FINRISKARENA: TIME SAFE AND MULTILAYER EVALUATION OF FINANCIAL LLM DECISION MAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly considered for high stakes financial decision support, yet conventional benchmark performance provides limited evidence of whether such systems are reliable. Financial risk evaluation must account for temporal validity, differences between imminent and latent risk, operational reliability, evidence grounded decisions, and the validity of the metrics used to assess these properties. We introduce **FinRiskArena**, a time safe and risk sensitive benchmark for multidimensional evaluation of financial LLM decision making. FinRiskArena contains 32,045 company–time episodes from Chinese A share firms constructed under strict point in time information constraints. It separates **imminent distress detection**, which retains legally available near term signals, from **latent distress forecasting**, which removes audit opinions and requires more than 90 days of prediction lead. The benchmark evaluates five complementary layers: **predictive discrimination and calibration, operational reliability, semantic evidence use, counterfactual behavior, and evaluator validity**. A strong structured baseline reaches an AUPRC of 0.439 in the imminent regime, falls to 0.162 after audit opinion information is removed, and reaches 0.047 in the latent regime, showing that superficially similar distress prediction tasks differ substantially in difficulty. LLMs retain meaningful risk ranking signal, but predictive performance does not consistently align with structured output reliability or semantic auditability. Counterfactual tests reveal heterogeneous behavior across perturbations, while a separately frozen blinded financial expert validation shows that selected automatic counterfactual scores diverge from expert semantic judgments. FinRiskArena therefore treats trustworthy financial LLM evaluation as a multilayer measurement problem in which predictive quality, operational compliance, semantic reliability, counterfactual behavior, and evaluator validity must be assessed separately.

1 INTRODUCTION

Large language models are moving from general text generation toward decision support in domains(Hu & Zhao, 2026) where errors can carry material consequences(Kamoi et al., 2024). Financial analysis is a demanding setting for this transition(Tang & Yang, 2025; Peng et al., 2026; Kim et al., 2025). A useful system must do more than produce a plausible answer or a well ranked risk score. The information used in a decision must have been available at the time of prediction. The model must distinguish emerging distress from genuinely early risk(Deng et al., 2025). Its output must be usable by downstream systems, and its stated evidence must support the final decision. These requirements make financial LLM evaluation broader than static question answering or ordinary classification(Yang et al., 2025).

Existing financial LLM benchmarks cover question answering(Zhu et al., 2021), document understanding(Zhu et al., 2024), numerical reasoning(Chen et al., 2021), investment analysis(Li et al., 2025), and increasingly realistic decision settings(Xie et al., 2024). These benchmarks establish that finance requires domain specific evaluation, but the central unit of comparison is usually a task response or task score. FinRiskArena studies a narrower longitudinal decision problem in which the validity of the evaluation depends on when evidence became available, how close the adverse event is, whether the response can be consumed operationally, and whether the supplied evidence supports the model's stated decision.

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.

Corporate distress prediction provides a useful setting because it combines rare adverse outcomes with changing information environments. Classical work progresses from accounting ratio models(Altman, 1968) to probabilistic(Ohlson, 1980), hazard based(Shumway, 2001), and market augmented distress prediction(CAMPBELL et al., 2008). FinRiskArena does not replace this literature with an LLM task. It uses distress prediction to study how benchmark construction, model behavior, and measurement validity interact.

Two temporal issues are central. First, chronological train and test splits do not guarantee point in time correctness(Zhang et al., 2026; Gusak et al., 2025). A financial statement can describe an earlier accounting period while becoming public only after the intended prediction date. A corrected report, audit opinion, or status transition can therefore leak into a retrospective benchmark even if all examples are ordered by year. Recent work on temporal reasoning and evolving knowledge in LLMs highlights the importance of time sensitive evaluation, but point in time data validity is a separate construction problem(Chu et al., 2024; Xiong et al., 2024). Second, the same one year distress label can describe different stages of risk. A firm entering distress within weeks of the cutoff may already have explicit public warnings, whereas a firm entering months later must be identified from weaker evidence. This distinction connects to early warning, survival analysis, and horizon dependent risk modeling(Huang et al., 2025). FinRiskArena therefore separates imminent distress detection from latent distress forecasting rather than treating horizon as a secondary ablation.

Reliability also extends beyond discrimination. Calibration(Huang et al., 2024), selective prediction(Luo et al., 2026), abstention(Sharma et al., 2026), robustness(Shi et al., 2026), and out of distribution behavior describe different failure modes(Zeng et al., 2025). LLMs add further requirements because structured output can be syntactically valid while its evidence use is semantically unsupported. Work on factuality, grounding, attribution, counterfactual testing, and automatic evaluators addresses parts of this problem. FinRiskArena combines these concerns in one financial decision setting and keeps the resulting measurements separate because they use different denominators and support different claims.

Across the study, predictive discrimination, structured compliance, semantic auditability, counterfactual behavior, and evaluator validity repeatedly diverge. FinRiskArena makes these distinctions explicit. Our contributions are fourfold.

1. **Time safe financial LLM evaluation.** We construct a company–time benchmark that enforces point in time information availability, announcement date alignment, censoring aware distress labels, anonymized identities, and fact level provenance. Temporal validity is part of the task definition rather than a post hoc leakage check.
2. **Risk sensitive task decomposition.** We separate imminent distress detection from latent distress forecasting through two regimes. Their performance gap shows that near event recognition and earlier risk forecasting should not be treated as interchangeable tasks.
3. **Multidimensional reliability auditing.** We evaluate financial LLMs across predictive discrimination and calibration, operational compliance, semantic evidence use, abstention and action behavior, and paired counterfactual response. Experiments with structured baselines and two frozen LLM deployments show that these properties can diverge substantially.
4. **Evaluator validity as an explicit audit target.** We extend the audit to the measurement mechanism itself. A blinded validation shows that some automatic counterfactual flags diverge from intended semantic judgments,demonstrating the need to audit construct validity.

2 FINRISKARENA

2.1 TIME SAFE DECISION EPISODES

We observe a population of firms at repeated annual decision cutoffs. The basic unit is a company–time episode

$$E_i = (c_i, t_i, X_i, M_i, Y_i), \tag{1}$$

where c_i is an anonymized company, t_i is the prediction cutoff, X_i contains model visible financial evidence available by t_i, M_i records missing information explicitly, and Y_i is the future outcome used only for evaluation. The current benchmark uses April 30 as the annual cutoff.

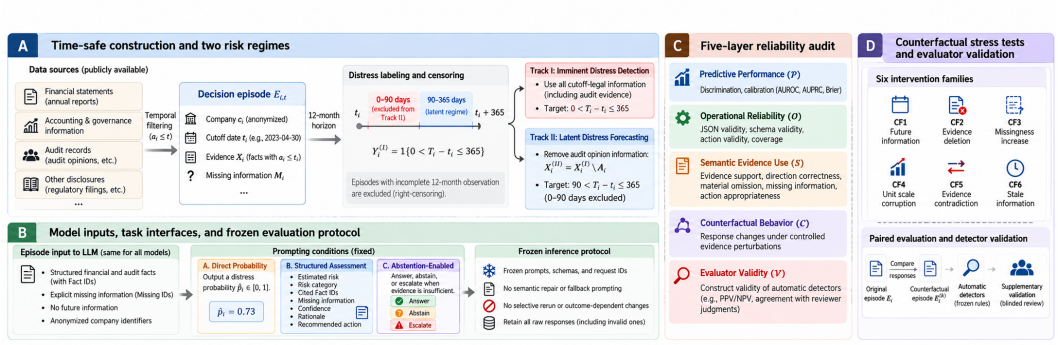

Figure 1: Overview of FinRiskArena Framework

Each fact $x_{ij} \in X_i$ has an availability date d_{ij}. It can enter the episode only when $d_{ij} \leq t_i$. This condition is stricter than checking whether the underlying accounting period ends before the cutoff. A December 31 statement is not considered available at April 30 unless the relevant version had actually been published. Model visible episodes exclude future distress states, post cutoff corrections, future audit information, and other target revealing fields. Company names and ticker symbols are replaced with opaque identifiers. Visible evidence receives Fact IDs, and unavailable fields receive Missing IDs so later audits can trace the evidence used by the model.

The benchmark studies future entry into ST or *ST status, a formally designated severe financial distress state in the Chinese A share market. This target is an institutional distress proxy. It is not treated as equivalent to bankruptcy, bond default, credit default, or corporate failure in general.

A firm is therefore not labeled negative merely because no event appears in the currently observed data. Firms that previously entered distress and later recovered can contribute later episodes if their state at the new cutoff is reliably normal. The primary target concerns future distress entry conditional on the current state rather than only the first lifetime event.

2.2 Imminent and Latent Risk Regimes

A one year label does not indicate how close the event is or what warning signals are already available. We define two evaluation regimes to separate these settings.

Track I, Imminent Distress Detection. Track I retains all information legally available by the cutoff, including eligible audit evidence. Its target is

$$Y_i^{(I)} = \mathbb{I}[T_i \in (t_i, t_i + 365]], \tag{2}$$

subject to the censoring rule above. This track represents realistic short horizon screening in which strong warning signals may already be public.

Track II, Latent Distress Forecasting. Track II removes audit opinion information from the visible evidence,

$$X_i^{(II)} = X_i^{(I)} \setminus A_i \tag{3}$$

where A_i denotes audit-opinion evidence available at the cutoff. Positive events are retained only when $90 < T_i - t_i \leq 365$. Episodes with distress entry within 90 days after the cutoff are excluded from the Track II evaluation population rather than relabeled as negatives. This regime therefore focuses on earlier warning cases in which highly proximate distress events and audit-opinion signals are excluded. Track I and Track II are not a controlled causal comparison. Track II changes both the information set and the event composition, so their difference is interpreted as risk regime heterogeneity rather than the isolated effect of audit opinion removal.

2.3 MULTILAYER RELIABILITY AUDIT

FinRiskArena $\mathcal{A}(f_\theta) = (\mathcal{P}, \mathcal{O}, \mathcal{S}, \mathcal{C}, \mathcal{V})$, represents reliability as a profile of distinct measurements, where $\mathcal{P}$ denotes predictive reliability, $\mathcal{O}$ operational reliability, $\mathcal{S}$ semantic auditability, $\mathcal{C}$ counterfactual behavior, and $\mathcal{V}$ evaluator validity. Figure 1 summarizes the audit layers.

Predictive reliability. The predictive layer measures whether model probabilities rank future distress and whether they are calibrated. We report AUROC, AUPRC, Brier score, and, where relevant, screening metrics at a fixed review budget. Because distress is rare, AUPRC is also interpreted relative to the target prevalence π, $\text{relative AUPRC} = \frac{\text{AUPRC}}{\pi}$. These metrics answer whether the model contains useful risk signal. They do not establish whether the output is operationally usable or semantically supported.

Operational reliability. Operational reliability measures whether planned requests produce usable responses. Let O_i^{json}, O_i^{schema}, $O_i^{\text{action}} \in \{0, 1\}$ indicate valid JSON, valid schema, and valid action fields where applicable. The natural denominator is all planned requests. This preserves a distinction between predictive performance conditional on an evaluable response and end to end deployment reliability.

Semantic auditability. For structured responses, we examine evidence support, evidence direction, material omission, missing information handling, action appropriateness, semantic hallucination, outcome leakage, and overall auditability. The audit concerns observable response content rather than hidden chain of thought. A syntactically valid Fact ID establishes traceability but does not establish that the cited fact supports the decision.

Counterfactual behavior. For selected episode E_i, a transformation T_k creates a counterfactual episode $E_i^{(k)} = T_k(E_i)$, where k indexes one of six stress tests: future information insertion, critical evidence deletion, increased missingness, unit scale corruption, contradictory evidence, and stale information. The model responses are compared as a pair, $\Delta_i^{(k)} = \Phi_k(R_i, R_i^{(k)})$, where Φ_k is specific to the construct. The objective is to diagnose how a targeted change in evidence affects the decision, not to aggregate all perturbations into one safety score.

Evaluator validity. Counterfactual analysis introduces a measurement problem because the intended semantic failure may not be directly observable at scale. Let $Z_i^{(k)}$ denote the semantic behavior of interest and $\hat{Z}_i^{(k)}$ the label produced by automatic detector h_k, $\hat{Z}_i^{(k)} = h_k(R_i, R_i^{(k)}, E_i, E_i^{(k)})$. Automatic rates over $\hat{Z}$ support semantic claims only to the extent that the detector aligns with the intended construct Z. Where separate semantic validation is available, we examine agreement, false positive and false negative patterns, and quantities such as PPV and NPV. This layer separates a valid stress test from a valid automatic interpretation of the response.

3 BENCHMARK CONSTRUCTION

3.1 DATA, LABELS, AND TEMPORAL SAFETY

FinRiskArena is constructed from longitudinal records for publicly listed Chinese A-share firms. The source data include annual financial statements, accounting and governance variables, audit information, and recorded distress status transitions. We align these sources into a company–time panel in which every observation has both an accounting period and an information availability time.

The final supervised panel contains 32,045 company–cutoff episodes from 2017 to 2024, including 625 positive future distress entries and 31,420 negative episodes. Severe distress is rare in the natural population, so the structured evaluation retains the natural imbalance.

April 30 was selected before model evaluation after auditing candidate cutoffs. The main role of the cutoff is to define a realistic information boundary while retaining broad annual report coverage. Figure 2 distinguishes disclosure time from prediction horizons.

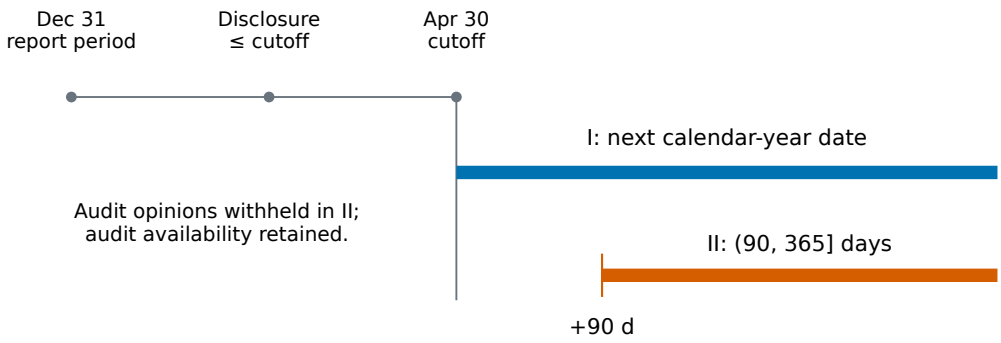

Figure 2: Temporal eligibility and frozen task endpoints.

Table 1: Benchmark populations. Train: 2017–2021; Dev: 2022; Test: 2023–2024. Representative rates are enriched sample fractions, not population prevalence.

Population	N	Positive	Rate
Train	17,407	351	2.02%
Dev	4,508	56	1.24%
Test	10,130	218	2.15%
Representative I	1,000	218	21.80%
Representative II	892	110	12.33%

3.2 Splits and Evaluation Populations

The chronological split uses 2017–2021 for training, 2022 for development, and 2023–2024 for testing. The Population Test contains 10,130 episodes and 218 positives. We additionally report a company disjoint test as an auxiliary generalization setting.

Repeated LLM evaluation over the entire structured test set is expensive across multiple interfaces and counterfactual variants. We therefore freeze representative LLM sets drawn from the Population Test. Track I contains 1,000 episodes, with 218 positives and 782 negatives. Track II contains 892 episodes, with 110 positives and the same 782 negatives. These sets deliberately enrich positive cases for diagnosis and do not preserve natural prevalence. Frozen weights restore the target prevalence for population level predictive metrics, approximately 2.15% for Track I and 1.10% for Track II.

Model visible episodes use opaque company identifiers, while outcome labels and evaluator only provenance are stored separately. Table 1 reports the population and enriched sample denominators. Complete cutoff selection results, population tables, feature details, diagnostic sampling rules, and reproducibility manifests are provided in the appendix.

4 Experimental Setup

We first establish a structured financial reference using conventional machine learning models under the chronological split. Model selection is performed only on the development year. The selected primary model is a regularized logistic regression baseline evaluated with full evidence, after removing audit information, and under the latent Track II regime. Company disjoint performance is reported separately.

We evaluate two frozen LLM deployments: DeepSeek-v4-Flash through an OpenAI compatible chat completions interface and Qwen3.8-27B-FP8 served locally with vLLM. The latter combines a quantized model with a local serving stack, so cross deployment differences are descriptive rather than universal model family rankings.

Each representative episode is evaluated under three fixed interfaces. **Condition A** requests only a distress probability. **Condition B** requests a structured assessment containing risk, cited Fact IDs, missing information references, confidence, a concise rationale, and a recommended action. **Condition C** extends the structured interface by allowing answer, abstain, or escalate actions.

The representative evaluation contains 1,892 track specific episodes. Under three interfaces, each deployment receives 5,676 formal requests. Prompts, schemas, and request identifiers are frozen before execution. We apply no semantic repair, fallback prompting, selective reruns, or outcome dependent postprocessing. Raw invalid and schema nonconforming outputs remain part of the evaluation.

4.1 Predictive, Operational, and Semantic Evaluation

Population level predictive evaluation reports AUROC, AUPRC, AUPRC normalized by prevalence, and Brier score. The structured baselines additionally report Recall@Top-5%, Precision@Top-5%, and Lift@Top-5% as fixed budget screening metrics. Predictive metrics for the enriched LLM samples use frozen population restoring weights.

Predictive and operational denominators are kept separate. Predictive metrics are computed when the required prediction is evaluable under the frozen parser. Operational metrics use all planned requests and report JSON validity, schema validity, action validity where applicable, and answer coverage under Condition C. This separation avoids both silently dropping unusable outputs and treating every formatting failure as an ordinary numeric prediction error.

Structured output validity does not establish correct evidence use. FinRiskArena therefore performs a semantic audit on a stratified diagnostic response set. The frozen audit covers 222 responses and the semantic audit was conducted by financial experts using a frozen eight-dimension rubric. Reviewers assessed observable response semantics, including evidence support, direction correctness, material omission, missing-information handling, action appropriateness, semantic hallucination, outcome leakage, and overall auditability. Disagreements were resolved under a predefined adjudication procedure. Each audited response was independently assessed by two financial experts, with disagreements resolved by adjudication under the frozen rubric.

4.2 Counterfactual and Evaluator Validation

Natural distribution evaluation cannot isolate how a model responds when a specific property of the evidence changes. We therefore construct paired counterfactual episodes. For an original episode E_i, perturbation family k produces $E_i^{(k)} = T_k(E_i)$. The six perturbation families target future information leakage, critical evidence deletion, increased missingness, unit scale corruption, evidence contradiction, and stale information. Each transformation follows a minimal pair principle and changes the target construct while preserving unrelated information as far as possible.

Each deployment receives 1,440 counterfactual requests, for 2,880 total. Original responses are reused from the frozen representative evaluation rather than regenerated. We report a conditional view over construct applicable, parser valid responses and an end to end view that retains the full request set under the frozen fail closed rule.

Automatic counterfactual labels are reported as detector positive outcomes unless semantic construct validity has been established. When early validation raised uncertainty for selected detectors, those rules were not retuned. A separate expert-validation protocol was frozen before the automatic labels were unblinded. It contains 96 stratified CF4/CF6 cases, each independently reviewed by two financial experts using behavior-based rubrics. Mandatory disagreements were adjudicated by a third financial expert, and expert consensus was frozen before the automatic detector labels were revealed. The stratified sample is used to assess construct validity rather than population failure prevalence.

5 RESULTS

5.1 RISK REGIME DETERMINES TASK DIFFICULTY

The structured baseline first establishes the difficulty of the underlying financial task. Under the full time safe information regime, the selected model reaches AUROC 0.886 and AUPRC 0.439 on the Population Test. With a distress prevalence of approximately 2.15%, this AUPRC is far above the random precision recall baseline.The structured baseline remains strong under the company-disjoint evaluation, reaching an AUROC of 0.877 and an AUPRC of 0.448, suggesting that the observed risk signal is not driven solely by repeated observations of firms seen during training.

Performance drops sharply when near term audit information is removed. On the same population test, AUPRC falls from 0.439 to 0.162. Under the latent regime, which removes audit opinion information and requires more than 90 days of event lead, AUPRC is approximately 0.047. Recall among the top 5% of predicted risk also decreases from about 59.6% in the full setting to 34.4% without audit information and 28.2% in the latent regime.

These results establish risk regime heterogeneity. They do not isolate the causal effect of audit opinions because Track II also changes minimum lead time and event composition. The relevant conclusion is that near term detection and latent forecasting measure different information environments. Table 2 gives the structured reference.

Table 2: Dev-selected logistic baselines. Full and No audit share the natural Population Test (10,130 episodes; 218 positives). Latent uses 10,022 episodes and 110 positives; it changes information and event composition.

Setting	AUROC	AUPRC	Brier	Recall@5%
Full	0.886	0.439	0.016	59.6%
No audit	0.842	0.162	0.020	34.4%
Latent	0.790	0.047	0.011	28.2%

Both frozen LLM deployments produce above random precision recall performance in Track I and Track II. In Track I, AUPRC relative to prevalence ranges from approximately 6.2 to 8.3 for DeepSeek and 10.4 to 16.1 for Qwen across prompting conditions. In Track II, the corresponding ranges are approximately 4.1 to 8.9 and 3.7 to 11.3. No prompting condition dominates both tracks and all metrics. Full model-by-track-by-condition results are reported in the appendix.

5.2 PREDICTIVE, OPERATIONAL, AND SEMANTIC RELIABILITY DIVERGE

Predictive signal does not guarantee reliable structured output. DeepSeek produces valid JSON for approximately 99.6% of planned representative requests and satisfies the full schema for approximately 98.3%. The corresponding Qwen rates are approximately 76.2% and 73.6%. Both deployments nevertheless retain nonrandom conditional risk ranking signal. Condition C also shows that selective behavior is not equivalent to improved predictive reliability: abstain or escalate cases are more risk enriched and more missing heavy, but the answered subset does not consistently achieve better precision recall performance. Figure 3 and Table 3 separate conditional prediction from operational validity.

The semantic audit reveals a second separation. Among finalized labels for each dimension, 72.9% receive the highest Evidence Support category, 47.2% the highest Direction Correctness category, 56.8% the highest Missing Information Accuracy category, 48.2% the highest Action Appropriateness category, and 54.7% the highest Overall Auditability category. Material omission is identified in 8.6% of audited cases. These rates use dimension specific finalized denominators and characterize the diagnostic audit rather than the population. Figure 4 shows finalized-label semantic denominators. A valid structured response can therefore remain semantically weak. Schema compliance establishes that the response satisfies the output contract.

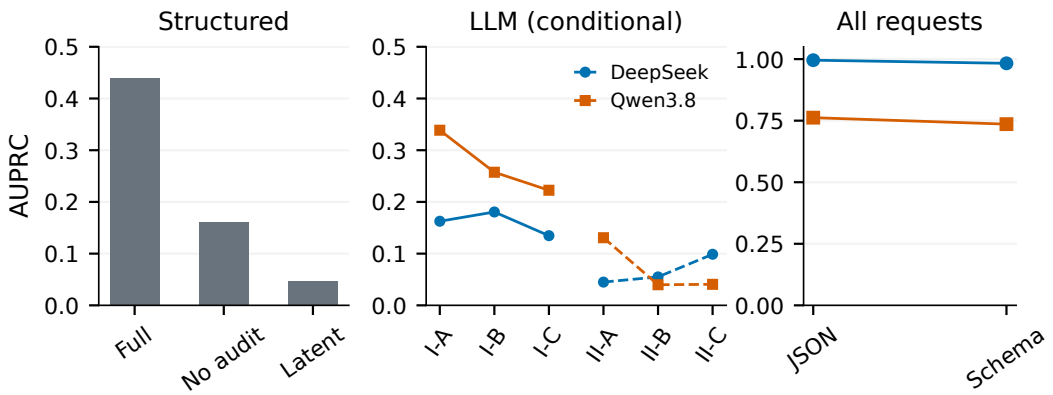

Figure 3: Distinct predictive and operational views. Left: natural-population structured tests (latent composition differs). Center: reweighted LLM AUPRC conditional on valid responses; solid/dashed lines denote Tracks I/II, with A/B/C interfaces. Right: JSON/schema rates over all planned requests.

Table 3: Operational reliability on all 5,676 planned requests per deployment. Full A/B/C predictive results are in Appendix C.

Deployment	JSON	Schema	C action
DeepSeek	99.6%	98.3%	98.7%
Qwen3.8	76.2%	73.6%	70.8%

5.3 Counterfactual tests reveal construct specific behavior

The six counterfactual families produce heterogeneous automatic detector profiles. Future information and critical evidence deletion have low positive rates under the frozen rules. Missingness produces moderate rates: the corrected conditional CF3 rates are approximately 18.1% for DeepSeek and 25.7% for Qwen. Evidence contradiction produces approximately 15.5% and 3.8%, respectively. The unit scale and stale information detectors produce much larger automatic positive rates. Full conditional and end to end results are reported in the appendix.

These results are construct specific. They do not support a single global counterfactual safety ranking, and automatic positive rates are not interpreted as semantic failure prevalence without validation.

A separately frozen supplementary study evaluates the two measurement sensitive detectors. The final semantic consensus contains 44 SAFE, 3 FAIL, and 1 INDETERMINATE case for CF4, and 48 SAFE with no FAIL cases for CF6. Among expert-reviewed cases with a determinate consensus label, none of the automatic FAIL labels are confirmed as semantic FAIL for either detector. Both detectors are therefore classified as not validated for the stronger semantic interpretation suggested by their raw positive rates. The original rules remain unchanged rather than being retuned on the validation sample.

Table 4 retains construct-specific automatic counts. The results support the central empirical claim of FinRiskArena $\mathcal{P} \neq \mathcal{O} \neq \mathcal{S} \neq \mathcal{C} \neq \mathcal{V}$, where the notation indicates that these reliability dimensions are not interchangeable, not that they are statistically independent.

6 Discussion, Limitations, and Conclusion

FinRiskArena shows why high stakes LLM evaluation should preserve separate reliability measurements rather than compress them into a single score. The information regime changes apparent task difficulty. LLMs can retain useful financial risk signal while differing sharply in structured output reliability. Structured compliance does not ensure semantically supported evidence use. Counter-

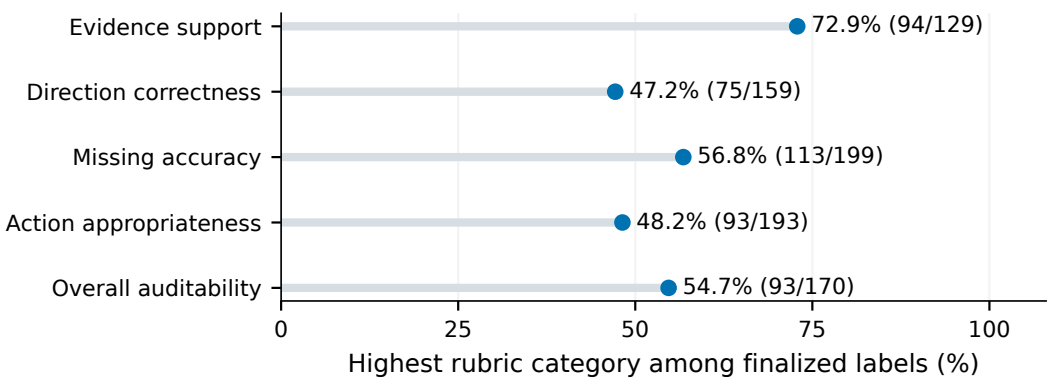

Figure 4: Financial expert semantic audit of 222 diagnostically sampled responses. Bars show the share receiving the highest rubric category among finalized labels; denominators are dimension specific.

Table 4: Conditional automatic counterfactual detector-positive counts. Full end-to-end rates and intervals are in Appendix E.

Model	CF1	CF2	CF3	CF4	CF5	CF6
DeepSeek	0/230	0/233	41/226	234/234	34/219	217/230
Qwen	5/159	0/160	29/113	147/164	6/159	136/157

factual behavior depends on the perturbation and evaluation denominator. Selected automatic safety rules can also fail to align with the semantic construct they are intended to measure.

This last result treats the evaluator as part of the measurement system. Validating an evaluator and developing a better evaluator are different activities. Retuning a detector on the same cases that revealed its measurement error would convert the validation set into a development set. The current study therefore preserves the original rules and reports their limitation rather than retrospectively optimizing them.

The study has several boundaries. FinRiskArena is based on Chinese A-share firms and an institutional ST/*ST distress proxy, so its numerical findings may not transfer directly to other markets or default definitions. We evaluate two primary LLM deployments, and the local Qwen results reflect both the model and serving stack. The semantic audit uses diagnostic sampling, so its reported rates characterize the audited sample rather than population-wide semantic error prevalence. Expanded construct validation currently covers selected counterfactual detectors rather than every semantic metric. Finally, the current benchmark evaluates single step episode-to-response decisions rather than multistep tool using agents.

We introduced FinRiskArena as a time safe benchmark for financial LLM decision making that separates imminent from latent distress and evaluates reliability beyond prediction alone. Across structured baselines and two LLM deployments, predictive quality, operational usability, semantic auditability, counterfactual behavior, and evaluator validity do not consistently align. High stakes LLM evaluation should therefore audit the decision task, the model response, and the measurement mechanism separately rather than treating any single benchmark score as sufficient evidence of reliability.

REPRODUCIBILITY STATEMENT

The study uses frozen splits, prompts, outputs, weighting rules, evaluator settings and consensus records. The accompanying offline audit records source hashes and table-level provenance. No inference or statistical model fitting was repeated for this manuscript. Appendix details distinguish conditional predictive performance from planned-request operational denominators. Upon completion of data curation and release preparation, we will make all experiment-ready data, evaluation artifacts, and reproducibility materials publicly available on GitHub.

AI USE STATEMENT

AI assistance was used for manuscript polishing, code debugging, and figure preparation. AI tools were also used to assist annotation workflow preparation, but all expert labels and adjudicated judgments reported in the paper were independently determined and finalized by financial experts.

REFERENCES

Edward I. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4):589–609, 1968. doi: https://doi.org/10.1111/j.1540-6261.1968.tb00843.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1968.tb00843.x.

JOHN Y. CAMPBELL, JENS HILSCHER, and JAN SZILAGYI. In search of distress risk. *The Journal of Finance*, 63(6):2899–2939, 2008. doi: https://doi.org/10.1111/j.1540-6261.2008.01416.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2008.01416.x.

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 Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3697–3711, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.300. URL https://aclanthology.org/2021.emnlp-main.300/.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. TimeBench: A comprehensive evaluation of temporal reasoning abilities in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1204–1228, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.66. URL https://aclanthology.org/2024.acl-long.66/.

Shangkun Deng, Yongqi Li, Yingke Zhu, Bingsen Wang, Hong Ning, Siyu Yi, and Tatsuro Shimada. Financial distress warning and risk path analysis for chinese listed companies: An interpretable machine learning approach. *Economic Modelling*, 152:107288, 2025. ISSN 0264-9993. doi: https://doi.org/10.1016/j.econmod.2025.107288. URL https://www.sciencedirect.com/science/article/pii/S0264999325002834.

Danil Gusak, Anna Volodkevich, Anton Klenitskiy, Alexey Vasilev, and Evgeny Frolov. Time to split: Exploring data splitting strategies for offline evaluation of sequential recommenders. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pp. 874–883. ACM, 2025. doi: 10.1145/3705328.3748164. URL http://dx.doi.org/10.1145/3705328.3748164.

Xiaoyu Hu and Jinman Zhao. Fin-bias: Comprehensive evaluation for LLM decision-making under human bias in finance domain. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*,

pp. 5678–5694, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.279. URL https://aclanthology.org/2026.findings-acl.279/.

Ling Huang, Yucheng Xing, Swapnil Mishra, Thierry Denœux, and Mengling Feng. Evidential time-to-event prediction with calibrated uncertainty quantification. *International Journal of Approximate Reasoning*, 181:109403, 2025. ISSN 0888-613X. doi: https://doi.org/10.1016/j.ijar.2025.109403. URL https://www.sciencedirect.com/science/article/pii/S0888613X25000441.

Yukun Huang, Yixin Liu, Raghuveer Thirukovalluru, Arman Cohan, and Bhuwan Dhingra. Calibrating long-form generations from large language models, 2024. URL https://arxiv.org/abs/2402.06544.

Ryo Kamoi, Sarkar Snigdha Sarathi Das, Renze Lou, Jihyun Janice Ahn, Yilun Zhao, Xiaoxin Lu, Nan Zhang, Yusen Zhang, Ranran Haoran Zhang, Sujeeth Reddy Vummanthala, Salika Dave, Shaobo Qin, Arman Cohan, Wenpeng Yin, and Rui Zhang. Evaluating llms at detecting errors in llm responses. *arXiv preprint arXiv:2404.03602*, 2024.

Alex Kim, Maximilian Muhn, and Valeri Nikolaev. Financial statement analysis with large language models, 2025. URL https://arxiv.org/abs/2407.17866.

Haohang Li, Yupeng Cao, Yangyang Yu, Shashidhar Reddy Javaji, Zhiyang Deng, Yueru He, Yuechen Jiang, Zining Zhu, K.P. Subbalakshmi, Jimin Huang, Lingfei Qian, Xueqing Peng, Jordan W. Suchow, and Qianqian Xie. INVESTORBENCH: A benchmark for financial decision-making tasks with LLM-based agent. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2509–2525, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.126. URL https://aclanthology.org/2025.acl-long.126/.

Gaoxiang Luo, Yifan Wu, Sinian Zhang, Aryan Deshwal, and Ju Sun. Aligning language models with selective prediction, 2026. URL https://arxiv.org/abs/2607.03528.

Ja Ohlson. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1):109–131, 1980. URL https://EconPapers.repec.org/RePEc:bla:joares:v:18:y:1980:i:1:p:109-131.

Xueqing Peng, Lingfei Qian, Yan Wang, Ruoyu Xiang, Yueru He, Yang Ren, Mingyang Jiang, Vincent Jim Zhang, Yuqing Guo, Jeff Zhao, Huan He, Yi Han, Yun Feng, Yuechen Jiang, Yupeng Cao, Haohang Li, Yangyang Yu, Xiaoyu Wang, Penglei Gao, Shengyuan Lin, Keyi Wang, Shanshan Yang, Yilun Zhao, Zhiwei Liu, Peng Lu, Jerry Huang, Suyuchen Wang, Triantafillos Papadopoulos, Polydoros Giannouris, Efstathia Soufleri, Nuo Chen, Zhiyang Deng, Heming Fu, Yijia Zhao, Mingquan Lin, Meikang Qiu, Kaleb E Smith, Arman Cohan, Xiao-Yang Liu, Jimin Huang, Guojun Xiong, Alejandro Lopez-Lira, Xi Chen, Junichi Tsujii, Jian-Yun Nie, Sophia Ananiadou, and Qianqian Xie. MultiFinBen: Benchmarking large language models for multilingual and multimodal financial application. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16934–16963, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.770. URL https://aclanthology.org/2026.acl-long.770/.

Ankit Sharma, Nachiket Tapas, and Jyotiprakash Patra. Improving llm reliability through hybrid abstention and adaptive detection, 2026. URL https://arxiv.org/abs/2602.15391.

Boshen Shi, Yize Liu, Chen Zhao, Ce Chi, Zhendong Wang, Xing Wang, and Junlan Feng. Trustdabench: Benchmarking reliability and robustness of llms for structured data analysis, 2026. URL https://arxiv.org/abs/2608.24145.

Tyler Shumway. Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, 74(1):101–124, 2001.

Yixuan Tang and Yi Yang. FinMTEB: Finance massive text embedding benchmark. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3620–3638, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.179. URL https://aclanthology.org/2025.emnlp-main.179/.

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, 2024. URL https://arxiv.org/abs/2402.12659.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. Large language models can learn temporal reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10452–10470, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.563. URL https://aclanthology.org/2024.acl-long.563/.

Yuzhe Yang, Yifei Zhang, Yan Hu, Yilin Guo, Ruoli Gan, Yueru He, Mingcong Lei, Xiao Zhang, Haining Wang, Qianqian Xie, Jimin Huang, Honghai Yu, and Benyou Wang. UCFE: A user-centric financial expertise benchmark for large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5444–5463, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.300. URL https://aclanthology.org/2025.findings-naacl.300/.

Cong Zeng, Shengkun Tang, Yuanzhou Chen, Zhiqiang Shen, Wenchao Yu, Xujiang Zhao, Haifeng Chen, Wei Cheng, and Zhiqiang Xu. Human texts are outliers: Detecting llm-generated texts via out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 163483–163513, 2025. doi: 10.52202/085713-5457.

Fan Zhang, Zhen Li, Sijia Peng, and Yu Chen. When alpha disappears: A one-switch benchmark for decision-time leakage in financial backtests, 2026. URL https://arxiv.org/abs/2605.23959.

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 Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3277–3287, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.254. URL https://aclanthology.org/2021.acl-long.254/.

Jie Zhu, Junhui Li, Yalong Wen, and Lifan Guo. Benchmarking large language models on CFLUE - a Chinese financial language understanding evaluation dataset. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5673–5693, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.337. URL https://aclanthology.org/2024.findings-acl.337/.

A Benchmark Construction Details

A.1 Data Source and Population

FinRiskArena is constructed from longitudinal records for publicly listed Chinese A-share firms. The source data include annual financial statements, accounting and reporting variables, governance information, audit records, and formal distress status transitions.

The primary structured task retains the natural rarity of distress rather than artificially balancing the evaluation population.

A.2 Point-in-Time Episode Construction

For company c at cutoff t, an episode is $E_i = (c_i, t_i, X_i, M_i, Y_i)$. Evidence eligibility is recomputed independently at every cutoff.

The benchmark distinguishes accounting period, fiscal year end, initial disclosure date, correction or restatement date, audit publication date, distress event date. A fact that becomes legal at a later cutoff is never copied retrospectively into an earlier episode.

For financial statements, the primary construction uses consolidated annual reports with December 31 fiscal year endpoints. Parent-only statements and records outside the intended annual period are excluded from the main feature panel.

When multiple versions exist, only the version available by the cutoff is eligible. Corrections released after the cutoff are not inserted into earlier episodes.

A.3 Annual Cutoff Selection

The annual cutoff is April 30. This boundary was selected before formal model evaluation after comparing March 31, April 30, May 31, and June 30. In the 2017–2024 cutoff audit, April 30 annual-report availability was 95.77%, while time-safe three-statement coverage was 85.50%. Moving the cutoff to May 31 or June 30 increased pooled time-safe coverage by only 0.21 and 0.51 percentage points, respectively.

The cutoff is therefore treated as a decision boundary rather than a simple calendar split. The purpose is to retain broad annual-report availability while preserving useful lead time before subsequent distress events.

A.4 Distress-State Reconstruction

Distress records are reconstructed as a temporal state sequence rather than as isolated binary labels. Conceptually, a firm can occupy states such as NORMAL, ST, *ST, other distress, UNKNOWN.

An episode enters the primary supervised population only when the firm's state at the cutoff is reliably known.

Let T_i denote the first eligible distress entry after cutoff t_i. Track I uses

$$Y_i^{(I)} = \begin{cases} 1, & 0 < T_i - t_i \le 365, \\ 0, & \text{the complete 365-day window is observed and no event occurs,} \\ \text{NA}, & \text{future observation is incomplete or the cutoff state is unresolved.} \end{cases}$$

This prevents right-censored firms from being silently labeled negative.

Firms that previously entered distress and later recovered can contribute later episodes when their state at the new cutoff is reliably normal. The primary task therefore predicts future distress entry conditional on the current state rather than only the first lifetime distress event.

A.5 Representative and diagnostic LLM evaluation sets

A separate diagnostic pool enriches false positives, false negatives, borderline predictions, missing heavy episodes, unusual audit evidence, latent positives, repeated distress histories, and company disjoint examples. This set is used for semantic auditing rather than population frequency estimation.

A.6 Identity protection, label isolation, and freeze discipline

Benchmark construction is frozen before formal LLM inference. The pipeline records source hashes, split manifests, episode identifiers, prompt versions, weighting rules, and label generation rules. Later analyses can change the interpretation of a metric, but they do not alter the underlying frozen responses, episode population, or label definitions. This separation between pre inference correction and post hoc adaptation is central to the benchmark audit trail.

B Experimental Protocol supplement

B.1 Structured financial baselines

The baseline is used as a reference rather than as a claim that logistic regression is the optimal distress prediction architecture. It quantifies the predictability of the financial task, measures dependence on near cutoff audit information, and provides a natural distribution reference for the LLM results. We evaluate the full information setting, a no audit setting, a standard audit only analysis where relevant, the latent risk regime, and the Company OOD split. Company-OOD Test: $N = 2{,}632$, 67 positives, AUROC $= 0.877$, AUPRC $= 0.448$.

B.2 LLM deployments and prompting conditions

Both deployments receive the same frozen episode representation, task instructions, Fact ID conventions, Missing ID conventions, and output schema definitions. Each representative episode is evaluated under the three conditions.

Condition A requests a direct distress probability $\hat{p}_i \in [0, 1]$. It places the least structure on the response and primarily measures predictive discrimination and calibration.

Condition B requests a structured risk assessment that includes evidence references, missing information, confidence, a concise assessment, and a recommended action. This interface makes the relation between supplied evidence and the observable decision fields auditable.

Condition C extends the structured interface by allowing abstention or escalation when the model judges the available evidence insufficient for a reliable decision. The benchmark evaluates whether this selective behavior is associated with genuinely difficult or insufficiently supported cases rather than assuming abstention is beneficial by definition.

The three conditions are evaluation interfaces applied to the same frozen model deployments. Performance differences across them are reported as prompt dependent observations, not causal effects of prompting. The formal inference protocol uses no response repair, fallback prompt, or selective rerun.

B.3 Predictive and operational metrics

This distinction prevents a deployment with many unusable responses from appearing reliable only because its valid subset performs well. It also avoids treating a formatting failure as if it were an ordinary probability error.

B.4 Selective prediction and abstention

Condition C allows the model to abstain or escalate. We measure answer coverage and compare the properties of answered and unanswered cases. Abstention receives no automatic safety credit. A useful selective policy should identify cases where evidence is genuinely insufficient or difficult, and any coverage quality tradeoff must be demonstrated empirically.

Table 5: DeepSeek-V4-Flash predictive results using frozen population weights, conditional on evaluator-included responses. AP/prev. uses weighted prevalence within that valid subset, not the complete target population.

Track	Cond.	Valid N	AUROC	AUPRC	Brier	AP/prev.
I	A	1000	0.683	0.163	0.087	7.6
I	B	970	0.707	0.181	0.078	8.3
I	C	975	0.599	0.135	0.085	6.2
II	A	892	0.661	0.045	0.126	4.1
II	B	871	0.652	0.055	0.106	5.1
II	C	871	0.664	0.099	0.095	8.9

Table 6: Qwen3.8-27B-FP8 predictive results using frozen population weights, conditional on evaluator-included responses. AP/prev. uses weighted prevalence within that valid subset, not the complete target population.

Track	Cond.	Valid N	AUROC	AUPRC	Brier	AP/prev.
I	A	831	0.826	0.339	0.049	16.1
I	B	699	0.832	0.257	0.072	12.8
I	C	719	0.805	0.223	0.072	10.4
II	A	741	0.750	0.131	0.053	11.3
II	B	595	0.808	0.040	0.067	4.2
II	C	592	0.791	0.041	0.089	3.7

B.5 Statistical uncertainty and reproducibility

Where appropriate, uncertainty estimates preserve the paired and stratified structure of the benchmark. Paired comparisons use the frozen bootstrap procedure rather than treating repeated episodes as unrelated observations. Counterfactual confidence intervals use the preregistered bootstrap configuration. Evaluator PPV and NPV use appropriate binomial intervals. We do not reinterpret tests on operational validity as evidence about semantic safety.

Every formal stage produces frozen artifacts, including episode manifests, request IDs, model visible inputs, raw responses, parser outputs, metric tables, bootstrap settings, reviewer workbooks, consensus records, and SHA 256 hashes. Later analyses reference these artifacts directly. This audit trail allows a reported statistic to be traced to the response and episode from which it was computed.

C Full deployment results

Predictive metrics are conditional on evaluable responses, with frozen population weights. Tables 5 and 6 retain all A/B/C settings.

D Semantic audit details

The semantic audit evaluates 222 diagnostically sampled responses along eight observable dimensions: evidence support, direction correctness, material omission, missing information accuracy, action appropriateness, semantic hallucination, outcome leakage, and overall auditability. A valid Fact ID provides traceability but does not prove that the referenced evidence supports the conclusion. Table 7 retains frozen finalized-label denominators.

E Counterfactual automatic endpoints

For original episode E_i, counterfactual family k produces $E_i^{(k)} = T_k(E_i)$. The six frozen families are shown in Table 8.

Table 7: Frozen financial-expert semantic audit. Category 2 is the highest ordinal category. Denominators are finalized labels, not applicable-only cases; missing-information and action categories each retain 54 explicit NA labels.

Dimension	Category	Count	Finalized N
evidence support	2	94	129
direction correctness	2	75	159
missing information accuracy	2	113	199
action appropriateness	2	93	193
overall auditability	2	93	170
material omission	yes	19	222
semantic hallucination	yes	0	222

Table 8: Counterfactual perturbation constructs

Construct	Perturbation
CF1	Future information leakage
CF2	Critical evidence deletion
CF3	Increased missingness
CF4	Unit scale corruption
CF5	Evidence contradiction
CF6	Stale information

Tables 9 and 10 retain the original automatic endpoint intervals, not semantic-failure intervals.

F EVALUATOR CONSTRUCT VALIDITY

A counterfactual experiment has two separate validity requirements. The intervention must instantiate the intended stress condition, and the scoring rule must identify the intended response behavior. Early validation raised measurement uncertainty for selected detectors. We therefore froze a supplementary validation protocol before additional review instead of changing the detector after seeing the results. Table 11 and Figures 5, 6, and 7 summarize construct alignment, detector-positive rates, and all-request operational reliability.

Predictive evaluation measures whether model generated risk estimates discriminate future distress. Because the representative LLM samples are enriched for positive cases, population level predictive metrics use the frozen population restoring weights defined in Section 4. We report AUROC, AUPRC, prevalence normalized AUPRC, and the Brier score. AUROC measures pairwise ranking performance,

$$\mathrm{AUROC} = P(\hat{p}^{+} > \hat{p}^{-}),$$

while AUPRC is emphasized because distress is rare and its random baseline is approximately the outcome prevalence π. We therefore also report

$$\frac{\mathrm{AUPRC}}{\pi},$$

which measures precision recall performance relative to the underlying class prevalence. Calibration is summarized with the weighted Brier score,

$$\mathrm{Brier} = \frac{1}{N}\sum_{i} w_i(\hat{p}_i - Y_i)^2,$$

with normalized weights where population restoration is required. For the structured baselines, we additionally report Recall@Top-5%, Precision@Top-5%, and Lift@Top-5% to represent a screening setting in which only a fixed fraction of firms can receive additional review. We distinguish conditional predictive performance from end to end operational reliability. Predictive metrics are computed only when the required prediction is evaluable under the frozen parser. Operational metrics instead retain all planned requests as the denominator. This separation prevents malformed

Table 9: Frozen conditional automatic detector-positive rates and original bootstrap intervals. No intervals were re-estimated. CF4/CF6 are not validated semantic rates.

Deployment	CF	Positive/N	Rate	95% interval
DeepSeek	CF1	0/230	0.0%	[0.0, 0.0]
DeepSeek	CF2	0/233	0.0%	[0.0, 0.0]
DeepSeek	CF3	41/226	18.1%	[13.3, 23.0]
DeepSeek	CF4	234/234	100.0%	[100.0, 100.0]
DeepSeek	CF5	34/219	15.5%	[11.0, 20.1]
DeepSeek	CF6	217/230	94.3%	[91.3, 97.0]
Qwen3.8	CF1	5/159	3.1%	[0.6, 6.3]
Qwen3.8	CF2	0/160	0.0%	[0.0, 0.0]
Qwen3.8	CF3	29/113	25.7%	[18.6, 33.6]
Qwen3.8	CF4	147/164	89.6%	[85.4, 93.9]
Qwen3.8	CF5	6/159	3.8%	[1.3, 6.9]
Qwen3.8	CF6	136/157	86.6%	[81.5, 91.7]

Table 10: Frozen end-to-end automatic detector-positive rates and original bootstrap intervals. No intervals were re-estimated. CF4/CF6 are not validated semantic rates.

Deployment	CF	Positive/N	Rate	95% interval
DeepSeek	CF1	10/240	4.2%	[1.7, 7.1]
DeepSeek	CF2	7/240	2.9%	[0.8, 5.0]
DeepSeek	CF3	55/240	22.9%	[17.5, 27.9]
DeepSeek	CF4	240/240	100.0%	[100.0, 100.0]
DeepSeek	CF5	55/240	22.9%	[17.9, 28.3]
DeepSeek	CF6	227/240	94.6%	[91.7, 97.1]
Qwen3.8	CF1	86/240	35.8%	[30.0, 41.7]
Qwen3.8	CF2	80/240	33.3%	[27.5, 39.2]
Qwen3.8	CF3	156/240	65.0%	[58.8, 71.2]
Qwen3.8	CF4	223/240	92.9%	[89.6, 96.2]
Qwen3.8	CF5	87/240	36.2%	[30.4, 41.7]
Qwen3.8	CF6	219/240	91.2%	[87.5, 95.0]

responses from being silently removed from system level evaluation while avoiding the opposite error of treating every formatting failure as an ordinary numeric prediction error. For request i, let

$$J_i = \mathbf{1}\{\text{valid JSON}\}, \qquad S_i = \mathbf{1}\{\text{schema valid}\},$$

and, where an action is required,

$$A_i = \mathbf{1}\{\text{valid action}\}.$$

Operational reliability is summarized by

$$\text{JSONValidity} = \frac{1}{N}\sum_i J_i, \qquad \text{SchemaValidity} = \frac{1}{N}\sum_i S_i,$$

together with the corresponding action validity rate. For Condition C, we additionally report substantive coverage,

$$\text{Coverage} = \frac{\#\text{answered decisions}}{\#\text{eligible requests}}.$$

Coverage is interpreted jointly with predictive quality because a lower answer rate does not by itself imply more reliable selective prediction. We therefore examine how often the model abstains or escalates, whether those cases are systematically more difficult or information poor, and whether the retained answered subset exhibits better predictive performance. Where uncertainty estimates are required, paired comparisons preserve the repeated episode structure through frozen paired bootstrap procedures. Counterfactual analyses use their prespecified bootstrap configuration, while proportions such as evaluator PPV and NPV use exact or otherwise appropriate binomial confidence intervals. We do not introduce post hoc significance tests for endpoints that were not specified in the corresponding analysis protocol.

Table 11: Automatic labels versus frozen reviewer consensus. CF4 excludes one indeterminate case (47 clear); CF6 has 48 clear cases. Both detectors are NOT VALIDATED. These stratified counts are not population failure estimates.

Detector	TP	FP	FN	TN	PPV	NPV
CF4	0	34	3	10	0.00%	76.92%
CF6	0	25	0	23	0.00%	100.00%

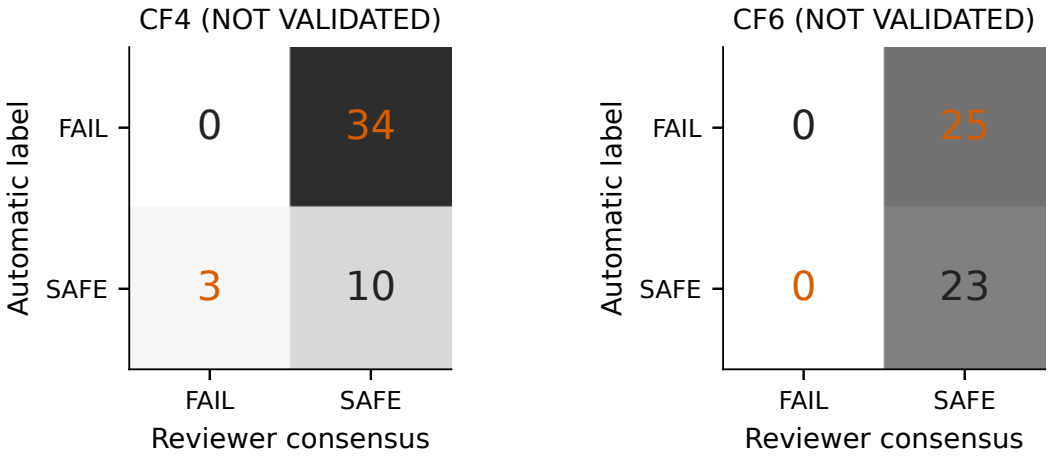

Figure 5: Reviewer-clear construct alignment in the supplementary stratified sample: CF4 has 47 clear cases and one indeterminate case excluded; CF6 has 48 clear cases. Neither automatic detector is validated. The absence of a CF6 consensus failure does not establish universal safety.

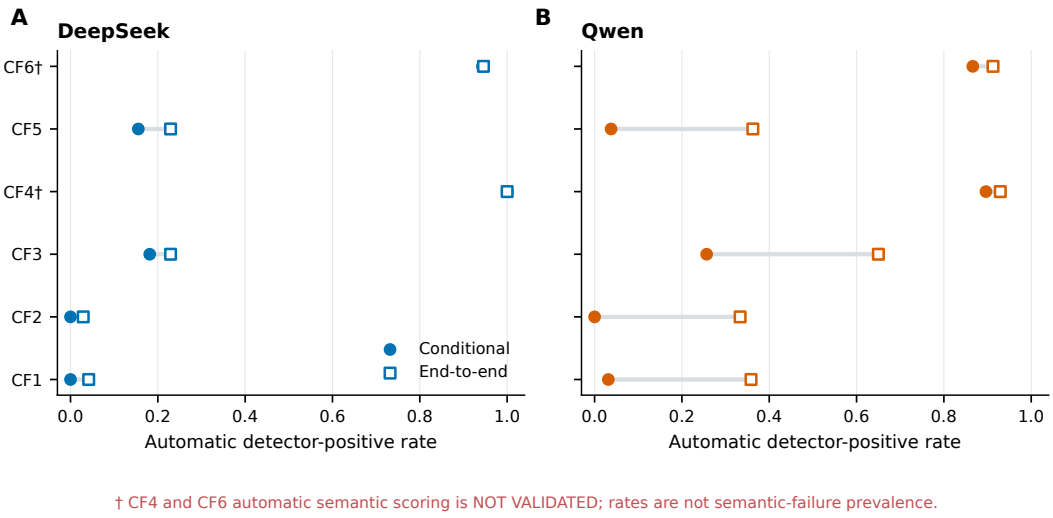

Figure 6: Frozen conditional and fail-closed end-to-end automatic detector-positive rates. These are not validated semantic-failure prevalence.

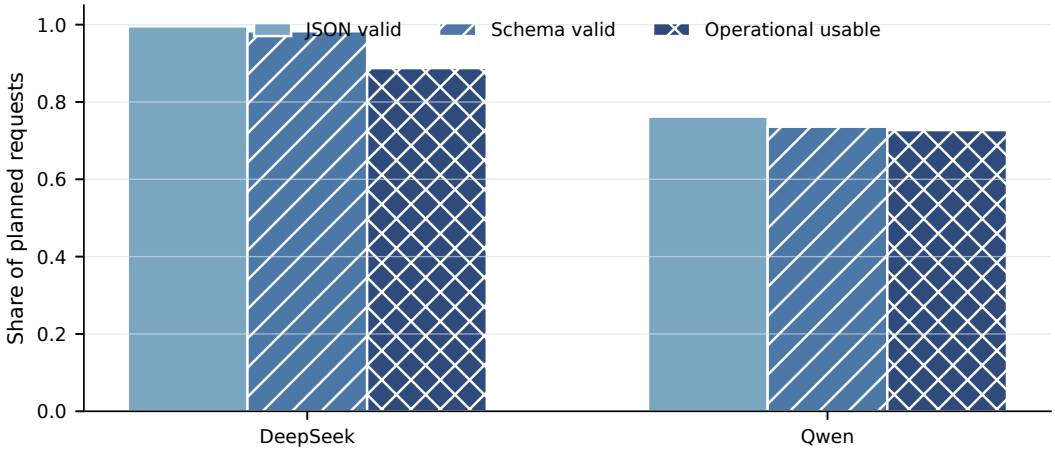

Figure 7: All-request operational funnel. Operational usability is cumulative and must not be confused with action-field parsing alone.