group columns show how knowledge performance varies within the screening score.

Table 13: Domain Knowledge results for 33 configurations. Overall is the unweighted 42-family macro.

Configuration	Risk & Compliance	Trade	FinTech & Security	Commerce & Payment	Finance & Law	Overall
Dianjin-R1-7B	64.00	63.43	70.25	72.50	68.00	65.00
FinR1-7B	66.35	68.00	78.25	75.50	76.00	68.93
Ling-2.6-Flash	71.80	74.71	80.00	79.00	77.50	74.17
Ling-2.6-1T	75.22	78.96	83.22	78.82	80.49	77.65
Seed-2.0-Mini	73.22	76.64	80.43	78.32	80.04	75.62
Seed-2.0-Lite	73.70	80.36	78.78	81.51	78.55	77.01
Nemotron-3-Nano	49.06	49.30	57.49	57.46	49.92	50.38
Nemotron-3-Super	64.74	68.49	73.37	70.36	72.16	67.43
Nemotron-3-Ultra	71.81	76.89	78.06	79.73	81.85	74.96
Qwen2.5-7B	62.23	61.79	67.11	71.85	70.94	63.42
Qwen3-1.7B	40.13	37.87	43.78	44.49	46.05	40.21
Qwen3-4B	63.17	60.38	68.45	69.46	69.10	63.32
Qwen3-8B	67.12	65.68	71.24	74.42	73.06	67.66
Qwen3-14B	67.92	70.93	76.49	73.66	77.65	70.48
Qwen3-32B	70.90	74.46	79.73	79.20	80.66	73.79
Qwen3-Next-80B	74.20	78.97	82.52	79.03	82.08	77.19
Qwen3-235B-A22B	73.25	79.02	83.02	82.08	81.26	76.90
Qwen3.7-Max	77.15	82.49	85.04	86.01	83.65	80.41
DeepSeek-V3-0324	70.44	75.65	78.33	74.99	79.48	73.57
DeepSeek-V4-Flash	71.33	75.49	79.98	76.30	77.65	74.08
DeepSeek-V4-Pro	72.11	76.87	82.73	81.55	79.88	75.53
Kimi-K2.5	75.23	81.95	82.15	80.25	79.21	78.56
Kimi-K3	78.11	83.05	84.07	82.80	80.95	80.68
GLM-5	73.53	78.14	80.27	79.93	78.81	76.26
GLM-5.2	74.65	79.18	80.33	79.43	79.02	77.13
GPT-5.4-Mini	74.38	79.48	80.59	84.34	80.16	77.42
GPT-5.4	74.46	81.17	84.69	84.34	85.34	78.66
GPT-5.6-Luna	72.49	77.61	77.59	78.81	78.53	75.27
GPT-5.6-Terra	74.87	81.58	81.19	82.73	80.24	78.34
GPT-5.6-Sol	76.30	83.00	80.23	82.42	76.71	79.21
Gemini-2.5-Flash	70.27	77.68	80.56	80.51	77.98	74.57
Gemini-3-Flash-preview	77.24	83.80	84.33	86.41	80.53	80.69
Claude-Opus-4.7	76.55	82.15	86.92	86.81	82.75	80.19

E.2 Structured Review Operations

Table 14 reports the five Evidence-Grounded Processing operations and three structured Applied Review operations. All 33 configurations are eligible. Extraction uses field-level agreement, quantitative reasoning uses item-specific answer scores, and provision selection uses partial-credit set scoring; the remaining operations use exact label agreement.

E.3 Open-Ended Applied Review

Table 15 reports the three generation operations. Within each repetition, coverage and correctness are averaged over the family before their harmonic mean is computed (Equation 8). The comparison pool contains 32 configurations, excluding the fixed evaluator's self-scored row. Together with the structured-operation scores, these results form the eleven-operation profiles used in RQ1 and RQ2.

E.4 FinRisk-Ask

Table 16 reports action agreement and request alignment. RAA averages agreement over all states, while BAcc weights the two recorded-action branches equally. ERA covers all recorded Ask states; CRA uses the subset of those states on which a configuration predicts Ask. DEH and REH distinguish direct from at least partial alignment. RQ3 compares the 32 non-evaluator configurations. Constant-action policies are shown separately, and "–" denotes an unspecified request policy or an undefined conditional metric.

Table 14: Structured-operation results for 33 configurations. Columns follow the operation definitions in Table 7.

	Evidence-Grounded Processing					Structured Applied Review		
Configuration	Case class.	Extraction	Inst. match.	Person match.	Quant. reason.	Risk class.	Legal outcome	Provision selection
Dianjin-R1-7B	76.63	17.51	89.60	64.20	69.83	60.11	53.86	35.65
FinR1-7B	76.18	9.75	82.40	84.40	77.59	66.39	63.60	34.90
Ling-2.6-Flash	80.00	38.21	78.00	80.00	62.00	74.59	60.00	40.00
Ling-2.6-1T	82.92	74.53	78.80	80.40	89.66	75.68	64.76	52.10
Seed-2.0-Mini	83.15	75.42	86.80	73.80	79.31	68.85	62.22	36.95
Seed-2.0-Lite	85.62	84.48	84.00	81.80	79.31	73.22	63.81	56.40
Nemotron-3-Nano	66.29	33.43	75.60	73.40	65.52	54.92	53.76	23.50
Nemotron-3-Super	79.78	72.44	82.40	79.20	75.00	74.32	56.72	30.15
Nemotron-3-Ultra	82.02	81.49	74.80	85.60	81.90	75.41	61.69	29.45
Qwen2.5-7B	76.63	32.14	78.00	82.00	75.00	68.03	62.96	40.55
Qwen3-1.7B	54.61	61.79	79.60	76.40	56.03	71.86	58.94	16.20
Qwen3-4B	76.40	75.72	80.00	84.00	75.00	71.31	62.33	12.80
Qwen3-8B	81.72	61.69	80.27	83.00	70.86	74.86	64.49	17.44
Qwen3-14B	81.80	76.82	78.80	85.40	74.14	68.85	64.97	33.40
Qwen3-32B	82.10	74.63	90.00	84.40	79.66	74.04	62.96	27.44
Qwen3-Next-80B	84.94	76.02	90.80	81.60	80.17	73.77	65.71	41.30
Qwen3-235B-A22B	83.37	71.74	88.00	83.40	81.03	67.21	65.08	40.50
Qwen3.7-Max	85.39	82.79	84.00	84.40	78.45	75.68	66.03	46.85
DeepSeek-V3-0324	83.15	79.12	84.40	80.40	76.72	74.86	63.70	34.15
DeepSeek-V4-Flash	82.92	51.84	88.00	83.40	72.41	77.05	63.70	53.00
DeepSeek-V4-Pro	85.62	76.02	86.80	85.20	68.97	70.49	64.97	41.45
Kimi-K2.5	88.09	84.28	87.20	83.60	75.86	73.22	65.93	46.40
Kimi-K3	87.19	74.43	84.80	87.00	77.59	74.86	68.15	51.65
GLM-5	84.72	81.39	87.60	84.20	84.48	73.22	64.76	51.50
GLM-5.2	84.94	81.69	85.60	83.60	78.45	71.86	65.71	51.15
GPT-5.4-Mini	85.62	83.78	81.60	85.80	76.72	71.86	64.97	55.35
GPT-5.4	86.29	83.98	82.00	85.60	75.86	74.04	67.51	60.45
GPT-5.6-Luna	84.04	85.97	85.60	85.80	75.00	71.86	65.93	50.80
GPT-5.6-Terra	85.62	83.48	90.40	85.80	76.72	73.77	67.83	56.95
GPT-5.6-Sol	87.42	85.07	90.00	86.00	77.59	74.59	67.09	67.55
Gemini-2.5-Flash	82.25	75.72	87.60	84.80	66.38	72.95	60.11	66.25
Gemini-3-Flash-preview	84.94	67.96	93.60	85.80	77.59	76.78	59.68	65.75
Claude-Opus-4.7	85.17	76.62	87.20	86.00	73.28	75.14	65.19	54.80

F Supplementary Capability Analyses

This section connects the complete scores in Appendix E to the operation profiles, screening results, and evidence-seeking diagnostics in Section 5. All calculations use the reported score matrix and the eligibility rules in Table 9.

F.1 Complete Capability Profiles

Figure 4(a) displays Domain Knowledge and all eleven operations. Within-column percentiles show relative standing under each scoring contract, and white circles mark the highest reported scores in the corresponding eligible pools. Native scores remain in Tables 13–15.

Panel (b) compares the native scores of Gemini-3-Flash-preview and Kimi-K3, whose Knowledge scores are 80.69 and 80.68. Their opposite advantages on applicable-provision selection and legal-outcome prediction illustrate how nearly equal Knowledge scores can accompany different operation profiles. The pairwise analysis below extends this comparison across the complete-profile pool.

Table 15: Open-ended Applied Review results. The evaluator's self-scored row (†) is excluded from comparisons.

Configuration	Legal-judgment generation	Decision-view generation	Disputed-issue generation
Dianjin-R1-7B	29.00	20.00	36.00
FinR1-7B	18.00	15.00	52.00
Ling-2.6-Flash	30.00	37.00	63.00
Ling-2.6-1T	40.27	55.69	63.30
Seed-2.0-Mini	28.95	47.11	72.73
Seed-2.0-Lite	29.66	58.73	73.35
Nemotron-3-Nano	21.48	17.09	39.57
Nemotron-3-Super	28.62	37.17	76.67
Nemotron-3-Ultra	30.68	52.00	71.58
Qwen2.5-7B	28.10	36.24	70.40
Qwen3-1.7B	16.18	7.76	41.13
Qwen3-4B	19.38	21.99	68.83
Qwen3-8B	17.37	30.05	37.30
Qwen3-14B	35.02	39.03	53.81
Qwen3-32B	34.54	35.60	53.30
Qwen3-Next-80B	32.43	53.10	67.51
Qwen3-235B-A22B	39.26	49.00	57.35
Qwen3.7-Max	36.22	57.84	70.58
DeepSeek-V3-0324	36.26	71.31	51.68
DeepSeek-V4-Flash†	25.16	54.86	81.23
DeepSeek-V4-Pro	30.67	54.48	77.12
Kimi-K2.5	34.49	62.74	61.26
Kimi-K3	35.05	64.13	74.84
GLM-5	31.60	57.16	72.49
GLM-5.2	31.59	58.59	68.37
GPT-5.4-Mini	39.86	48.77	62.13
GPT-5.4	46.36	62.91	54.04
GPT-5.6-Luna	41.70	52.77	61.33
GPT-5.6-Terra	40.98	54.72	58.29
GPT-5.6-Sol	42.09	62.25	58.11
Gemini-2.5-Flash	45.29	65.27	74.05
Gemini-3-Flash-preview	40.48	64.61	74.69
Claude-Opus-4.7	35.53	51.97	74.06

F.2 Static Operations and Evidence-Seeking Behavior

Figure 5(a) compares AskR and ProceedR for each configuration. Panel (b) separates recorded Ask states into direct alignment (DEH), partial alignment ($\mathrm{REH} - \mathrm{DEH}$), zero-credit Ask predictions ($\mathrm{AskR} - \mathrm{REH}$), and no Ask prediction ($100 - \mathrm{AskR}$). Zero-credit Ask predictions include invalid and unaligned requests; no Ask prediction includes Proceed and Invalid actions. The shares sum to 100, and ERA equals the direct share plus half the partial share, up to reporting precision. The decomposition distinguishes missing Ask predictions from partial or zero-credit requests.

These additive relationships also hold after repetition averaging. The product of mean AskR and mean CRA, divided by 100, need not equal mean ERA. The curves in Figure 3(e) therefore show reference values, following Equation 26.

Panel (c) reports Spearman correlations between each operation score and each FinRisk-Ask metric across the same 32 configurations. Legal-judgment generation correlates with CRA at $\rho = 0.80$ and with BAcc at $\rho = 0.19$. Decision-view generation shows a similar contrast, with correlations of 0.80 and 0.10, respectively. For these operations, the association with conditional request alignment is stronger than with balanced action agreement. Generation and request alignment use the same fixed evaluator (Appendix B).

Panel (d) contrasts state-weighted RAA with branch-balanced BAcc. Always-Ask attains RAA of 85.74 because 583 of 680 references are Ask, but its BAcc is 50.00. This difference follows the

Table 16: FinRisk-Ask results and constant-action references. Metrics follow Appendix C; the fixed evaluator (†) is excluded from joint comparisons.

	Action agreement				Request alignment			
Configuration	BAcc	AskR	ProceedR	RAA	ERA	CRA	DEH	REH
Always-Ask	50.00	100.00	0.00	85.74	–	–	–	–
Always-Proceed	50.00	0.00	100.00	14.26	0.00	–	0.00	0.00
Dianjin-R1-7B	45.97	91.94	0.00	78.82	57.72	62.78	51.97	63.46
FinR1-7B	50.00	100.00	0.00	85.74	48.46	48.46	45.45	51.46
Ling-2.6-Flash	56.03	75.99	36.08	70.29	51.80	68.17	49.40	54.20
Ling-2.6-1T	51.89	96.57	7.22	83.82	79.76	82.58	76.50	83.02
Seed-2.0-Mini	48.65	75.64	21.65	67.94	54.55	72.10	52.32	56.78
Seed-2.0-Lite	44.36	60.89	27.84	56.18	42.97	70.55	40.31	45.63
Nemotron-3-Nano	39.48	40.82	38.14	40.44	24.79	60.70	22.98	26.59
Nemotron-3-Super	46.77	56.43	37.11	53.68	36.28	64.27	34.99	37.56
Nemotron-3-Ultra	55.69	82.50	28.87	74.85	67.75	82.12	66.21	69.30
Qwen2.5-7B	43.52	45.80	41.24	45.15	26.84	58.61	25.04	28.64
Qwen3-1.7B	51.38	94.51	8.25	82.21	31.39	33.20	29.33	33.45
Qwen3-4B	54.68	55.75	53.61	55.44	34.31	61.53	32.93	35.68
Qwen3-8B	55.09	77.19	32.99	70.88	48.80	63.22	46.14	51.46
Qwen3-14B	54.83	75.64	34.02	69.71	50.77	67.12	48.89	52.66
Qwen3-32B	58.19	65.87	50.52	63.68	48.54	73.69	45.97	51.11
Qwen3-Next-80B	56.21	73.24	39.18	68.38	57.72	78.79	56.09	59.35
Qwen3-235B-A22B	53.63	79.42	27.84	72.06	62.86	79.15	61.58	64.15
Qwen3.7-Max	43.85	57.80	29.90	53.82	50.69	87.67	49.91	51.46
DeepSeek-V3-0324	49.52	57.80	41.24	55.44	45.80	79.22	44.43	47.17
DeepSeek-V4-Flash†	50.19	81.82	18.56	72.79	59.52	72.73	57.63	61.41
DeepSeek-V4-Pro	50.71	75.64	25.77	68.53	60.29	79.70	58.49	62.09
Kimi-K2.5	55.68	83.53	27.84	75.59	66.81	79.97	65.35	68.27
Kimi-K3	59.46	89.02	29.90	80.59	77.10	86.61	76.50	77.70
GLM-5	55.26	73.41	37.11	68.24	55.66	75.82	54.20	57.12
GLM-5.2	53.37	71.70	35.05	66.47	58.49	81.58	57.29	59.69
GPT-5.4-Mini	48.15	51.97	44.33	50.88	39.28	75.56	38.25	40.31
GPT-5.4	50.45	72.04	28.87	65.88	59.18	82.13	57.98	60.38
GPT-5.6-Luna	54.21	95.03	13.40	83.38	77.87	81.95	75.81	79.93
GPT-5.6-Terra	56.11	91.60	20.62	81.47	75.64	82.58	74.10	77.19
GPT-5.6-Sol	54.05	83.36	24.74	75.00	67.75	81.27	66.55	68.95
Gemini-2.5-Flash	49.33	74.96	23.71	67.65	62.61	83.51	61.23	63.98
Gemini-3-Flash-preview	52.51	77.19	27.84	70.15	64.49	83.55	62.95	66.04
Claude-Opus-4.7	56.11	88.51	23.71	79.26	75.99	85.85	74.10	77.87

weighting identity in Equation 18 and shows how overall agreement can conceal limited agreement on recorded Proceed states.

F.3 Knowledge Proximity and Shortlist Coverage

Panels (a,b) use the common pool $\mathcal{M}_{\text{complete}}$ of 32 non-evaluator configurations with all eleven operation scores. Let $M = |\mathcal{M}_{\text{complete}}| = 32$, and let $R_{m,o}$ be the descending average rank of configuration m on operation o within this pool. Tied scores receive average ranks. For the profile comparison, define

$$P_{m,o} = 100\frac{M - R_{m,o}}{M - 1}, \qquad D^K_{mn} = |K_m - K_n|. \tag{27}$$

The mean difference between the two operation profiles is

$$G_{mn} = \frac{1}{|\mathcal{O}|}\sum_{o\in\mathcal{O}} |P_{m,o} - P_{n,o}|, \tag{28}$$

where $\mathcal{O}$ contains the eleven operations. Each operation contributes equally, and the common rank scale permits comparison across different native scoring rules.

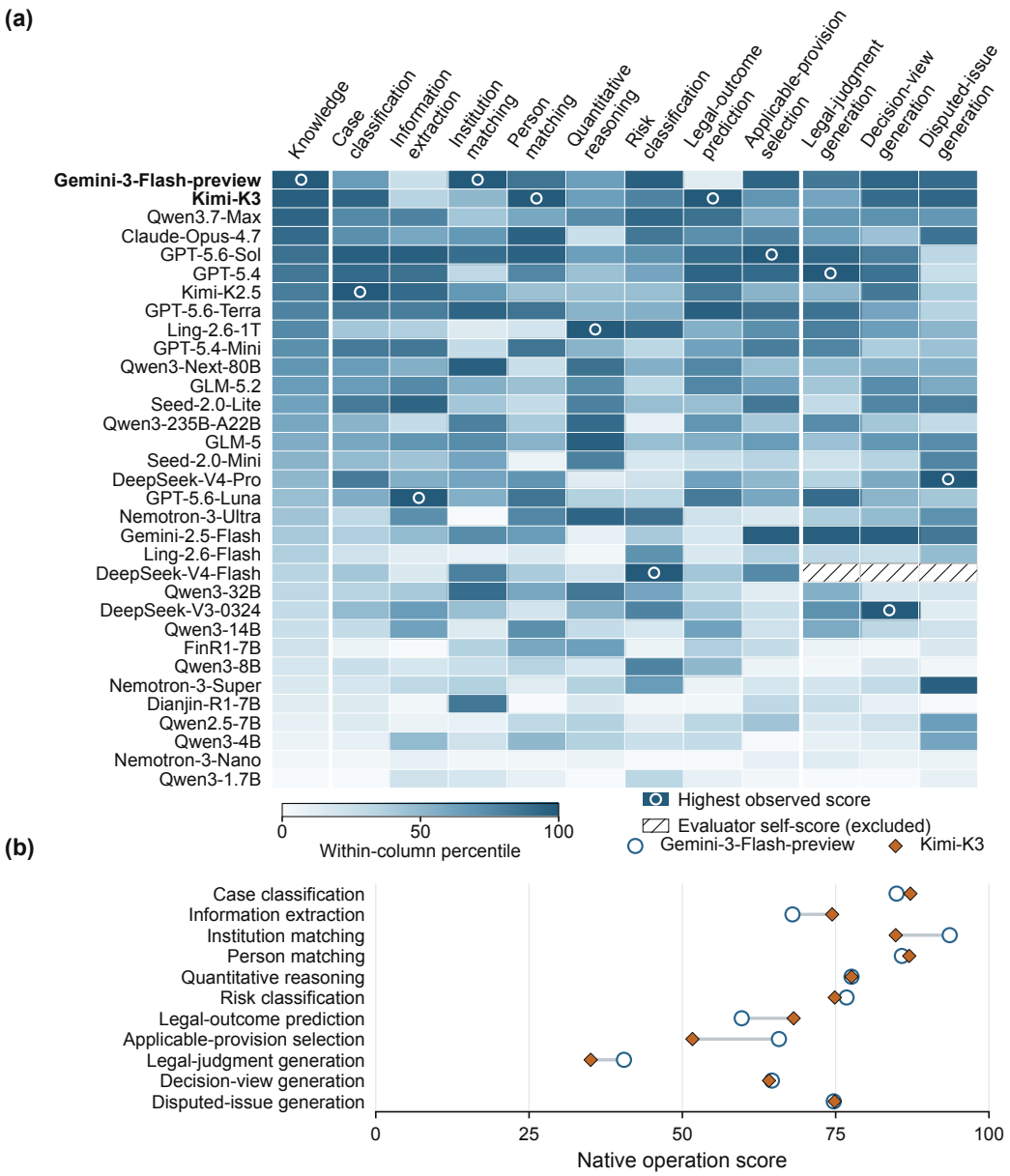

Figure 4: Configuration profiles. (a) Within-column percentiles ordered by Knowledge; white circles mark eligible maxima. (b) Native operation scores of the two Knowledge leaders.

There are $\binom{32}{2} = 496$ configuration pairs, of which 30 have $D^K_{mn} \leq 0.5$. Their median profile gap is 24.0 percentile points, compared with 34.3 across all pairs. Knowledge-near configurations thus have a lower median profile gap, while still differing in their standing on individual operations.

For panel (b), let $\mathbf{C} \in \mathbb{R}^{11\times 11}$ be the Spearman correlation matrix over the same complete profiles. Its ordered eigenvalues $\lambda_1 \geq \cdots \geq \lambda_{11} \geq 0$ define the component shares

$$\eta_j = \frac{\lambda_j}{\sum_{\ell=1}^{11} \lambda_\ell}, \qquad \sum_{\ell=1}^{11} \lambda_\ell = 11. \tag{29}$$

The leading eigenvalue accounts for 50.1% of the total, and the first five account for 84.5%. Together with the mean pairwise operation correlation of 0.42, this summary shows shared ranking structure alongside variation across operations.

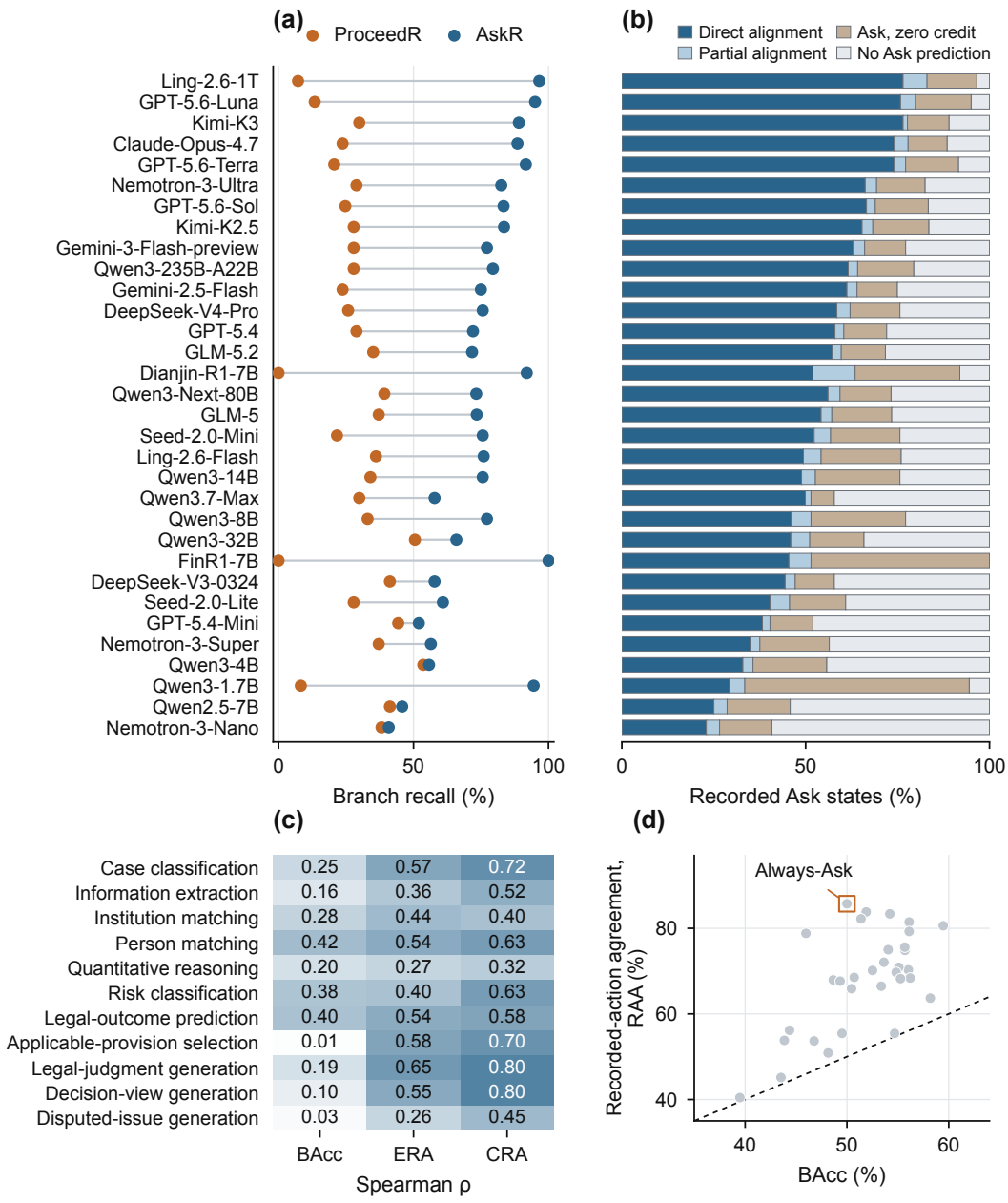

Figure 5: Evidence-seeking diagnostics for 32 non-evaluator configurations. (a) AskR and ProceedR ordered by ERA. (b) Outcomes on recorded Ask states. (c) Spearman correlations of operation scores with BAcc, ERA, and CRA. (d) RAA versus BAcc with the Always-Ask reference.

Panels (c,d) use the operation-specific pools $\mathcal{M}_o$ and competition ranks $r^K_{m,o}$ from Section 5.3. An operation maximum is recovered when the shortlist contains any configuration attaining that maximum. The count in panel (c) is

$$\text{Recovered}(k) = \sum_{o \in \mathcal{O}} \mathbb{I}[\text{Reg}_o(k) = 0]. \tag{30}$$

Ties at the Knowledge cutoff are retained. Within the evaluated pools, the counts in panel (c) are one, seven, and eleven recovered maxima at $k = 1$, $k = 10$, and $k = 23$, respectively. Panel (d) reports the Knowledge rank of each operation leader, including rank 18 for information extraction and rank 23 for decision-view generation.

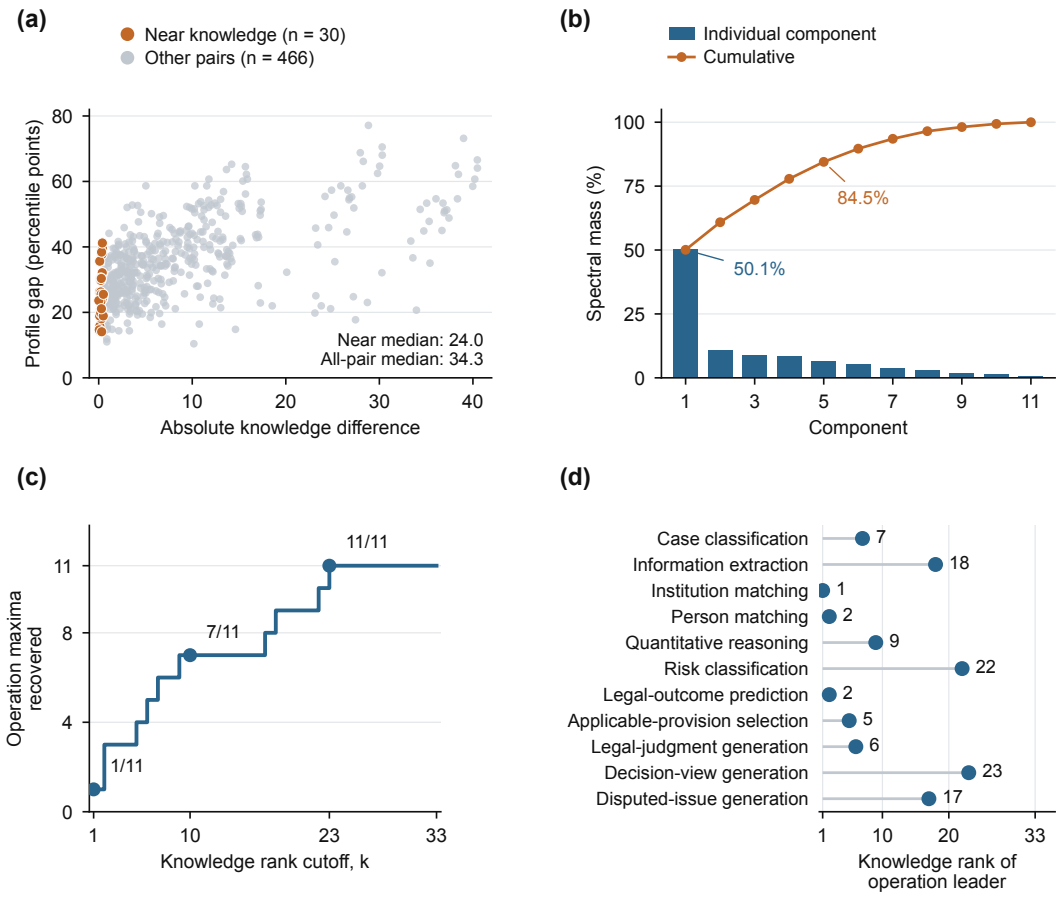

Figure 6: Knowledge proximity and operation profiles. (a) Knowledge differences versus profile gaps. (b) Eigenvalue shares of the operation-rank correlation matrix. (c) Operation maxima recovered at each Knowledge cutoff. (d) Knowledge ranks of eligible operation leaders. Panels (a,b) use 32 complete non-evaluator profiles; (c,d) use operation-specific eligibility.

The ranks locate the leading configurations, while the gaps quantify the operation performance excluded at each cutoff. Extraction is within 0.90 points of its maximum at $k = 5$, although its leader has Knowledge rank 18. Decision-view generation retains a 6.70-point gap at $k = 15$. Reading the ranks alongside these gaps distinguishes exclusion of an exact leader from the magnitude of the remaining performance gap.

F.4 Descriptive Layer Summaries

Figure 3(a) uses the Knowledge macro K_m from Equation 9 and two descriptive means. Let $\mathcal{O}_E$ contain case classification, information extraction, institution matching, person matching, and quantitative reasoning. Let $\mathcal{O}_A$ contain risk classification, legal-outcome prediction, applicable-provision selection, legal-judgment generation, decision-view generation, and disputed-issue generation. Then

$$E_m = \frac{1}{5} \sum_{o \in \mathcal{O}_E} S_{m,o}, \qquad A_m = \frac{1}{6} \sum_{o \in \mathcal{O}_A} S_{m,o}. \tag{31}$$

These means give equal weight to constituent operations, whereas Knowledge gives equal weight to its 42 families. They summarize performance within each layer; operation-specific comparisons retain the underlying scores and contracts.

Knowledge and Evidence-Grounded Processing summaries admit all 33 configurations. Applied Review excludes the fixed evaluator and admits 32. The overview displays eleven selected con-

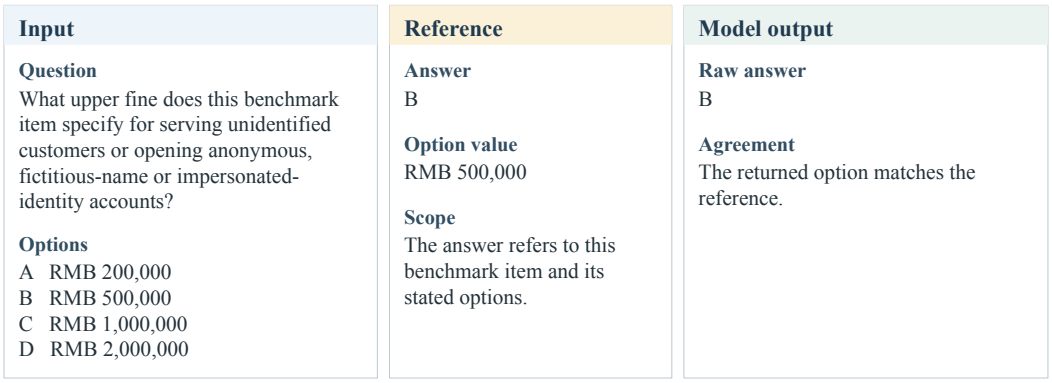

Figure 7: Anti-money-laundering customer-identification compliance under the source-specified benchmark question.

figurations and includes the highest eligible reported score in each layer. The Qwen3-235B-A22B configuration corresponds to the Instruct-2507 endpoint.

G Qualitative Case Studies

The following cases illustrate how evaluation contracts apply to individual inputs, model responses, and scoring decisions. Displayed responses were obtained under the evaluation protocol and complement the aggregate results in Appendix E. Static-case responses are from GPT-5.6-Sol. Chinese inputs and generated passages are translated and abridged for display, with original strings retained for extraction and institution matching. Ellipses indicate omissions. Explanations in reference panels are provided for interpretation and are not model-generated reasoning.

G.1 Domain Knowledge

Figure 7 asks for the source-specified upper bound of an administrative fine associated with anti-money-laundering customer-identification obligations. The recorded response is B (RMB 500,000), matching the reference for this benchmark item. The response contains no additional explanation.

The trade-knowledge item in Figure 8 requires computing an effective protection rate from final-good and input tariffs of 20% and 5%, respectively, with an imported-input share of 0.6. The calculation $(20\% - 0.6 \times 5\%)/(1 - 0.6) = 42.5\%$ uses the domestic value-added share as its denominator. The recorded response, 42.5, matches the reference, with the percentage unit specified in the question. This example belongs to a trade-knowledge family rather than the downstream quantitative-reasoning family.

G.2 Evidence-Grounded Processing

The maritime example in Figure 9 combines waterway transport contracts, delivery records, payment history, and procedural events. The recorded response is D (maritime cargo contract dispute), matching the reference label. The reference panel summarizes the transport relationship underlying the unpaid freight.

Figure 10 requests five ordered fields from a record containing competing amounts and dates. The response returns all five reference values in the required order with the prescribed # delimiter, including the purchase-payment amount of RMB 391,700 rather than the property-offset amount of RMB 361,700. The example contributes five correct matches to the field-weighted score in Equation 6.

The cross-lingual example in Figure 11 requires an identity decision from Chinese and English institution names. The reference panel aligns the location, the transliterated name Saite, the business

Input

Question
Compute the effective protection rate (EPR), as a percentage to one decimal place.

Given values
Final-good tariff: 20%
Input tariff: 5%
Input share of the output price: 0.6

Reference

Answer
42.5%

Calculation
$(20 - 0.6 \times 5) / (1 - 0.6)$
$= 17 / 0.4$
$= 42.5$

Interpretation
The denominator is the domestic value-added share.

Model output

Raw answer
42.5

Agreement
The value matches the reference. The percentage unit is specified in the question.

Figure 8: Effective protection rate in a trade-knowledge item, with the calculation shown as reference analysis.

Input

Case evidence (excerpt)
Two 2012 contracts covered 18,643.44 tonnes of tailings from Laizhou to Huangshi by sea and inland waterway. The agreed freight rate was RMB 85/tonne. Unpaid freight was RMB 814,692.40, with RMB 2,250 in insurance charges. …

Selected options
D Maritime / navigable-water cargo contract
I General contract
E Insurance dispute
Z Loan contract

Reference

Answer
D

Category
Maritime / navigable-water cargo contract

Evidence summary
The dispute concerns freight due for carriage of goods by water.

Model output

Raw answer
D

Agreement
The returned category matches the reference.

Figure 9: Maritime-contract classification from an abridged record, with selected answer options shown.

Input

Case evidence (excerpt)
Purchase debt: RMB 391,700.
Transfer agreement: 4 May 2008.
Property seizure: 9 April 2008.
IOU: RMB 30,000.
Guardianship notarization: 24 April 2008. …

Task
Extract these five fields in this order, using the original Chinese spans and "#" as the delimiter.

Reference

Ordered values
391700 元
2008 年 5 月 4 日
2008 年 4 月 9 日
3 万元
2008 年 4 月 24 日

Serialization
Join these values with "#" in the requested order.

Model output

Raw answer
391700 元 #
2008 年 5 月 4 日 #
2008 年 4 月 9 日 #
3 万元 #
2008 年 4 月 24 日

Field agreement
5 / 5 fields match.

Figure 10: Ordered extraction of five legal fields. Original Chinese spans are retained, with display line breaks after #.

descriptor, and the legal-entity suffix. The recorded Chinese answer means Yes and matches the reference identity label.



Figure 11: Cross-lingual institution matching with names and the model answer in their original form.

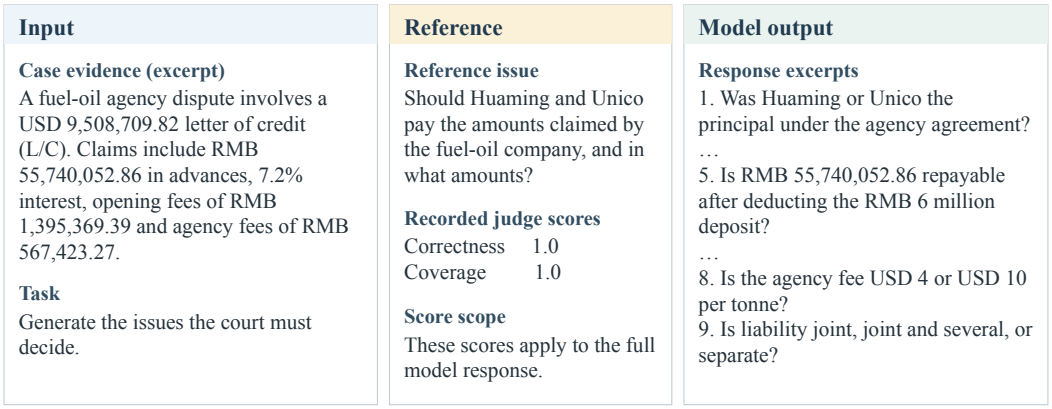

Figure 12: Disputed-issue generation in a letter-of-credit dispute. Excerpts retain the original issue numbers. Recorded scores refer to the full response.

G.3 Applied Review

The two generation examples require an issue list and a reasoned legal analysis. Coverage and correctness are assessed for each response under its family rubric. Within each repetition, the two component means are computed over the family and combined by their harmonic mean (Equation 8).

Figure 12 concerns a fuel-agency and letter-of-credit dispute involving guarantees, insurance, and payments. The response identifies contested questions about the principal's identity, guarantee liability, and the components of the claimed amount, addressing the reference questions of who owes the letter-of-credit-related sums and how much.

Figure 13 concerns a creditor's challenge to an enforcement settlement that transferred the debtor's operating assets. The reference establishes that the settlement is subject to substantive judicial review. The model's full response also analyzes the creditor's standing, the asset transfer, valuation evidence, and the transferee's knowledge before concluding that the settlement should be revoked. This disposition is generated by the model and is not a disposition stated in the supplied reference.

G.4 FinRisk-Ask

The replay cases separate the pre-action context C_t from later evidence used to verify request targets. Later-evidence panels are provided for interpretation and were hidden during model inference.

In Figure 14, the 180-day and 365-day outbound-payment queries return no records. The model selects Ask and requests confirmation of whether the empty results reflect no outbound payments, a query failure, or missing data. Its action matches the recorded Ask transition, and the request receives full alignment credit ($z_t = 1$).

Input

Case evidence (excerpt)
A creditor held a pre-existing RMB 1,893,780 construction claim. A settlement dated 14 September 2011 transferred nearly all operating assets for RMB 1,571,020. The transferee had participated in financing and management. Other debts remained unpaid. …

Task
Generate a reasoned civil judgment addressing the review route and the creditor's avoidance claim.

Reference

Reference excerpt
The execution ruling recorded the settlement and released the seizure. It did not confirm the settlement's substantive validity. The plaintiff may challenge the agreement, and the court should review whether it is revocable.

Recorded judge scores
Correctness 1.0
Coverage 1.0

Model output

Response excerpts
"The settlement has the nature of a civil contract."
…
"The disposition reduced the property available to satisfy creditors."
…
"The settlement … should be revoked in its entirety."

Generated disposition
Revoke the settlement signed on 14 September 2011.

Figure 13: Legal judgment generation in a creditor-avoidance dispute. The reference analysis and generated disposition are distinct. Recorded scores refer to the full response.

Input

Condensed context: Outbound-payment queries for the past 180 days and 365 days both returned empty lists. No actual outbound-payment records were present in the model-visible context.

Reference target excerpt

The system tool did not capture payout activity to a company with the same name as the merchant.

Model response (GPT-5.6-SOL)

The outbound-payment tools for the past 180 days and 365 days both returned zero records. Please confirm whether this merchant truly had no outbound payments, or whether there was a query timeout or missing data.

Recorded action: Ask Model action: Ask
Credit: 1.0

Figure 14: An Ask response about empty outbound-payment query results, shown with request credit $z_t = 1$.

Input

Condensed context: Approximately USD 362,599.11 in outbound payments over the past 365 days was concentrated among three overseas companies. The 180-day query returned no records. Contracts, invoices, and logistics documents for these transactions were not included.

Reference target excerpt

[...] The outbound-payment reviewer replied: transaction records need to be obtained to investigate the purpose of the outbound payments. [...]

Model response (GPT-5.6-SOL)

Approximately USD 362,600 in the merchant's outbound payments over the past 365 days was concentrated among three overseas companies, and the query for the past 180 days returned no outbound-payment records. Are there contracts, invoices, and logistics documents demonstrating genuine trade relationships between the recipients and the merchant?

Recorded action: Ask Model action: Ask
Credit: 0.5

Figure 15: An Ask prediction with partial evidence-request alignment ($z_t = 0.5$).

In Figure 15, the visible record shows outbound payments concentrated among three overseas companies, but contains no contracts, invoices, or logistics documents for these transactions. The model requests these documents to verify the trade relationships between the recipients and the merchant. Its Ask prediction agrees with the recorded action, while the displayed evaluation assigns partial alignment credit ($z_t = 0.5$).

Input

Condensed context: The merchant is linked to a previously terminated merchant across multiple attributes, including device, identity, and contact information.

Reference target excerpt

[...] Exact matches across multiple attributes, including devices, the legal representative's identity document, phone number, and email. [...]

Model response (GPT-5.6-SOL)

<stop />

Recorded action: Ask Model action: Proceed
Credit: 0.0

Figure 16: A Proceed prediction at a recorded Ask state: action disagreement and zero request credit ($z_t = 0$).

Figure 16 shows a later state from the same recorded trajectory as Figure 15, replayed with its own pre-action context containing links to a terminated account associated with money-laundering risk. The model selects Proceed, whereas the recorded action is Ask. No request is generated, so the response receives zero request credit ($z_t = 0$) without invoking the semantic evaluator.

H Limitations and Responsible Use

FinRiskAtlas evaluates Chinese-language financial knowledge, eleven review operations, and evidence-seeking decisions at reconstructed enterprise review states. Its coverage reflects the selected source materials and professional settings. Generalization to other languages, jurisdictions, and review workflows requires further evaluation. Input checks remove answer-bearing metadata and information outside each task's temporal boundary; overlap between public source material and model pretraining data remains unmeasured.

Rankings and screening gaps describe the evaluated configurations under the reported inference settings and scoring contracts. Layer means provide descriptive summaries of native operation scores. Quantitative reasoning contains 58 items, and the paired resampling intervals in Appendix B.6 quantify item variation for fixed configuration contrasts rather than uncertainty in the identity of the best configuration. FinRisk-Ask uses state-weighted metrics, so trajectories with more eligible states contribute more to the corresponding scores.

FinRisk-Ask uses recorded local actions and later-supported evidence needs as references. On the context-only audit subset, sensitivity analyses based on action acceptability preserve the reported leaders and branch-recall pattern (Appendix D). Reasonable alternative actions or requests can extend beyond the recorded trajectory. Offline replay does not execute requests, so their acquisition costs and effects on subsequent review decisions require prospective interactive evaluation.

The three generation operations and request alignment share a fixed semantic evaluator. Blinded validation measures its agreement with expert labels on sampled outputs; the shared evaluator may also contribute to cross-task correlations. Construction and evaluator audits assess reference support and scoring reliability, while expert task-performance baselines under the candidate-model protocol remain to be established.

Institutional approval and de-identification preceded inclusion of enterprise trajectories. Data handling and access arrangements are described in Appendix B.7. FinRiskAtlas supports research evaluation and operation-specific configuration comparison. Use in automated credit, onboarding, or legal decisions requires independent validation, appropriate safeguards, and professional oversight.