

A ROSEY RESEARCH REPORT

Benchmarking Large Language Models on Document Accuracy Verification

*A Precision–Recall–F1 Evaluation Using an LLM-Generated
Cracker Barrel Old Country Store Quarterly Earnings Summary*

Published by InSummary, Inc.

April 2026

Report v1.0 · Ground Truth v003 (2026-04-22) · Rosey v0.3.2

Abstract

We evaluate the ability of large language models (LLMs) to detect numerical errors and unsupported claims in a professional document. The test document is an LLM-generated first draft of a Q2 FY2026 quarterly earnings summary for Cracker Barrel Old Country Store, Inc. (NASDAQ: CBRL), assessed against the company's quarterly earnings release, investor presentation, and earnings call transcript. A human review of every numerical claim in the test document produced a Ground Truth set of **51 verified accuracy errors** – 21 numerical contradictions and 30 unsupported claims – drawn from a total of 330 numerical claims (an error rate of 15.5%).

We test **15 LLM configurations** spanning three providers (Anthropic, Google, OpenAI), three Model Tiers (Lite, Balanced, Max), and two Reasoning Tiers (Baseline, Extended). All LLM tests were conducted via each provider's web-based user interface. Each configuration was run 3 independent times, yielding 45 total runs. Runs taking longer than 12 minutes were dropped, leaving 12 configurations and 33 independent runs. For each included configuration we report their average Precision, Recall, and F1 against the Ground Truth, with finding-to-Ground-Truth matching adjudicated by a human audit.

We find that no LLM configuration tested achieved an F1 above 0.41. The median LLM F1 is 0.26. Five of the twelve included configurations score below F1 0.20 – a scenario in which most Ground Truth errors are missed and a majority of raised findings are false positives. Increasing reasoning allocation does not produce monotonic improvement; Claude Opus Baseline (F1 0.36) outperforms Claude Opus Extended (F1 0.31) on the same document. The highest-precision LLM configuration, Claude Sonnet Extended, achieves 87% precision – matching Rosey – but catches only 28% of Ground Truth errors. The highest-recall LLM configuration, GPT Thinking Baseline, achieves 33% recall – half of Rosey – but only 54% of findings match Ground Truth errors.

We additionally test Rosey, a purpose-built document quality assurance system developed by InSummary, Inc., on the same task, under the same protocol, and scored by the same adjudication process. Rosey achieves F1 0.76 (Precision 87%, Recall 67%) – approximately 2.4 times the recall of the highest-precision LLM at equivalent precision, and nearly double the F1 of the best LLM tested.

In coordination with this report, we have released everything required to reproduce the study: the test document, source materials, evaluation prompt, scoring rubric, full configuration matrix, the complete Ground Truth, and a public interactive results explorer. We intend to regenerate this benchmark periodically on new documents as LLMs (and Rosey) continue to progress.

1. Introduction

The professional document is one of the most rapidly adopted uses of generative AI. Investment committee memos, earnings summaries, consulting deliverables, scientific manuscripts, regulatory filings, technical reports – written artifacts that once took several days to produce are now drafted with large language models in minutes. The outputs are fluent, well-organized, and superficially faithful to the source materials. They are also, with uncomfortable frequency, wrong.

Generative models are trained to produce plausible continuations of text, not verified ones. When an LLM writes “Q4 revenue rose 14% year-over-year,” it is generating a probable completion given the context it has been shown. Sometimes the 14% figure is in the source. Sometimes the source says 4% and the model has transposed digits or inverted directionality. Sometimes the figure is an invention with no basis in the source at all. The errors tend to be subtle and unevenly distributed: not every number is wrong, but enough of them are wrong to matter for any downstream decision that rests on the document.

This verification gap is well-known to anyone who has systematically checked LLM output against source material. It is also under-measured. Most published LLM benchmarks target capabilities that are adjacent but not identical to document-level numerical verification: code generation, multi-step reasoning, short-form question answering, retrieval, agentic workflows. Few test whether a model, given a draft document and its source materials, can reliably find the numerical errors a human reviewer would flag.

This study addresses this gap in research.

1.1 Why earnings summaries

We use a quarterly earnings summary as the test document for five reasons.

- **Density.** A typical earnings summary contains several hundred specific numerical, directional, and categorical claims in the space of a few pages, making it a rich environment for measuring quantitative accuracy.
- **Accessibility.** The input materials into this test are publicly available which makes this test transparent and reproducible.
- **Closed source universe.** The earnings release, investor presentation, and call transcript form a small, authoritative, self-contained set of sources. Every claim in the summary can be definitively classified as supported, contradicted, or unsupported against this set.
- **High-stakes.** Earnings summaries feed investor decisions, sell-side research, investment committee discussions, and media coverage. Errors in these documents can lead to material consequences.

- **Extensibility.** The error modes present in an LLM-drafted earnings summary – transposed digits, inverted directionality, invented metrics, miscalculated ratios, and hallucinated attributions – are the same error modes that appear in LLM-drafted IC memos, diligence reports, scientific manuscripts, and technical documents. A benchmark that exposes these modes here exposes them in general.

1.2 What this study contributes

1. A replicable benchmark of LLM performance on document-level numerical verification, grounded in a single, realistic test document and a transparent scoring methodology.
2. Evaluation conducted in each provider's **consumer user interface** – the environment with which most sophisticated end users interact – rather than via raw API. We have confirmed that accessing via the UI – as compared to the APIs of these LLMs – produces superior results and have therefore focused our measurements to align with real-world use of LLMs via the UI.
3. A rigorous Ground Truth construction that distinguishes between numerical contradictions and unsupported claims, adjudicated through a human review process, with every claim in the test document individually classified against the source materials.
4. A comparative result for Rosey, a document quality assurance system designed on the thesis that reliable verification must be structurally separate from generation.

We release the methodology, test document, source materials, prompt, scoring rubric, models tested, the complete Ground Truth, and an interactive results explorer so that any interested party can reproduce, challenge, or extend our findings. This is an open benchmark. We recognize that full release creates the possibility of future GenAI models incorporating the Ground Truth into training data; our response to this concern is operational rather than defensive – we are prepared to repeat this benchmark on new test documents periodically.

2. Methodology

2.1 Test document

The test document is an LLM-generated first-draft Q2 FY2026 quarterly earnings summary for Cracker Barrel Old Country Store, Inc. (NASDAQ: CBRL). The draft was produced with a single prompt to a leading frontier model, with the three source documents ([Section 2.2](#)) and a formatting example supplied as context. No human editing was applied to the draft prior to evaluation; the document evaluated is the model's direct output.

Cracker Barrel was selected for three reasons: it is a mid-cap company with an active investor audience and publicly available filings; the quarter in question contained enough narratively substantive material (operational changes, comparable sales dynamics, capital allocation commentary) to support a summary of non-trivial length and numerical density; and the company is outside the cluster of hyperscale technology firms that disproportionately populate LLM training data, reducing the risk that models are relying on memorized financial data rather than grounded retrieval from the supplied sources.

2.2 Source materials

Three primary source documents, all publicly available:

- The quarterly earnings press release issued by Cracker Barrel.
- The investor presentation accompanying the earnings release (and also presented during the earnings call).
- The earnings call transcript, covering prepared remarks and analyst Q&A.

Every numerical claim in the test document is evaluated exclusively against this set. Claims that cannot be resolved with the provided sources – and cannot be derived arithmetically from the data within them – are marked **unsupported**, even where they are plausibly correct in the broader world. This study measures document-grounded numerical verification, not factual truth in general; a model that relies on memorized external knowledge to verify a claim absent from the supplied sources is evaluated as failing the task.

2.3 Ground Truth construction

The Ground Truth set was constructed through a human review of every numerical claim in the test document. The process was as follows:

1. **Extraction.** A reviewer extracted every numerical, quantifiable claim in the test document. Claims were recorded verbatim alongside the source location (paragraph and sentence index) from which they were drawn.

2. **Classification.** Each extracted claim was individually checked against the three source documents and classified as **supported** (the claim is supported by at least one source), **contradicted** (the claim directly conflicts with source data), or **unsupported** (the claim appears nowhere in the provided sources).
3. **Adjudication.** A second pass audited the classifications independently. Disagreements, borderline calls, and paraphrased-but-technically-true claims were escalated for group adjudication. Within the error-bearing classifications:
 - **Contradicted claims** include incorrect numerical values (e.g., the document states 14% where the source states 15%), miscalculated derived figures (ratios, growth rates, margins), directional errors (reversal of increase/decrease), and internal inconsistencies within the document itself (conflicting figures in different paragraphs).
 - **Unsupported claims** include absent figures (values the source does not state and cannot be derived from source data), terminology misrepresentations (e.g., labeling a non-GAAP figure as GAAP), and attributed quantitative statements that do not appear in any source.

The review classified **330 total quantitative claims** in the test document: **279 supported**, **30 unsupported**, and **21 contradicted**. The 51 error-bearing claims (unsupported + contradicted) constitute the Ground Truth against which all verification systems are evaluated. This represents a 15.5% error rate – a figure roughly consistent with what we have observed in other LLM-generated drafts of similar documents, and a useful baseline for any practitioner reasoning about the expected prevalence of numerical errors in AI-generated professional content.

2.4 Model and configuration matrix

We test three providers – Anthropic, Google, and OpenAI – across three Model Tiers and two Reasoning Tiers, defined below. The matrices below describe the UI configurations on which the primary study is conducted; we focus on the UI delivery mode for the reasons set out in [Section 2.5](#), and we therefore do not catalogue API-only Reasoning Tiers (e.g., “Full”: Max Effort / Deep Think / XHigh) in this report. Configurations that are not exposed in the consumer UI for a given provider are omitted.

Configurations attempted: 15 LLM configurations, spanning nine distinct (provider × Model Tier) combinations with Baseline and, where available, Extended reasoning, were each run three times.

Configurations included after timing exclusions (as described in [Section 2.7](#)): 12 LLM configurations, plus Rosey. Three OpenAI configurations (GPT Pro Baseline, GPT Pro Extended, GPT Thinking Extended) were excluded because all three of their runs exceeded a 12-minute cutoff. Three additional configurations (Claude Sonnet Extended, Claude Opus Extended, GPT Instant Extended) had one of three runs excluded for the same reason and are reported on $n = 2$.

Model Tiers

Tier	Anthropic	Google	OpenAI	Description
Lite	Claude Haiku (4.5)	Gemini Flash-Lite (3)	GPT Instant (5.3)	Efficient, cost-optimized models
Balanced	Claude Sonnet (4.6)	Gemini Flash (3)	GPT Pro (5.4)	General-purpose performance models
Max	Claude Opus (4.6)	Gemini Pro (3)	GPT Thinking (5.4)	Frontier capability models

Reasoning Tiers

Tier	Anthropic	Google	OpenAI	Description
Baseline	Extended Off	Default	Standard	No extended reasoning enabled in the UI
Extended	Extended On	– (not tested)	Extended	Extended reasoning mode engaged in the UI

2.5 Delivery mode: consumer user interface

All LLM configurations in this study were tested through each provider's web-based user interface – claude.ai, chatgpt.com, and gemini.google.com – rather than through the corresponding API endpoints. This is a deliberate methodological choice.

A parallel API-only study, holding the prompt and source materials constant, was also conducted and can be elaborated upon at a future time. Initial findings show that the APIs for these same LLMs in the same single-shot fashion significantly underperform compared to the UI (the best API configuration scored an F1 of 0.05), which is why that data is excluded.

2.6 Evaluation protocol

Each model configuration received a single structured prompt accompanied by the complete test document and all three source documents concatenated in a consistent order. The prompt instructed the model to identify contradicted and unsupported numerical claims in the test document vis-à-vis the source materials, and to produce a structured, enumerated list of findings. The prompt is shown verbatim in [Appendix D](#).

No configuration had access to the Ground Truth during evaluation. Each configuration was run three times independently in “incognito” or “temporary” mode to measure both central tendency and variance. Runs were conducted over a contiguous window in April 2026 to minimize the risk of provider-side model version changes mid-study.

2.7 Timing and exclusions

We imposed a 12-minute duration cutoff per run. Any run that exceeded 12 minutes – whether due to an API latency spike, a UI response stall, a reasoning trace that failed to converge, or an unrecoverable tool call – was discarded and did not count toward the configuration's averages. Runs that returned a non-parseable output, an explicit refusal, or a request for clarification were also discarded. In this study, three configurations (GPT Pro Baseline, GPT Pro Extended, GPT Thinking Extended) had all three runs exceed the cutoff and were fully excluded from reported results. Three additional configurations (Claude Sonnet Extended, Claude Opus Extended, GPT Instant Extended) had one run of three exceed the cutoff and are reported on $n = 2$. A total of twelve LLM runs were dropped against the cutoff. Full exclusion details are reported in [Appendix A](#) and the interactive results explorer.

The 12-minute threshold is practical rather than theoretical. A verification system with a median per-document latency exceeding 12 minutes is unlikely to see adoption in professional workflows where review cycles are measured in minutes per document. Our goal is to benchmark *usable* verification, not theoretical ceiling performance. The latency of a model under reasoning load is therefore evaluated here as a capability, not an incidental operational detail.

2.8 Scoring

For each “arm” (configuration × delivery mode) we compute Precision, Recall, and F1 against the Ground Truth, per-run, and then average across the arm's retained runs. Definitions:

- A **True Positive (TP)** occurs when a model raises a finding that matches with a Ground Truth error.
- A **False Positive (FP)** occurs when a model raises a finding that does not match a Ground Truth error.
- A **False Negative (FN)** occurs when a given Ground Truth error is not matched to any findings from the model under measurement.
- **Precision** = $TP / (TP + FP)$. Of the findings raised by the model, what fraction correspond to a verified Ground Truth error? Or, for each time an LLM raised a proposed error, how often were those errors in the Ground Truth? Higher is better.
- **Recall** = $TP / (TP + FN)$. Of the verified Ground Truth errors, what fraction did the model find? Higher is better.
- **F1** = harmonic mean of Precision and Recall. Reported on a 0–1 scale; higher is better.

Matching rule. A finding is classified as a True Positive for a given Ground Truth error if it substantively identifies the same underlying accuracy issue. Semantic equivalence is required; exact wording is not. The enumerated output format described in [Section 2.6](#) enforces a strict pairing: a single finding matches at most one Ground Truth error, and a single Ground Truth error matches at most one finding

within a given run. This 1:1 pairing avoids double-credit in both directions (a model cannot claim credit for two errors with one sweeping statement, and a model that raises two distinct findings about the same underlying error collects credit only once).

2.9 Adjudication

Finding-to-Ground-Truth matching was adjudicated through a human audit. Each arm's accuracy findings (flagged contradicted or unsupported claims) were compared pairwise against every Ground Truth error. A human reviewer evaluated each pairing to produce a True Positive (TP), False Positive (FP), or False Negative (FN) designation.

2.10 Rosey

Rosey is a document quality assurance system built and maintained by InSummary, Inc. Rosey was tested on the same task, under the same evaluation protocol, with the same timing and exclusion rules, and its findings were scored against the same Ground Truth by the same adjudication process. Rosey was run three times with the same single-pass protocol used for every LLM configuration, and the version tested in this benchmark is identified in [Appendix D](#). The system's underlying model routing, prompt orchestration, and internal verification logic are proprietary; to support independent evaluation we publish sufficient configuration detail for a user to reproduce Rosey's output on the test document by subscribing to the product under standard commercial terms.

3. Results

Full per-configuration results – including Precision, Recall, F1, raised findings, true positives, false positives, false negatives, retained runs, and excluded runs – are available in the interactive results explorer at <https://www.insummary.com/prf-cbri/prf-cbri.html>. This section reports the headline leaderboard, summary statistics by provider, model tier, and reasoning tier, and the patterns that inform the discussion in [Section 4](#).

3.1 Headline

Of the twelve LLM configurations that completed within the 12-minute cutoff, **none achieved an F1 above 0.41**. The median LLM F1 was 0.26. The best LLM configuration – Claude Sonnet with Extended reasoning – achieved Precision 87% and Recall 28%. Five of twelve LLM configurations scored below F1 0.20. The worst-performing LLM configuration, GPT Instant Baseline, returned F1 0.02 – catching 1% of Ground Truth errors at 12% precision, a result in which the system is effectively silent.

Rosey achieved F1 0.76 on the same task (Precision 87%, Recall 67%). Rosey's precision matches the best LLM configuration; its recall is 2.4× the best LLM and 2.6× the LLM median. Rosey completed in approximately five minutes per run, well within the 12-minute operational cutoff.

3.2 Full leaderboard

The complete leaderboard of included configurations, ranked by F1, appears below. Configurations excluded for exceeding the 12-minute cutoff are listed in [Appendix A](#).

#	Configuration	Precision	Recall	F1	Min
1	Rosey	87%	67%	0.76	5
2	Claude Sonnet · Extended (Anthropic, Balanced)	87%	28%	0.41	9
3	GPT Thinking · Baseline (OpenAI, Max)	54%	33%	0.40	5
4	Claude Opus · Baseline (Anthropic, Max)	66%	25%	0.36	6
5	GPT Instant · Extended (OpenAI, Lite)	47%	28%	0.34	6
6	Gemini Flash · Baseline (Google, Balanced)	60%	22%	0.32	3
7	Claude Opus · Extended (Anthropic, Max)	53%	22%	0.31	9
8	Claude Sonnet · Baseline (Anthropic, Balanced)	53%	13%	0.21	8
9	Gemini Pro · Baseline (Google, Max)	50%	12%	0.20	2
10	Claude Haiku · Extended (Anthropic, Lite)	33%	10%	0.15	3

#	Configuration	Precision	Recall	F1	Min
11	Gemini Flash-Lite · Baseline (Google, Lite)	16%	4%	0.06	1
12	Claude Haiku · Baseline (Anthropic, Lite)	7%	2%	0.03	2
13	GPT Instant · Baseline (OpenAI, Lite)	12%	1%	0.02	1

Values are per-run averages across three runs (two for configurations with one run excluded at the 12-minute cutoff). “Min” = mean time per run in minutes.

3.3 By provider

Averaging the F1 across all configurations within each provider, the three providers cluster within a narrow band: Anthropic at mean F1 0.25 (six configurations), OpenAI at mean F1 0.25 (three configurations included – see note below), and Google at mean F1 0.19 (three configurations). These averages are close enough that provider choice, independent of tier and reasoning settings, is not the strongest lever on evaluation performance.

Each provider's best configuration performs comparably: Anthropic's Claude Sonnet Extended (F1 0.41), OpenAI's GPT Thinking Baseline (F1 0.40), and Google's Gemini Flash Baseline (F1 0.32). The meaningful variation in performance is within provider – between model and reasoning tier – not between providers.

Note on OpenAI. Three of the six OpenAI configurations we attempted were fully excluded for exceeding the 12-minute cutoff on all three runs: GPT Pro Baseline, GPT Pro Extended, and GPT Thinking Extended. The OpenAI provider average above reflects only the three configurations that completed. The inability to complete a single-pass document verification within 12 minutes is itself a measurement – discussed in [Section 4.3](#) and treated as a capability result, not a missing data point.

3.4 By Model Tier

Averaging F1 across all configurations at each Model Tier: Lite tier mean F1 0.12 (5 configurations), Balanced tier mean F1 0.31 (3 configurations), and Max tier mean F1 0.32 (4 configurations). The relationship is not strictly monotonic. Stepping up from Lite to Balanced approximately doubles mean F1; stepping from Balanced to Max produces no appreciable improvement.

Based on this data for document verification, we observe the Max tier offers no significant advantage over the Balanced tier on this task.

3.5 By Reasoning Tier

At the aggregate level, enabling Extended reasoning improves mean F1 on average. Averaging across all configurations, Baseline reasoning produces mean F1 0.20 (eight configurations) and Extended reasoning produces mean F1 0.30 (four configurations). Additional Extended configurations (GPT Pro

Extended, GPT Thinking Extended, and Gemini extended models) could not be measured – either because all three runs exceeded the cutoff or because the consumer UI did not expose an Extended toggle at the time of evaluation.

However, the within-model picture is not monotonic, and in at least one case is reversed: **Claude Opus Baseline (F1 0.36) outperforms Claude Opus Extended (F1 0.31) on the same document and the same task.** A notion that “more reasoning is better” is not necessarily true. We return to this result in [Section 4.3](#).

3.6 Precision vs. Recall

A recall-focused view of the results is misleading. High recall paired with low precision means that every finding the model raises must be checked by a human – and checking dozens of false positives is typically slower than reviewing the document from scratch, particularly when the false positives are plausibly worded and require returning to the source material to confirm/refute. An undesirable outcome for a document-verification system is not a missed error but rather a flood of false alarms that trains the user to ignore the system entirely.

At the opposite extreme, high precision paired with low recall means that the findings the system produces are reliable, but most errors in the document go uncaught. For a one-shot review – the typical workflow for a document under deadline – a high-precision, low-recall reviewer misses most of the errors the user needs to fix.

The Cracker Barrel data shows LLMs distributed across both failure modes. Eight of twelve included configurations have precision below 60% – i.e., more than two of every five findings they raise are false positives. The highest-precision LLM configuration (Claude Sonnet Extended, 87% precision) catches only 28% of the Ground Truth errors; if this reviewer went over the document, 37 of 51 errors would remain in it. The highest-recall LLM configuration (GPT Thinking Baseline, 33% recall) has precision 54% – so roughly half the findings it raises are wrong, and two-thirds of real errors are still missed. No LLM configuration achieves simultaneously strong precision and strong recall. [Section 4.2](#) develops this finding into its implication.

3.7 Rosey

Rosey achieved F1 0.76 on this benchmark – Precision 87%, Recall 67%. In each run, Rosey raised 39 findings on average; approximately 34 matched a Ground Truth error and approximately 5 did not. Rosey missed approximately 16 of 51 Ground Truth errors on average. It completed each run in approximately five minutes, within the 12-minute operational cutoff.

Rosey's precision of 87% matches the highest-precision LLM configuration (Claude Sonnet Extended). The differentiating dimension is recall: Rosey caught approximately 2.4× as many Ground Truth errors at

equivalent precision. Rosey's findings are presented directly within MS Word with tracked changes, citations, and commentary on potential weaknesses or omissions.

4. Discussion

4.1 The “verification gap” is significant and structural

No LLM configuration tested achieved an F1 above 0.41 on an earnings summary containing 51 known errors among 330 numerical claims. The median LLM F1 was 0.26. Five of twelve included configurations scored below F1 0.20 – an outcome in which the system is arguably worse than a careful reader with no tool at all, as it produces confident findings that point the user's attention away from the real errors. This is not a story about a few weak models; it is a pattern across twelve configurations from three leading providers, tested on the consumer interfaces those providers ship to professional users.

Generative models are trained to produce plausible completions of text, not verified ones. When the prompt asks for errors, the model produces errors, and a meaningful fraction of those outputs have no relationship to the source material. Conversely, the model's attention during the verification pass is distributed across an entire document – and without a verification-specific objective, it will tend to identify the most syntactically salient potential issues rather than the ones that matter. Both failure modes show up in the data: low-precision configurations raise confident nonsense, and low-recall configurations leave most of the real errors in the document.

Even as models, reasoning, and prompting practices improve, it is fair to expect this gap to persist given how generative models are trained. Based on this study, when a firm is evaluating whether to rely on an LLM to check an LLM-generated document, we believe that the answer, on the present evidence, is that it should not.

4.2 Precision without recall is not verification

A finding that reads most clearly from the leaderboard: Claude Sonnet with Extended reasoning achieves **the same precision as Rosey** – 87% in both cases. It is the highest-precision LLM configuration in the study. Its recall, however, is 28% against Rosey's 67%. At equivalent precision, Rosey catches approximately 2.4× as many Ground Truth errors.

This matters because document verification is not a retrieval task. The user of a verification system should not be required to interact with it over many turns, refining queries and keeping the results they like. They expect to read a document once, run the verifier, review its findings, and send the document to its next destination. If the verifier's recall is 28%, approximately 37 of 51 errors pass through untouched. While the findings list the user acts on is fairly accurate – 87% of what was raised was correctly raised – the document goes out with nearly three-quarters of its errors intact. In verification, precision without recall creates a false sense of confidence: the user has worked through a clean list of fourteen issues and feels the document has been reviewed, when in fact most of the errors remain.

The inverse failure is also undesirable. The higher-recall, low-precision configurations in the leaderboard (e.g., GPT Instant Extended at 28% recall and 47% precision, or GPT Thinking Baseline at 33% recall and

54% precision) produce findings lists that are, by construction, majority noise. The verifier's time is spent disqualifying false alarms rather than fixing real errors. In practice, this failure mode causes users to disengage from the tool entirely – an outcome worse than simply missing errors, because it discourages the habit of checking.

The verification task therefore has a simultaneous precision-and-recall requirement that is strictly more demanding than either metric in isolation. We believe this data shows that prompting general-purpose LLMs for errors is not a drop-in verification layer. On the present data, no LLM tested is close to meeting the simultaneous bar. Rosey does by a wide margin.

4.3 More reasoning is not monotonically more accurate – and some reasoning does not finish

The intuition that additional reasoning allocation should monotonically improve verification performance is wrong, at least on this task. Averaging across all configurations for which both modes were tested, Extended reasoning does produce a modest F1 improvement over Baseline (0.30 vs. 0.20). But the improvement is not uniform, and in at least one case reverses cleanly: **Claude Opus Baseline achieves F1 0.36 against Claude Opus Extended's 0.31 on the same document and the same task.**

This is consistent with a simple mechanistic hypothesis: reasoning training rewards the production of long, plausible chains of thought, not the suppression of unsupported claims. A verification task – where the correct behavior is sometimes to conclude “I find no error here” – does not align with the training objective of extended reasoning.

Three OpenAI configurations – GPT Pro Baseline, GPT Pro Extended, and GPT Thinking Extended – had all three of their runs exceed the 12-minute cutoff, with typical per-run times between 22 and 50 minutes. A verification system that takes 45 minutes to analyze a short professional document is not a verification system that achieves adoption in a professional workflow with real document volume. The reasoning budget itself is a capability gate: at high reasoning allocations, the question is not only whether the model can verify the document but also whether it can do so inside any practically bounded time. On the present data, for a meaningful fraction of the configurations tested, the answers to both are no.

4.4 The case for an independent review layer

The three findings above point to a consistent conclusion: no frontier model today can reliably verify numerical accuracy in a document touched by AI.

Mature domains that involve critical processes that have achieved this level of reliability have done so by separating the producer from the verifier. Financial statements are prepared by management and audited by a separate firm. Code is written by engineers and reviewed by different engineers against a separate test suite. Drug trials are run by the sponsor and reviewed by an independent regulator. Legal

briefs are drafted by one team and cite-checked by another. The separation of creation from verification is not a bureaucratic nicety in these fields; it is the structural mechanism by which errors are caught. It works because the incentives, training, tooling, and attentional focus of a verifier are fundamentally different from those of a creator.

The same argument applies to AI-generated professional documents. The verification layer should not be the same model that produced the draft, should not be in the same conversational context, and should not share the creator's prior of "this document is probably correct." It should be an independent system with a verification-specific objective, its own evaluation infrastructure, and measurable accuracy against a known Ground Truth. Rosey is implemented on that thesis.

5. Implications for Practitioners

The findings above have direct operational implications for any organization that produces numerically-dense professional documents with the assistance of LLMs – which, increasingly, is every such organization. We highlight four domains in which the implications are immediate.

5.1 Private equity and M&A diligence

Investment committee memos, diligence reports, and deal summaries are increasingly drafted with LLM assistance. The documents are numerically dense, consequential, and often produced under time pressure. The data reported here should be read as a direct warning against relying on the same LLMs that draft memos to check them. Even the best LLM configuration in this study catches only 28% of the errors a human-validated review surfaces; eight of twelve configurations produce findings lists that are more than 40% false positives. Firms should either maintain a human verification pass on every LLM-drafted document before it reaches an investment committee, or adopt an independent verification system with measurable accuracy against a known benchmark. Verification-by-AI-review is not a reliable practice.

5.2 Management consulting

The consulting deliverable is, operationally, the product; it is the artifact on which brand, fees, and renewals ride. Firms that integrate LLM-assisted drafting into their client-facing workflows are exposed to the error modes documented in this report. The cost of a single shipped deliverable with even one material numerical error can exceed the efficiency gains from AI-assisted drafting across a portfolio of engagements. At the precision levels observed here, it is clear that most of the errors present in the document go unflagged by an LLM.

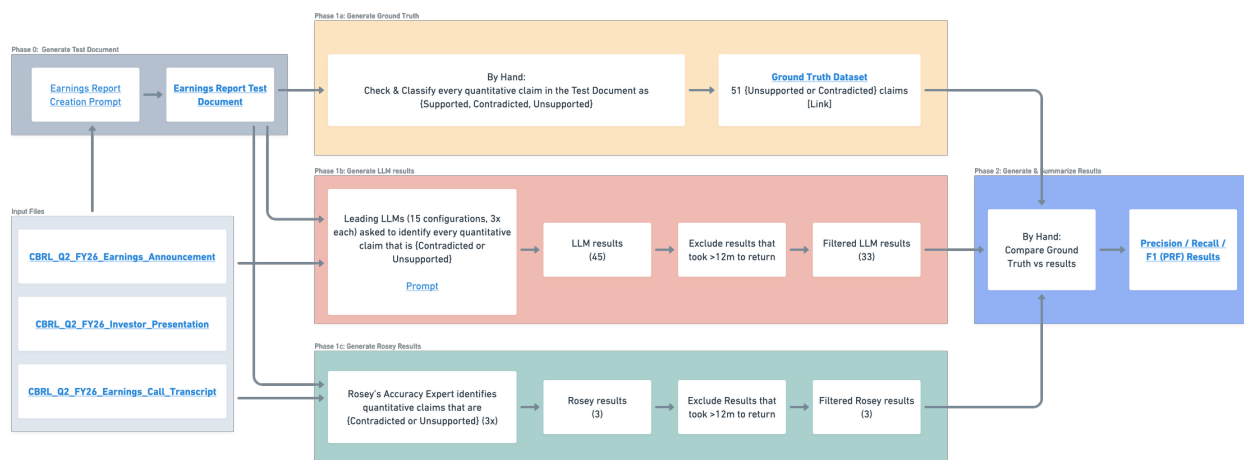
5.3 Corporate strategy, IR, and finance

Investor relations, FP&A, and corporate strategy teams produce earnings summaries, board materials, management commentary, and strategic memos that flow upward to senior executives and outward to external constituencies. These are exactly the documents this study's test artifact is representative of – the test document in this benchmark is an earnings summary. A numerical error in an IR-drafted earnings summary, or an unsupported claim in board materials, has reputational and legal consequences that dwarf the drafting efficiency gains from LLM use. The data in this report is a direct input to that risk calculus.

5.4 Scientific and technical publishing

Peer review has, for centuries, been the separation of creation from verification in science. LLM-assisted drafting of scientific manuscripts has arrived faster than review infrastructure has adapted, and the categories of error documented in this report (numerical contradictions against cited sources, unsupported claims presented as established) map directly onto the failure modes most damaging to scientific trust. Journals, preprint servers, and institutional research offices have a direct interest in the same kind of independent, measurable document verification layer this report argues for.

6. Replication Protocol



To support independent replication, audit, and extension of this study, we release the following in full:

- The [test document](#) (the LLM-generated Cracker Barrel earnings summary) in its original form.
- The three source documents ([earnings announcement](#), [investor presentation](#), [earnings call transcript](#)), or direct citations to their public sources.
- The [prompt](#) and [example file](#) used to produce the test document.
- The [prompt](#) presented to each candidate LLM.
- [Ground Truth Viewer](#) – all 51 verified numerical accuracy errors – along with all 279 supported numerical claims – with source citations, classification, and adjudication notes.
- An interactive [Results Explorer](#) enabling filtering by provider, model tier, and reasoning tier.

7. About Rosey

Rosey is an independent document quality assurance system for professional documents that have been touched by AI. It lives within the existing document workflow (delivered today as a Microsoft Word Add-in) and reviews financial, scientific, and technical documents for errors, omissions, and weaknesses.

Rosey does not generate documents; it reviews them. This separation is the operational expression of the conclusions drawn from this study.

Rosey is built and maintained by InSummary, Inc. More information and a product demo are available at **insummary.com** or by contacting **sales@insummary.com**.

Appendices

A. Configurations excluded for timing

The following configurations exceeded the 12-minute cutoff on all three runs and are not represented in the [Section 3.2](#) leaderboard.

Configuration	Provider	Runs excluded	Time per run (min)
GPT Pro · Baseline	OpenAI (Balanced)	3 / 3	45 / 36 / 22
GPT Pro · Extended	OpenAI (Balanced)	3 / 3	50 / 38 / 34
GPT Thinking · Extended	OpenAI (Max)	3 / 3	17 / 12 / 14

Three additional configurations had a single run of three exceed the cutoff and are reported on n = 2 in the leaderboard: Claude Sonnet · Extended, Claude Opus · Extended, and GPT Instant · Extended (each with one run of 12+ minutes; two runs retained).

B. Error taxonomy, with examples

The Ground Truth error taxonomy distinguishes between contradicted and unsupported numerical claims. The examples below illustrate the shape of each category. The complete Ground Truth, with source citations and adjudication notes for each of the 51 errors, is included in the released benchmark package.

Contradicted numerical claims (21 errors)

- *Numerical value error.* The test document stated a net income figure of \$47.1M; the earnings release disclosed a different value.
- *Derived figure miscalculation.* The test document reported a year-over-year basis-point change in operating margin, derived from two source figures, where the arithmetic was incorrect.
- *Directional inversion.* The test document characterized a line item as growing year-over-year where the source data showed a decline.
- *Internal inconsistency.* The test document stated different values for the same metric in two places within itself.

Unsupported numerical claims (30 errors)

- *Absent figure.* The test document states a numerical value or range that appears in no source document and cannot be derived from source data (e.g., GT015–GT021 – various guidance-range statements).

- *Terminology misrepresentation.* The test document labeled a non-GAAP metric as GAAP, or conflated two different measures.
- *Frequency/count misstatement.* The test document asserts a count (e.g., GT046, GT047 – “2,” “1”) that cannot be verified against any source.
- *Attributed statement not in source.* The test document presented a quoted or attributed statement that does not appear in the earnings release, investor presentation, or call transcript.

C. Test document creation prompt

The prompt used to generate the test document is shown below. The example file is included in the replication kit.

```
Attached is an Earnings Summary prepared for VMH for Q4 2024
[VMH_Q4_2024_Earnings_Summary.docx].

Attached are the resources from the CBRL Q2 2026 Earnings Announcement:
the earnings announcement, earnings call transcript, and earnings
investor presentation.

Put together an Earnings Summary for CBRL Q2 2026 similar to the one
prepared for VMH Q4 2024.
```

D. Evaluation prompt

The evaluation prompt used in every model configuration is reproduced below. Minor formatting adjustments (line breaks, source-document delimiters) were made to accommodate each provider's preferred input format; the substantive instructions were identical across providers.

```
You are a senior reviewer conducting a thorough accuracy review of a
document before it is sent to its intended audience. Your review should
reflect the standard of care expected in professional services - the
kind of review a managing director would do before a document goes to a
client, a board, or an investment committee.

---

## The Document Under Review

test_document.docx **(Attached)**

---

## Source Materials

These are the authoritative source documents. Use them to verify every
factual claim, figure, and data point in the draft.

- CBRL-Q2-FY2026_EA.pdf
- CBRL-Q2-FY2026_IP.pdf
- CBRL-Q2-FY2026_Transcript.pdf
```

Your Review

Produce the output section described below.

Accuracy Findings

****Be exhaustive and comprehensive. Flag every verifiable discrepancy between the document and the source materials, no matter how minor it appears.****

Scope constraint: Flag only claims that contain a specific number, figure, date, count, percentage, ratio, or derived calculation. Do not flag claims that are purely qualitative or characterizing, even if they appear imprecise or misleading.

Review every verifiable claim in the document against the source materials. Only two types of issues belong in this section:

****Type 1 - Factual Contradictions:**** A claim in the document that directly conflicts with a source document.

- Dollar amounts, percentages, ratios, counts, and dates that differ from the sources
- Derived calculations (YoY changes, margins, growth rates) that do not match what can be computed from source data
- Directionality errors where a specific quantified figure is also cited (e.g., document says revenue 'grew 12 percent' when source shows a decline)
- Internal inconsistencies where one section contradicts another within the document

****Type 2 - Unsupported Claims:**** A specific numeric or quantitative assertion that cannot be verified against any provided source document.

- A specific figure cited that does not appear in any source
- Quantified claims where the source materials are silent
- Quantitative labels that are technically wrong (e.g., 'CAGR' applied to a single-period comparison where a specific figure is cited)

****Do not include**** - these are findings that are out of scope for this review:

- Style, grammar, spelling, punctuation
- Formatting, layout, visual presentation
- Structural or organizational observations (e.g., missing sections, section ordering)
- Tone or word choice
- Analytical depth or completeness of observations

- Clarity, ambiguity, or logical gap observations
- Any assessment of whether the document achieves its purpose with its intended audience
- Qualitative or descriptive characterizations that do not involve a specific numeric claim (e.g., calling a trend 'strong' or 'accelerating' without a figure attached)
- Any finding that is not a factual contradiction or unsupported claim

For each finding, provide:

- **Title:** A concise, specific name for the discrepancy
- **Type:** `'factual_contradiction'` or `'unsupported_claim'`
- **Claim Text:** The verbatim text as it appears in the document that represents the claim.
- **Location:** Section name and, where applicable, the specific figure, row, or quoted text
- **Detail:** What the document claims and what the source shows (or why it cannot be verified). For contradictions, state both the document value and the correct source value. Quote the source directly where possible.
- **Source quote:** The verbatim source text (or "N/A - not found in any source" for unsupported claims)
- **Suggested fix:** The specific correction to apply to the Claim Text.

IMPORTANT: You must enumerate a finding for each and every Unsupported Claim or Factual Contradiction. If the same error occurs multiple times in the document, then each instance must be represented as its own finding.

If you are genuinely uncertain whether something is an error - for example, if source materials are ambiguous or you cannot locate the relevant data point - include it as a finding and mark it "unverifiable" rather than declaring it correct or incorrect.

OUTPUT 1: Accuracy Findings

Present your findings using the following format:

...

Accuracy Findings

[All factual contradictions and unsupported claim findings, numbered A1, A2, A3, ...]

...

*Label each finding with its type: ****[Factual Contradiction]**** or ****[Unsupported Claim]****.*

Begin your review now.

E. Rosey configuration

Rosey v0.3.2 was exercised on the test document using the commercially available product in April 2026.