FINAGENT-RELIANCE: BENCHMARKING HUMAN RELIANCE AND ERROR INHERITANCE IN FINANCIAL LLM DECISION SUPPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing financial LLM benchmarks typically evaluate whether a model produces the correct answer, but stop before asking whether model failures propagate into human decisions. We introduce **FinAgent-Reliance**, a source-grounded benchmark and experimental framework for evaluating reliability at the human-AI system level. The framework links authoritative financial evidence to human-validated answers, natural and controlled agent failures, blinded correctness assessment, and separate coding of AI-specific error inheritance. Its frozen Stress-Test v2 contains 47 human-validated records: 12 correct parent responses and 35 isolated controlled failures spanning ten failure categories, with leakage-resistant splits across parent families, companies, and source documents. Stress-Test v2 is difficult for both evaluated open-weight 7B instruction models. Qwen2.5-7B achieves balanced accuracy 0.576 (95% parent-family cluster-bootstrap CI: 0.486–0.681), while Mistral-7B reaches 0.515 on parseable outputs and 0.475 when unparseable outputs are counted as failures. At the failure-typing level, neither model correctly identifies the target category for any of the 20 controlled failures spanning seven semantically subtle categories. Exploratory human studies show that the reliance measurement and evidence-inspection protocols can be implemented in practice; these studies are not used to estimate intervention effects. FinAgent-Reliance evaluates an additional question beyond model correctness: whether a known model error is detected, resisted, corrected, or inherited after it is presented to a human decision-maker.

1 INTRODUCTION

Large language models (LLMs) are increasingly used to retrieve, analyse, and reason over financial disclosures. Existing financial LLM benchmarks evaluate tasks such as numerical reasoning, evidence-grounded question answering, agentic retrieval, and cross-entity analysis (Chen et al., 2021; Islam et al., 2023; Choi et al., 2025; Jiang et al., 2026). These evaluations mainly ask whether the model retrieved the right evidence, reasoned correctly, or produced the right answer. In a decision-support setting, this leaves an important part of the problem unmeasured: an incorrect model answer may still influence the person who receives it.

This paper studies that downstream stage. We introduce **FinAgent-Reliance**, a source-grounded benchmark and experimental framework for measuring whether errors produced by a financial LLM agent are detected, resisted, corrected, or inherited after they are presented to a human decision-maker. The distinction between answer correctness and error inheritance is important. A participant may move from a correct answer to an incorrect one without adopting the AI's particular error, or may reach the correct answer without using the AI's reasoning. We therefore score answer correctness separately from textual evidence of AI-specific error inheritance, AI-supported correction, resistance, and recovery.

FinAgent-Reliance contains two complementary resources. The **Financial-Agent Failure Library** contains 75 naturally generated financial-agent responses and 14 human-validated controlled failures. **Stress-Test v2** is constructed separately from 12 human-validated parent responses and 35 controlled single-error variants tied to authoritative financial evidence. The resulting 47-record stress

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.

test spans ten failure categories. Parent families, companies, and source documents are grouped across development, public-test, and held-out partitions so that related variants of the same financial evidence do not cross evaluation splits.

We use Stress-Test v2 to examine whether open-weight LLM evaluators can detect and classify these failures. Qwen2.5-7B-Instruct achieves balanced accuracy of 0.576 (95% parent-family cluster-bootstrap CI: 0.486–0.681; MCC 0.134). Mistral-7B-Instruct-v0.3 achieves balanced accuracy of 0.515 on parseable outputs; when its three unparseable outputs are counted as unsuccessful predictions, balanced accuracy is 0.475 and MCC is -0.053. Failure typing is more difficult. Across seven semantically subtle categories containing 20 controlled failures, neither model correctly identifies the target failure category. Because Stress-Test v2 contains only 12 independent parent families and the evaluation covers two 7B models, we treat these results as evidence about the tested configurations rather than as a general estimate of LLM capability.

The benchmark also includes a human-evaluation protocol based on judgements collected before and after AI assistance. The protocol records answer correctness, AI-specific error inheritance, correction, resistance, confidence, and evidence-inspection behaviour as separate outcomes. Two exploratory pilot studies were used to test whether these measurements could be collected reliably in the study workflow. In Pilot v2, evidence was viewed in all structured-verification trials, where inspection was required by the interface, compared with 59.4% of voluntary-evidence trials. We therefore use the human studies to assess the feasibility of the reliance measurement protocol rather than to make confirmatory claims about verification interfaces.

This paper makes four contributions:

- **Evaluation of downstream error propagation.** We extend financial-agent evaluation beyond model correctness by measuring whether a known model error is inherited, resisted, or corrected after the response reaches a human decision-maker.
- **A source-grounded financial failure benchmark.** FinAgent-Reliance links each evaluated response to authoritative evidence and includes independently reviewed controlled single-error failures across ten financial failure categories.
- **Reliance measurement separated from answer accuracy.** We score correctness independently from AI-specific error inheritance and AI-supported correction, avoiding the assumption that every before–after change in answer quality represents reliance on the AI.
- **Stress-test evidence from two open-weight LLM evaluators.** Stress-Test v2 identifies failure types that remain difficult for the two evaluated 7B models, including seven categories for which both models have zero exact-category recall under the tested evaluator configurations.

2 Related Work

Financial LLM benchmarks primarily evaluate model-level reasoning and evidence use. FinQA focuses on numerical reasoning over financial reports (Chen et al., 2021), FinanceBench evaluates evidence-grounded financial question answering (Islam et al., 2023), FinBen provides broad evaluation across financial NLP and decision tasks (Xie et al., 2024), FinAgentBench studies multi-step agentic retrieval (Choi et al., 2025), and Fin-RATE extends evaluation to longitudinal and cross-entity reasoning over SEC filings (Jiang et al., 2026). FinAgent-Reliance addresses a different stage: whether a known, source-grounded model error is inherited, resisted, or corrected after reaching a human decision-maker. A detailed benchmark comparison is provided in Appendix F.

Human-AI decision-making studies show that model accuracy alone does not determine team performance (Bansal et al., 2021a), and that users may follow incorrect AI advice even when explanations are available (Buçinca et al., 2021; Bansal et al., 2021b). Appropriate-reliance research distinguishes between accepting useful advice and rejecting incorrect advice (Schemmer et al., 2023). More recent interaction-centred evaluation similarly argues that LLM reliability should account for how users rely on model responses, not only properties of the model output (Zhou et al., 2025). Related decision-support studies also show that explanation design can affect over-reliance when AI advice is incorrect (Kayser et al., 2024). Our setting begins after AI advice has been delivered and traces whether its specific error propagates into the final human judgement.

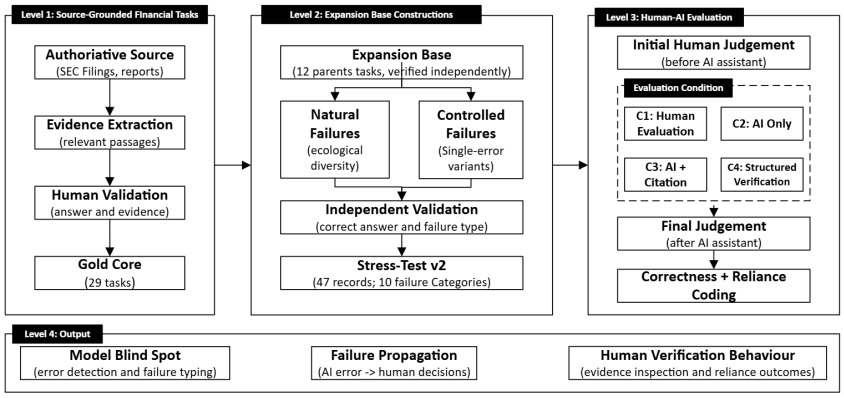


Figure 1: **FinAgent-Reliance architecture.** Layer 1 constructs source-grounded, human-validated financial tasks and the Gold Core. Layer 2 independently builds the 12-task Expansion Base, derives and validates controlled failure variants, and forms the frozen Stress-Test v2 (47 records; 10 failure categories). Layer 3 evaluates human judgements under four assistance conditions and separately scores correctness and AI-specific reliance. Layer 4 summarises model blind spots, failure propagation, and human verification behaviour.

Selective prediction and learning-to-defer study when a model should abstain or transfer a decision to another decision-maker (Geifman & El-Yaniv, 2017; Mozannar & Sontag, 2020). Our setting begins after the AI response has already been delivered and asks whether its error propagates through the subsequent human decision. The controlled-failure methodology is related to targeted behavioural evaluation, including CheckList (Ribeiro et al., 2020) and contrast-set evaluation (Gardner et al., 2020), which use controlled test cases or minimal modifications to expose failures that may be hidden by aggregate accuracy. FinAgent-Reliance applies this diagnostic principle to source-grounded financial errors and links those errors to downstream human reliance.

Recent work has shown that LLM-based evaluators can be sensitive to prompt wording, input artefacts, and evaluation instructions (Chen & Goldfarb-Tarrant, 2025; Murugadoss et al., 2025; Liu et al., 2025). Our model evaluation addresses a narrower setting: whether open-weight evaluators can identify source-grounded financial errors whose target failure type is known in advance. The prompt-robustness analysis tests whether the observed error-detection and failure-typing patterns persist under a second independently worded evaluator instruction.

3 FinAgent-Reliance Framework

FinAgent-Reliance evaluates what happens after a financial-agent response enters a human decision process. The framework starts from a financial task linked to authoritative evidence and a human-validated Gold answer, presents a fixed AI response under a specified assistance condition, and records the participant's judgement before and after that exposure. The aim is to separate two questions that are often treated as equivalent: whether the final answer is correct, and whether information or an error from the AI response was actually carried into that answer. Figure 1 summarises the FinAgent-Reliance pipeline.

3.1 Problem Formulation

For financial task j, let

$$\mathcal{T}_j = (q_j, y_j, E_j, a_j, f_j),$$

where q_j is the financial question, y_j is the human-validated Gold answer, E_j is the authoritative supporting evidence, a_j is the AI response presented to the participant, and f_j is the validated failure type when the response contains a controlled or identified error.

Participant i first provides an initial judgement h_{ij}^0 before the assigned intervention and a final judgement h_{ij}^1 afterwards. Let c_i denote the participant's experimental condition and $\mathcal{I}(c_i; a_j, E_j)$ the corresponding intervention. The interaction is

$$(q_j, E_j) \to h_{ij}^0 \to \mathcal{I}(c_i; a_j, E_j) \to h_{ij}^1.$$

For C1, $\mathcal{I}$ consists of human-only reconsideration and no AI response is shown. Conditions C2–C4 expose the participant to a_j under the assistance settings defined in Section 5.1.

Initial and final responses are evaluated using the task-specific scoring function

$$S_j(h) \in \{0, 1, 2\},$$

corresponding to incorrect, partially correct, and correct responses, respectively. The scoring rubric is fixed for each task before participant responses are evaluated.

A change in $S_j(h)$ describes a change in answer quality, but it does not by itself establish reliance on the AI response. For example, a participant may move to an incorrect answer without reproducing the AI's specific error, or may reach the correct answer without using the reasoning contained in the AI response. FinAgent-Reliance therefore treats answer correctness and AI-specific reliance as separate measurements.

3.2 Reliance Outcomes

FinAgent-Reliance records five reliance-related outcomes over their applicable opportunities: beneficial adoption, harmful adoption, resistance, under-reliance, and failure recoverability. These outcomes are derived from the relationship among the participant's initial response, the presented AI response, the final response, the Gold answer, and the supporting evidence, rather than from the correctness transition alone.

Beneficial adoption concerns cases in which useful information from the AI response contributes to an improved final judgement.

Harmful adoption captures the consequential form of over-reliance in which the participant carries the AI-specific material error into the final response. A reduction in answer score alone is not sufficient: the final response must contain the error attributable to the presented AI response.

Resistance concerns opportunities in which the participant does not inherit a material error contained in the AI response.

Under-reliance is reported as an exploratory pilot outcome when useful AI advice is available and the participant's initial answer is not already fully correct. The implemented rule requires all four of the following: no AI incorporation, no AI-supported correction adoption, no improvement in correctness score, and a final response that remains below the fully correct score. This operational definition refines an earlier conceptual specification and is therefore not treated as a fully prespecified estimand. Further details are provided in Appendix A.3.

Failure recoverability records whether a participant exposed to an incorrect AI response is able to arrive at a correct final judgement rather than carrying the model failure forward.

The operational coding procedure is described in Section 5, with the detailed coding rules and pilot-specific implementation reported in Appendix A.3.

4 Benchmark Construction

4.1 Source Tasks and Human-Validated Gold

FinAgent-Reliance begins with financial question-answering tasks that can be resolved against authoritative source evidence. Candidate answers are checked for entity identity, reporting period, metric definition, units and scale, numerical consistency, and evidential support. Tasks that cannot be grounded unambiguously are excluded or revised before use.

The resulting **Gold Core v1** contains 29 frozen, human-validated tasks. Each task links a financial question to a Gold answer, supporting evidence, source provenance, and a task-specific scoring rubric. Scoring criteria are fixed before model or participant responses are evaluated, so correctness definitions are not adjusted after observing downstream outputs.

4.2 Financial-Agent Failure Library

Studying error propagation requires incorrect as well as correct AI responses whose errors can be identified. We therefore maintain a **Financial-Agent Failure Library** containing both naturally occurring model responses and controlled failures. The frozen library contains 89 responses: 75 natural responses generated by the fixed financial-agent pipeline and 14 human-validated controlled failures.

The two sources serve different purposes. Natural responses capture errors produced by the actual pipeline, whereas controlled failures isolate a known material error while preserving unrelated content and source provenance. We use natural failures to characterise observed model behaviour and controlled failures when attribution to a specific error mechanism is required. Here, "financial agent" refers to the fixed LLM-based decision-support workflow that produces the presented responses; autonomous tool-action trajectories are not evaluated.

Failures are organised into ten categories: arithmetic/numerical, unit/percentage-point, temporal, entity/classification, metric-definition, comparative reasoning, unsupported inference, citation/evidence weakness, incomplete response, and inappropriate certainty.

4.3 Controlled Failure Construction and Validation

Natural model errors do not provide uniform coverage of the failure taxonomy. We therefore construct controlled variants from independently validated parent answers using a single-error rule:

$$\text{correct parent} \rightarrow \text{one material modification} \rightarrow \text{controlled failure}.$$

A variant is retained only when the intended error is financially meaningful, detectable from the authoritative evidence, and does not introduce a second material error. Depending on the task, the modification may affect a numerical result, unit, reporting period, entity, metric definition, comparison, evidential support, completeness, or degree of certainty.

To obtain additional parent tasks without modifying Gold Core v1, we created a separate 12-task **Human-Validated Expansion Base**. Controlled variants derived from these parents were independently reviewed for financial materiality, failure-category validity, single-error isolation, preservation of unrelated content, and consistency with the source evidence. Of 36 candidate variants, 35 were accepted and one was rejected for introducing a second material error. The two reviewers agreed on all 36 decisions (raw agreement 100%, Cohen's $\kappa = 1.0$). The full review and adjudication procedure is reported in Appendix A.1.

4.4 Stress-Test v2

Stress-Test v2 combines the 12 correct Expansion Base parent responses with the 35 validated controlled failures, yielding 47 records across all ten failure categories. It is a diagnostic challenge set: its purpose is to test whether specified, source-grounded errors can be detected and classified, not to estimate their natural prevalence.

The benchmark is divided into development, public-test, and held-out sets of 27, 8, and 12 records, respectively. A parent task and all of its controlled variants are assigned to the same split. Parent families that share a company or source document are also grouped before assignment, giving zero parent-family, company, or source-document overlap across splits.

The 47 records derive from only 12 parent families, so variants from the same parent are not treated as independent for uncertainty estimation. Point estimates use all records, while confidence intervals resample complete parent families as described in Section 5.

5 Evaluation Protocol

5.1 Human Study Design

The human pilots use the frozen 24-item **Experimental Stimuli v1**. The automated evaluator experiments use the separately constructed 47-record **Stress-Test v2**. These resources serve different purposes: the former supports the human reliance protocol, while the latter provides controlled coverage for model-level failure detection and classification.

Participants complete each task in two stages. They first provide an initial answer before receiving assistance and then submit a final answer after the assigned intervention. Initial responses are locked once assistance begins, and no condition provides correctness feedback.

We use four between-participant conditions: **C1**, human-only reconsideration; **C2**, a fixed AI response without citations; **C3**, the same AI response with optional access to its cited evidence; and **C4**, the same response and evidence with required evidence inspection, checklist completion, and a neutral challenge prompt before final submission. The AI response is identical across C2–C4, and C3 and C4 use the same citation set. Participants provide a final answer and confidence score; AI-assisted conditions additionally record an advice-action judgement, while citation conditions record evidence interaction.

5.2 Participants

Pilot v1 included 24 completed participants. Pilot v2 included 16 completed participants, with eight assigned to C3 and eight to C4; each participant completed eight tasks. Pilot v2 was introduced after Pilot v1 recorded no evidence opening in 96 citation-eligible C3/C4 trials and was used to test the revised evidence-access and structured-verification workflow. Both pilots are treated as protocol-validation studies rather than powered tests of intervention efficacy. Additional participant and ethics details are reported in Appendix A.2.

5.3 Correctness and Reliance Coding

Initial and final answers are scored against the frozen task-specific Gold rubric as 0 (incorrect), 1 (partially correct), or 2 (correct). Correctness scorers are blinded to experimental condition, response stage, AI output, confidence, and participant identity.

Reliance is coded separately from the textual relationship among the participant's initial response, the presented AI response, the final response, the Gold answer, and the supporting evidence. The coding records AI incorporation, AI-specific error inheritance, AI-supported correction, and independent correction. This separation prevents a change in correctness from being treated automatically as evidence of reliance on the AI.

For C3 and C4, we additionally record citation opening, successful evidence viewing, viewing duration, and, in C4, checklist and challenge completion. These measures describe verification behaviour and are not used as proxies for correctness. Pilot reliance labels were produced by a single blinded coder and are therefore treated as exploratory. The detailed coding rules, including the operational definition of under-reliance, are provided in Appendix A.3.

5.4 Model Evaluation Protocol

Stress-Test v2 is evaluated with three non-generative baselines and two open-weight instruction models: Qwen2.5-7B-Instruct and Mistral-7B-Instruct-v0.3. Each system receives only the participant-facing information for a record: the financial question, presented answer, and authoritative evidence. Gold answers, expected labels, failure metadata, parent-family identifiers, and benchmark origin are withheld.

The non-generative baselines are: *Always trust*, which predicts that every presented answer is correct; *Majority class*, which predicts an error for every record; and an embedding-similarity baseline that flags an error when answer–evidence similarity falls below a development-tuned threshold. The embedding model and threshold-selection procedure are reported in Appendix C.4.

The generative evaluators predict whether the presented answer contains a material error and, when applicable, assign one of the ten frozen failure categories. Model-specific prompt delivery differs only where required by the respective chat template; the evaluator instruction and output schema are held fixed. Full prompts, decoding settings, parsing rules, and cross-model delivery details are reported in Appendix C.

The model evaluation is intended to test whether Stress-Test v2 exposes diagnostic differences under the reported configurations. It is not used to rank current language models or estimate a general performance ceiling.

5.5 Evaluation Metrics and Statistical Analysis

Stress-Test v2 contains 35 controlled failures and 12 correct responses, so raw accuracy and detection F1 alone can be misleading. We therefore report balanced accuracy, Matthews correlation coefficient (MCC), detection F1, sensitivity, and specificity.

Because the 47 records derive from only 12 parent families, uncertainty is estimated with a parent-family cluster bootstrap. Each of 10,000 bootstrap replicates samples complete parent families with replacement and includes all records belonging to each sampled family. Point estimates are computed on the full benchmark, and 95% confidence intervals are obtained from the 2.5th and 97.5th percentiles of the bootstrap distribution. This preserves dependence among controlled variants derived from the same source task. Implementation details are given in Appendix C.3.

5.6 Prompt-Robustness Protocol

We conduct a post-hoc sensitivity analysis to test whether evaluator behaviour depends on the wording of the instruction. A second, independently worded prompt is run once on the same frozen 47-record benchmark. Model weights, quantization, context length, decoding settings, participant-facing inputs, output schema, parser, and random seed are unchanged. The alternative prompt was not revised after its outputs were inspected.

For comparison across prompts, both evaluations use the same 47-record scope, with unparseable outputs counted as unsuccessful predictions. We treat this analysis as a robustness check rather than an independent confirmatory experiment. Both evaluator prompts are reproduced in Appendix D.

6 Results

6.1 Benchmark Validation

Stress-Test v2 contains 47 human-validated records: 12 correct parent responses and 35 controlled single-error failures spanning all ten failure categories. Of the 36 controlled-failure candidates reviewed, 35 were retained and one was rejected because it introduced a second material error. The two reviewers agreed on all 36 decisions (raw agreement 100%, Cohen's $\kappa = 1.0$).

The final benchmark contains 27 development, 8 public-test, and 12 held-out records. Parent tasks and all variants derived from them are assigned to the same split. Grouping additionally by company and source document produces zero parent-family, company, or source-document overlap across the three partitions.

6.2 Model Performance on Stress-Test v2

Table 1 reports the three non-generative baselines and the two open-weight 7B evaluators.

Qwen2.5-7B achieves balanced accuracy 0.576 (95% CI: 0.486–0.681), MCC 0.134 (95% CI: -0.024–0.330), and detection F1 0.607 (95% CI: 0.372–0.765). Its sensitivity is 0.486 (95% CI: 0.242–0.722). The confidence intervals remain wide, and neither the balanced-accuracy nor MCC interval supports a strong above-chance claim.

On the full 47-record evaluation, with unparseable outputs counted as unsuccessful predictions, Mistral-7B obtains balanced accuracy 0.475 (95% CI: 0.361–0.580), MCC -0.053 (95% CI: -0.307–0.182), and detection F1 0.311. Specificity is 0.750. As a secondary diagnostic, Mistral

Table 1: Stress-Test v2 model results. Balanced-accuracy intervals are 95% parent-family cluster-bootstrap confidence intervals based on 10,000 resamples. MCC is undefined for the two constant-prediction baselines and is reported as "–".

Model	Bal. Acc.	MCC	Det. F1	Spec.
Always trust	.500	–	.000	1.000
Majority class	.500	–	.854	.000
Embedding similarity	.485	-.044	.805	.083
Qwen2.5-7B	.576 [.486,.681]	.134	.607	.667
Mistral-7B	.475 [.361,.580]	-.053	.311	.750

produces parseable outputs for 44 of the 47 records; restricted to those outputs, balanced accuracy is 0.515, MCC is 0.033, and detection F1 is 0.333.

6.3 Failure-Type Analysis

Failure typing is uneven across categories. Under the primary evaluator prompt, the models achieve non-zero exact-category recall only for arithmetic/numerical, metric-definition, and unit/percentage-point failures.

Across the remaining seven categories: citation/evidence weakness, comparative reasoning, entity/classification, inappropriate certainty, incomplete response, temporal, and unsupported inference—neither model correctly identifies the target failure category for any of the 20 controlled failures. Because individual categories contain only $n = 2$–9 examples, we treat these results as diagnostic rather than population-level estimates. Full category-level results are reported in Appendix C.5.

6.4 Prompt-Robustness Analysis

Changing the evaluator instruction affects binary error detection. On the matched 47-record evaluation, Qwen's balanced accuracy changes from 0.576 under the primary prompt to 0.616 under the alternative prompt. Its sensitivity decreases from 0.486 to 0.314, while specificity increases from 0.667 to 0.917.

For Mistral, balanced accuracy changes from 0.475 to 0.530, sensitivity from 0.200 to 0.143, and specificity from 0.750 to 0.917.

The category-level result is unchanged: under both prompts, neither model correctly classifies any of the controlled failures in the same seven difficult categories. Binary detection is therefore sensitive to evaluator wording, whereas the observed failure-typing pattern is stable across the two tested prompts. The full matched comparison is reported in Appendix D.2.

6.5 Human Reliance and Verification Behaviour

Pilot v1 included 24 participants and 192 completed trials. None of the 96 citation-eligible C3/C4 trials opened the supplied evidence. A technical audit classified these events as `TRUE_ZERO_INTERACTION`, rather than a telemetry failure, which motivated the revised evidence-access workflow used in Pilot v2.

Pilot v2 included eight participants in C3 and eight in C4, with 64 completed trials per condition. Evidence was viewed in 38/64 C3 trials (59.4%) under voluntary access and in all 64 C4 trials, where evidence inspection was required by the interface. Checklist and challenge completion in C4 were also 100%. Median task time was 82.4 s in C3 and 131.9 s in C4.

Final accuracy was 76.6% in both conditions. The exploratory reliance measures were also identical across C3 and C4: beneficial adoption 75.0%, under-reliance 0.0%, harmful adoption 6.2%, resistance 62.5%, and failure recoverability 62.5%. Thus, mandatory verification in C4 produced complete recorded evidence viewing but was not accompanied by a difference in final accuracy or the coded reliance outcomes in this pilot. Given the small sample and exploratory design, these results are not treated as evidence of intervention efficacy.

7 DISCUSSION

On the two evaluated 7B models, arithmetic and some definitional failures were easier to identify than several semantically subtle error types. Binary error detection changed when the evaluator instruction was reworded, showing that prompt wording is part of the evaluation configuration. The category-level result was more consistent: under both prompts, neither model correctly classified any of the 20 failures drawn from the same seven difficult categories. Because Stress-Test v2 contains only 12 independent parent families and the evaluation covers two models, we treat this as a result for the tested configurations rather than a general statement about LLM capability.

Stress-Test v2 is intentionally failure-heavy, so accuracy and detection F1 can give a distorted view of evaluator performance. The majority-class baseline, for example, obtains detection F1 of 0.854 by predicting the error class for every record, but its specificity is 0 and its balanced accuracy is 0.500. An evaluator that rejects every answer is therefore not useful for this task.

Pilot v2 also illustrates why evidence interaction and decision quality should be reported separately. Evidence was viewed voluntarily in 59.4% of C3 trials and was mandatory in all C4 trials. Despite complete recorded evidence viewing in C4, final accuracy and the exploratory reliance outcomes were unchanged between the two conditions. The pilot is too small to estimate an intervention effect, but it shows that observed evidence interaction should not be used as a proxy for improved judgement.

FinAgent-Reliance is intended to identify where failure propagation occurs after an AI response is produced. A model error may be detected by the evaluator, resisted by the user, corrected during verification, or inherited in the final answer. The benchmark is not intended as a model leaderboard, and the current human pilots do not establish that any verification interface prevents harmful reliance.

8 LIMITATIONS AND RESPONSIBLE USE

Stress-Test v2 contains 47 records derived from only 12 independent parent families, with per-category support of $n = 2$–9. We therefore estimate uncertainty by resampling parent families rather than individual records; however, the small number of clusters still limits precision and population-level inference. Model findings are limited to two Q4_K_M-quantised 7B open-weight models and should not be generalised to larger or proprietary systems. Controlled failures isolate specific error types, but their distribution does not represent the frequency of naturally occurring model errors, and the human pilots are protocol-validation studies rather than powered intervention tests. Upstream licensing restrictions, including CC-BY-NC-4.0 FinanceBench content, must be preserved in redistribution. The benchmark is intended for reliability research and evaluation, not financial advice. Reliance outcomes in the pilot studies were coded by a single reviewer and were not independently double-coded; the human-reliance results should therefore be interpreted as exploratory pending inter-rater validation. Participant financial expertise was not measured, further limiting generalisation of the human-study findings.

9 CONCLUSION

We presented FinAgent-Reliance, a benchmark for studying what happens when a financial-agent response reaches a human decision-maker. Each benchmark task is tied to authoritative source evidence, and the controlled cases contain a known material error that has been independently reviewed. Stress-Test v2 evaluates whether models can detect and classify these errors, while the human protocol separately measures answer correctness and AI-specific error inheritance.

The current experiments show that the two evaluated 7B models have difficulty identifying several semantically subtle failure types. The human pilots also show that increased evidence inspection does not necessarily coincide with better final answers. These results are limited to the tested models, 12 parent families, and exploratory human samples, but they provide a concrete setup for studying error propagation in financial human–AI decision support.

Ethics Statement

The human pilot studies were conducted under institutional ethics approval. Participant responses were stored using anonymous study identifiers. Participant-facing materials did not expose Gold answers, failure labels, natural-versus-controlled origin, model/provider metadata, reviewer information, or research hypotheses. Identifying details of the ethics approval are withheld during double-blind review. The human studies are reported as protocol-validation pilots rather than powered tests of intervention efficacy.

Reproducibility Statement

Sections 4–6 describe the benchmark construction, evaluation protocol, metrics, and statistical analysis. Appendices A–D provide the human-review procedure, benchmark composition, evaluator prompts, decoding and parsing configuration, cluster-bootstrap procedure, embedding-similarity baseline, category-level results, and prompt-robustness analysis. The frozen research artifacts preserve task and parent identifiers, source provenance, validation status, and evaluation split assignments.

AI Use Statement

The authors used ChatGPT and Claude for literature summarisation, methodological checking, code-development support, and language editing. AI-generated material was not accepted without author review. The authors independently checked the experimental procedures, numerical results, references, and scientific claims and take responsibility for the final manuscript.

References

Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S. Weld. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11405–11414, 2021a. doi: 10.1609/aaai.v35i13.17359.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2021b. doi: 10.1145/3411764.3445717.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021. doi: 10.1145/3449287. URL https://doi.org/10.1145/3449287.

Hongyu Chen and Seraphina Goldfarb-Tarrant. Safer or luckier? LLMs as safety evaluators are not robust to artifacts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19750–19766, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.970. URL https://aclanthology.org/2025.acl-long.970/.

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 the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3697–3711. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.300. URL https://aclanthology.org/2021.emnlp-main.300/.

Chanyeol Choi, Jihoon Kwon, Alejandro Lopez-Lira, Chaewoon Kim, Minjae Kim, Juneha Hwang, Jaeseon Ha, Hojun Choi, Suyeol Yun, Yongjin Kim, and Yongjae Lee. FinAgentBench: A benchmark dataset for agentic retrieval in financial question answering. *arXiv preprint arXiv:2508.14052*, 2025. URL https://arxiv.org/abs/2508.14052.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117/.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023. doi: 10.48550/arXiv.2311.11944. URL https://arxiv.org/abs/2311.11944.

Yidong Jiang, Junrong Chen, Eftychia Makri, Jialin Chen, Peiwen Li, Ali Maatouk, Leandros Tassiulas, Eliot Brenner, Bing Xiang, and Rex Ying. Fin-rate: A real-world financial analytics and tracking evaluation benchmark for llms on sec filings. *arXiv preprint arXiv:2602.07294*, 2026. doi: 10.48550/arXiv.2602.07294. URL https://arxiv.org/abs/2602.07294.

Maxime Kayser, Bayar Menzat, Cornelius Emde, Bogdan Bercean, Alex Novak, Abdala Espinosa, Bartlomiej W. Papiez, Susanne Gaube, Thomas Lukasiewicz, and Oana-Maria Camburu. Fool me once? contrasting textual and visual explanations in a clinical decision-support setting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18891–18919, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1051. URL https://aclanthology.org/2024.emnlp-main.1051/.

Shuliang Liu, Xinze Li, Zhenghao Liu, Yukun Yan, Cheng Yang, Zheni Zeng, Zhiyuan Liu, Maosong Sun, and Ge Yu. Judge as a judge: Improving the evaluation of retrieval-augmented generation through the judge-consistency of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5788–5807, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.301. URL https://aclanthology.org/2025.findings-acl.301/.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087. PMLR, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Bhuvanashree Murugadoss, Christian Poelitz, Ian Drosos, Vu Le, Nick McKenna, Carina Suzana Negreanu, Chris Parnin, and Advait Sarkar. Evaluating the evaluator: Measuring LLMs' adherence to task evaluation instructions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19589–19597, 2025. doi: 10.1609/aaai.v39i18.34157. URL https://doi.org/10.1609/aaai.v39i18.34157.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. Appropriate reliance on ai advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pp. 410–422. Association for Computing Machinery, 2023. doi: 10.1145/3581641.3584066.

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*, 2024. doi: 10.52202/079017-3033. URL https://mlanthology.org/neurips/2024/xie2024neurips-finben/.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky, and Maarten Sap. REL-A.I.: An interaction-centered approach to measuring human-LM reliance. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11148–11167, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.556. URL https://aclanthology.org/2025.naacl-long.556/.

A Human Review and Participant Details

A.1 Benchmark Review Procedure

Benchmark construction used separate review stages for candidate validation, adjudication, and controlled-failure isolation. Expansion candidates were evaluated independently by two reviewers using a fixed VALIDATE/REVISE/REJECT rubric; cases requiring resolution were passed to a separate adjudication stage before the 12-task Human-Validated Expansion Base was frozen.

Controlled-failure candidates were subsequently assessed in a separate review round for financial materiality, failure-category validity, single-error isolation, preservation of unrelated content, and consistency with the bound source evidence. Review decisions were recorded independently before reconciliation. For the 36 controlled-failure candidates, the two review passes agreed on all cases: 35 were retained and one was rejected, corresponding to 100% raw agreement and Cohen's $\kappa = 1.0$.

Reviewer identities are represented by role-coded identifiers in the research artifacts and are withheld from the manuscript for double-blind review.

A.2 Pilot Participants

Because these pilots were designed for protocol validation rather than population inference, detailed demographic and financial-expertise variables were not collected; this limits generalisation of the human results.

Pilot v1 included 24 completed participants. Pilot v2 included 16 completed participants, allocated evenly between C3 and C4 ($n = 8$ per condition), with each participant completing eight tasks.

Pilot v2 was designed as an operational protocol-validation study to evaluate the revised citation and structured-verification workflow rather than as a powered efficacy comparison. The study was conducted under institutional ethics approval; identifying approval details are withheld for double-blind review.

Participant responses were stored under anonymous study identifiers. Participant-facing materials did not expose Gold answers, failure labels, natural-versus-controlled origin, model/provider metadata, reviewer information, or research hypotheses. Under the same operational definition, Pilot v1 under-reliance was 1/39 (2.6%) across assisted conditions overall (C2: 0/14; C3: 1/15; C4: 0/10).

A.3 Reliance-Coding Procedure

The coding guide defined four textual judgements: AI incorporation, AI-specific error inheritance, AI-supported correction adoption, and independent correction. Pilot v1 and Pilot v2 were coded using the same schema-versioned procedure by a single reviewer. For under-reliance, the implemented operational rule combines two textual criteria (no AI incorporation and no AI-supported correction adoption) with two correctness criteria (no score improvement and a final response that remains not

fully correct). This four-condition rule is stricter than the earlier two-condition conceptual specification and is therefore treated as an operational refinement rather than a fully prespecified estimand. The same effective eligibility population is used in both pilots. The reviewer had access to the participant's initial response, fixed AI response, final response, Gold answer, supporting evidence, and approved rubric, but not experimental condition, behavioural telemetry, confidence, participant identity, or research hypotheses.

No independent second-coder pass was performed for these pilot annotations; accordingly, no raw inter-rater agreement or Cohen's κ is reported.

B Benchmark Release Composition

Table 2 summarises the frozen artifacts used in the FinAgent-Reliance pipeline. Gold Core v1 and Stress-Test v2 are maintained as separate releases: Gold Core establishes source-grounded reference tasks, whereas Stress-Test v2 is a diagnostic challenge set constructed from the separately reviewed Expansion Base.

Table 2: FinAgent-Reliance benchmark releases used in this work.

Artifact	Records	Role
Gold Core v1	29	Source-grounded Gold tasks
Failure Library v1	89	Natural + controlled responses
Experimental Stimuli v1	24	Frozen human-study stimuli
Expansion Base v1	12	Validated stress-test parents
Stress-Test v2	47	12 correct + 35 controlled failures

B.1 Natural Response Generation

The 75 natural responses in Failure Library v1 were generated using the frozen financial-agent configurations documented in the release artifacts. Responses were retained without manual error injection and subsequently human-adjudicated for correctness and failure type. Model identities, generation settings, response counts, and provenance are provided with the released benchmark metadata.

C Stress-Test v2 Evaluation Details

C.1 Primary Evaluator Prompt

The primary Stress-Test v2 evaluation used the following system instruction verbatim:

> You are a careful financial-document auditor. You will be shown a question, an AI-generated presented answer, and authoritative evidence excerpts. You do NOT know the correct answer in advance. Decide, using ONLY the supplied evidence, whether the presented answer is fully supported by that evidence or whether it contains a material error.
>
> Respond with EXACTLY one JSON object and nothing else, using this schema:
>
> `{"evidence_supported": true|false, "error_present": true|false, "failure_category": "<one of the categories below, or null if error_present is false>", "abstain": true|false}`
>
> Set abstain to true only if the evidence is too incomplete to judge at all.
>
> Valid failure categories (use exactly one of these strings, or null):
>
> `"arithmetic / numerical"`
> `"unit / percentage-point"`
> `"temporal"`
> `"entity / classification"`
> `"metric-definition"`
> `"comparative reasoning"`
> `"unsupported inference"`
> `"citation / evidence weakness"`

```
"incomplete response"
"inappropriate certainty"
```

The corresponding user message was constructed as:

```
Question: {question}

Presented answer: {presented_answer}

Authoritative evidence: {evidence_text}

Respond with the JSON object only.
```

The evidence text was taken from the participant-facing `evidence_boundary` when available, otherwise from the concatenated evidence excerpts. Evidence was truncated at 2,400 characters, with a trailing `...[truncated]` marker when necessary. No Stress-Test v2 record required evidence truncation; all evaluator-visible inputs therefore retained the complete frozen evidence boundary.

The evaluator was not provided with the reference answer, expected label, failure metadata, benchmark origin, or parent-family identifier.

Output parsing. Structured output was requested through the prompt rather than function calling or a constrained decoding grammar. The evaluation harness extracted the first JSON object from the generated response and parsed `evidence_supported`, `error_present`, `failure_category`, and `abstain`. Outputs without a parseable JSON object were treated as abstentions. Failure-category strings outside the frozen ten-category vocabulary were mapped to null.

C.2 Cross-Model Prompt Delivery

Qwen and Mistral received the same evaluator instruction, task-specific user template, failure-category vocabulary, structured-output schema, and decoding configuration. Message placement differed only because of model-specific chat-template mechanics.

For Qwen2.5-7B-Instruct, the evaluator instruction was supplied as a system message and the task-specific content as a separate user message:

```
{"role": "system", "content": SYSTEM_PROMPT}
{"role": "user", "content": USER_PROMPT}
```

For Mistral-7B-Instruct-v0.3, the `llama-cpp-python` Mistral chat formatter used by the GGUF accepts user/assistant turns but drops system-role messages. To ensure that the same evaluator instruction reached the model, the system instruction and task-specific prompt were concatenated into a single user turn:

```
{"role": "user", "content": SYSTEM_PROMPT + "\n\n" +
USER_PROMPT}
```

This transformation changes message-role placement but not evaluator content. Both models therefore received the same frozen instruction text and evaluation schema.

Decoding configuration. Both generative baselines used deterministic decoding with `temperature` $= 0.0$, `top_p` $= 1.0$, `max_tokens` $= 200$, and `seed` $= 42$. Inference used a context window of 2,048 tokens, four CPU threads, and zero GPU layers.

For every record, the evaluator received only the participant-facing task view: the financial question, presented answer, and authoritative evidence. Gold answers, expected labels, failure metadata, parent-family identifiers, and benchmark-origin fields were withheld.

Because 35 of the 47 Stress-Test v2 records contain controlled failures, raw accuracy and detection F1 can be misleading. For example, the majority-class baseline achieves detection F1 0.854 while having specificity 0.000 and balanced accuracy 0.500. Class-imbalance-aware measures are therefore emphasised in the primary analysis.

C.3 Cluster-Bootstrap Uncertainty

Because multiple records derive from the same parent task, we use `parent_family_id` as the bootstrap unit. For each of 10,000 replicates, 12 parent families are sampled with replacement and all associated records are included together. Point estimates are computed on the original benchmark; 95% CIs are the 2.5th and 97.5th percentiles of the clustered bootstrap distribution.

C.4 Embedding-Similarity Baseline

The embedding-similarity baseline encodes the presented answer and authoritative evidence separately using `intfloat/e5-base-v2` (revision `f52bf8ec8c7124536f0efb74aca902b2995e5bcd`), with the E5 `query:` and `passage:` prefixes, attention-mask-weighted mean pooling, and L2 normalization. Cosine similarity is computed between the two representations, and an error is flagged when similarity is below 0.90. The threshold was selected by maximising detection F1 on the development split over candidate thresholds from 0.70 to 0.98 in 0.01 increments, then frozen before evaluation on the public-test and held-out splits. The question text is not embedded, and this baseline does not predict a failure category.

C.5 Per-Category Failure Detection

Table 3 reports exact-category recall for the two generative models under the primary evaluator prompt.

Table 3: Exact-category recall on Stress-Test v2 under the primary evaluator prompt.

Failure type	n	**Qwen**	**Mistral**
Arithmetic / numerical	9	.222	.111
Metric definition	2	.500	.000
Unit / percentage-point	4	.250	.000
Citation / evidence weakness	3	.000	.000
Comparative reasoning	3	.000	.000
Entity / classification	2	.000	.000
Inappropriate certainty	3	.000	.000
Incomplete response	4	.000	.000
Temporal	2	.000	.000
Unsupported inference	3	.000	.000

Across citation/evidence weakness, comparative reasoning, entity/classification, inappropriate certainty, incomplete response, temporal, and unsupported inference, neither model correctly identifies the target failure category. Together these seven categories contain 20 controlled failures.

Because category support is small ($n = 2$–9), these values are interpreted as diagnostic results rather than population-level estimates.

D Evaluator-Prompt Robustness

D.1 Alternative Robustness Prompt

The post-hoc robustness analysis changed only the natural-language evaluator instruction. The alternative prompt was:

> You are evaluating whether a financial answer is supported by the supplied evidence.
>
> Read the question, the proposed answer, and the evidence carefully.
>
> Decide whether the proposed answer contains a material error or unsupported claim. Consider numerical calculations, units, reporting periods, entity identity, metric definitions, comparisons, evidential support, completeness, and unjustified certainty.
>
> Do not assume that the answer is wrong merely because this is an evaluation task. A fully supported answer should be accepted as correct.

Return only the required structured output using the same schema as the primary evaluation.

Model weights, quantization, context length, thread count, GPU configuration, temperature, seed, participant-facing input representation, JSON schema, parser, and benchmark records were identical to the primary evaluation. The alternative prompt was executed once and was not modified after inspecting its results.

D.2 Prompt-Robustness Results

Both prompts are evaluated on the same 47-record coverage-adjusted population; unparseable outputs are scored as failures rather than excluded.

Table 4: Matched-scope prompt-robustness comparison on all 47 Stress-Test v2 records.

Model	Prompt	Sens.	Spec.	Bal. Acc.	MCC	Det. F1
Qwen2.5-7B	Primary	.486	.667	.576	.134	.607
Qwen2.5-7B	Alternative	.314	.917	.616	.231	.468
Mistral-7B	Primary	.200	.750	.475	-.053	.311
Mistral-7B	Alternative	.143	.917	.530	.078	.244

The alternative prompt shifts both models toward higher specificity and lower sensitivity. Nevertheless, both models retain zero exact-category recall under both prompts across the same seven difficult categories, comprising 20 controlled failures. We therefore treat binary detection as prompt-sensitive and the persistent failure-typing blind spot as the more stable diagnostic finding. This analysis is post-hoc and is not treated as an independent confirmatory experiment.

E Detailed Pilot v2 Outcomes

Pilot v2 included 16 completed participants, with eight assigned to C3 (voluntary evidence) and eight assigned to C4 (structured verification). Each participant completed eight tasks, yielding 64 completed trials per condition.

Table 5: Exploratory Pilot-v2 outcomes. Values are descriptive and are not confirmatory estimates of intervention efficacy.

Metric	C3	C4
Final accuracy	76.6%	76.6%
Beneficial adoption	75.0%	75.0%
Under-reliance	0.0%	0.0%
Harmful adoption	6.2%	6.2%
Resistance	62.5%	62.5%
Failure recoverability	62.5%	62.5%
Evidence viewed	59.4%	100.0%
Median task time	82.4 s	131.9 s

Evidence was viewed in 38/64 C3 trials (59.4%) under voluntary access and in all 64 C4 trials, where evidence inspection was required by the interface. Checklist and challenge completion in C4 were also complete. Median task time was 82.4 s in C3 and 131.9 s in C4, a difference of approximately 49.5 s per task.

Final accuracy and the exploratory reliance outcomes were identical across C3 and C4. These descriptive pilot results do not establish that structured verification improves decision quality or reduces harmful reliance.

F Comparison with Financial LLM Benchmarks

Table 6: Positioning of FinAgent-Reliance relative to representative financial LLM benchmarks. A checkmark denotes an explicit evaluation component; "–" indicates that the component is not a stated evaluation target of the benchmark.

Benchmark	**Source grounding**	**Financial reasoning**	**Controlled failures**	**Human reliance**	**Error inheritance**
FinQA (Chen et al., 2021)	✓	✓	–	–	–
FinanceBench (Islam et al., 2023)	✓	✓	–	–	–
FinAgentBench (Choi et al., 2025)	✓	✓	–	–	–
Fin-RATE (Jiang et al., 2026)	✓	✓	–	–	–
FinAgent-Reliance	✓	✓	✓	✓	✓