

Definitional or Computational? Decomposing LLM Numeric Fidelity in Financial-Statement Analysis with an XBRL Cascade Oracle

Oussama Chahed

Bulb’s Head SAS, Paris, France

Université Paris Dauphine–PSL, Paris, France

chahedous@gmail.com oussama.chahed@dauphine.eu

Abstract

Large language models (LLMs) analyze corporate financial statements fluently, yet routinely emit wrong figures while preserving a confident, well-formatted narrative—the *false-precision* problem. Prior benchmarks measure answer accuracy or attach claims to source cells, but do not explain *why* a financial number is wrong. We introduce a deterministic, XBRL-grounded *oracle* that recomputes the intermediate-management-balances (SIG) and ratio *cascade* from SEC EDGAR facts, and use it to score numeric fidelity per financial concept, with fact-level attribution and error propagation along the ratio dependency DAG. A *pinned/unpinned* prompt contrast, an *any-standard-definition* oracle for convention-dependent composites, and a rule-based *error-mode classifier* jointly separate *definitional divergence* (the model applies a different but sometimes valid convention) from *unexplained residual numerical error*. Across six open and proprietary models on 729 firm-years (and a 345 firm-year subset common to all six), we find: (i) simple two-fact ratios are computed faithfully by all but the weakest model, while multi-input composites separate the panel by capability—the residual hardest concept (ROCE) is executed reliably only by the strongest models; (ii) supplying the exact formula *and* source tags (pinning) helps every capable model and only hurts the weakest, inverting an apparent “computational limit” that turns out to be a weak-model artifact; and (iii) source-fact attribution and fidelity are distinct and can decouple—one model computes correctly yet cites its source facts almost never (“right answer, no receipts”), while the strongest models pair high fidelity with high attribution. Program-of-thought prompting does not change fidelity, and most wrong composites are local errors rather than propagated ones. Code, frozen model responses and full provenance are released under a DOI.

1 Introduction

Financial-statement analysis is deterministic: profitability, liquidity, solvency, leverage and value-creation metrics are closed-form functions of reported line items (Penman, 2013). Recent work shows LLMs can perform such analysis usefully—Kim et al. (2024) report GPT-4 predicting next-year earnings direction from anonymized statements above analyst consensus. Yet the same models are unreliable on the numbers themselves: on FinanceBench a strong retrieval-augmented model is wrong or abstains on 81% of filing questions (Islam et al., 2023), the errors clothed in fluent prose. The field has documented *that* LLMs hallucinate financial numbers; it has not characterized *why*.

We reframe the question. LLM financial-fidelity failures are not a single phenomenon but a mixture of mechanisms—convention/*definitional* divergence, magnitude/scale errors, and residual numerical failure—and disentangling them is both possible and necessary. The enabling idea exploits the determinism of the discipline: the SIG and ratio machinery forms a dependency DAG that recomputes exactly from XBRL, giving a free verification oracle. The oracle lets us measure fidelity per concept, attribute each claim to its source fact, and observe error *propagation* along the cascade.

We run each model *pinned* (the exact formula and source tags are supplied) and *unpinned* (the model uses its own definitions). Pinning is a *compound* intervention: it fixes the definition, but it also names the source facts and removes planning ambiguity. The contrast therefore measures the combined benefit of an explicit formula and explicit source specification; we use it as an *upper bound* on the definitional component and rely on the any-standard-definition oracle and the error-mode classifier to sharpen the decomposition (§3.3). For composites with several legitimate conventions (e.g. gearing), the any-standard oracle credits a valid al-

ternative rather than scoring it wrong; the classifier labels every wrong claim, turning the decomposition into a measurement.

Contributions. (1) An XBRL cascade oracle and per-concept fidelity benchmark—the SIG/ratio dependency DAG as ground truth, scored as a profile rather than a single rate. (2) A method that separates definitional from residual numerical failure (pinned/unpinned contrast, any-standard oracle, rule-based error-mode classifier). (3) A cascade error-propagation metric, reported per model and condition. (4) Findings across six models, on the full sample and on a common subset: composite fidelity separates the panel by capability; pinning helps every capable model (the apparent “computational limit” is a weak-model artifact); attribution and fidelity are distinct and can decouple; program-of-thought does not improve fidelity. (5) A reproducible, governance-gated, open-data pipeline with frozen model outputs released under a DOI.

2 Related work

Financial numeric reasoning and QA. FinQA, ConvFinQA, TAT-QA, MultiHiertt and DocFinQA evaluate numerical reasoning over tables and text (Chen et al., 2021, 2022; Zhu et al., 2021; Zhao et al., 2022; Reddy et al., 2024); FinanceBench evaluates open-book filing QA (Islam et al., 2023); FinBen aggregates financial tasks (Xie et al., 2024); and, most recently, FinanceReasoning (Tang et al., 2025) pairs financial numerical problems with executable Python solutions, and FAITH (Zhang et al., 2025) measures intrinsic tabular hallucination as calculation depth grows. These score a final answer (or an answer-level hallucination rate), not a *recomputed accounting cascade*, and do not *decompose* the error into definitional versus residual causes—the gap we fill.

LLMs for financial-statement analysis. Kim et al. (2024) establish predictive analysis ability; FinRobot (Yang et al., 2024) generates analyst reports but provides no fidelity oracle.

Grounding, attribution and XBRL recompute. FinGround performs atomic-claim verification with table-cell citations and a computational claim type (Guo et al., 2026); FinDVer and FinLFQA address explainable claim verification and attributed generation (Zhao et al., 2024; Long et al., 2025); FinTagging links facts to the US-GAAP taxonomy (Wang et al., 2025); XBRL Agent compares chain-of-thought with tool-augmented formula compu-

tation (Han et al., 2024); AuditFlow recomputes expected values from a US-GAAP/XBRL graph for audit verification (Li et al., 2026). We reuse this grounding substrate but differ on three axes: we recompute the *continental SIG cascade* (value added, EBITDA/EBE, capital employed, working-capital requirement, EVA) as a *dependency DAG*; we report a *per-concept fidelity profile*; and we *separate definitional from residual failure*, which none of the above isolate. Concretely, the closest concurrent systems verify or recompute *individual* claims and report pass/fail—AuditFlow (Li et al., 2026) against a US-GAAP/XBRL graph, FinGround (Guo et al., 2026) via atomic-claim types—but neither yields a per-concept fidelity *profile* nor an explained-vs-local error propagation along the cascade; and FAITH (Zhang et al., 2025) measures how hallucination grows with calculation depth without distinguishing a wrong *convention* from a wrong *number*.

Numeric reasoning and financial hallucination.

Chain-of-thought (Wei et al., 2022), program-of-thought (Chen et al., 2023) and PAL (Gao et al., 2023) improve multi-step computation; we use the CoT/PoT contrast. Hallucination is broadly surveyed by Ji et al. (2023). Most directly, Zhang et al. (2025) quantify *intrinsic tabular hallucination* on S&P 500 filings and find accuracy collapsing as tasks move from direct lookup to multivariate calculation—a calculation-depth effect. Our oracle *localizes* that effect to specific concepts (ROCE the residual), *separates* its definitional from its residual component, and shows it depends on model capability rather than being a fixed limit. This diagnosis also motivates the neuro-symbolic prescription of relieving LLMs of arithmetic via deterministic ledgers; our XBRL cascade oracle is exactly such a ledger, used here for *measurement* rather than as a runtime tool.

3 Method

3.1 Firm-year pivot and the XBRL oracle

The unit of analysis is the firm-year (CIK, fiscal year, accession). From SEC EDGAR *companyfacts* we extract us-gaap concepts via a candidate-tag map (first available tag wins; misses logged) and recompute, in closed form, gross/operating/EBITDA margins, ROE, ROCE, current and quick ratios, net debt, gearing, debt/EBITDA, working-capital-requirement in days, and EVA (Penman, 2013; Stewart, 1991). We use these intermediate-

management-balance aggregates (the continental *soldes intermédiaires de gestion*) not as a jurisdictional reporting requirement but as a deterministic analytical cascade over reported line items—a dependency structure that recomputes exactly from US-GAAP XBRL facts and that we exploit purely as a verification oracle. Each derived value records its *lineage* (tags, raw values, period, unit). The ratios form a dependency DAG—e.g. EBITDA-margin and debt/EBITDA descend from EBITDA-proxy; EVA descends from ROCE and capital employed—the object of the propagation analysis (§3.5). Division by zero or missing inputs yield a null, never an exception; conventions (e.g. current debt defaults to zero when untagged) are fixed and documented.

The oracle is deterministic, but determinism is not the same as accounting truth: candidate-tag selection and fixed conventions are modelling choices, and a firm whose current debt is reported under an unmapped tag receives a precise but convention-dependent reference value. The any-standard-definition oracle (§3.3) absorbs the most common alternative conventions; the residual mapping risk is stated in the Limitations and is the reason we report fidelity as a threshold *sweep* rather than a single exact-match rate.

3.2 Fidelity scoring

Each model returns, per ratio, a structured *Numeric-Claim* {concept, value, unit, cited_fact_id}. Value fidelity is the relative error against the oracle, reported as a sweep at 0.1%/1%/5% (a single threshold is misleading); the primary threshold is 1%. A unit/scale normalization precedes comparison: because the percent/decimal label is unreliable, the divide-by-100 decision is taken by *magnitude*, not by the reported unit (a margin reported as 38.6 with an empty unit is read as 0.386; a value of 0.386 is left unchanged; a scale error is then a value that remains off by a power of ten *after* this normalization). Attribution is scored as *partial source attribution*: it is correct when the single cited_fact_id resolves in the fact catalog to a tag in that ratio’s lineage. This establishes that the model can point to a genuine input, not that it cites the complete lineage of a multi-input ratio; full-lineage coverage would require a per-input citation schema, which we leave to future work and do not claim here. We record three mutually exclusive outcomes—parsed answer, refusal/unknown, unrecoverable parse failure—and report parse failure as its own per-model metric.

3.3 Pinned vs. unpinned; the any-standard oracle

Unpinned, the model is asked for a ratio with no formula; *pinned*, the exact formula is supplied, naming the precise catalog tags and conventions verbatim. Pinning changes several things at once—it resolves definitional ambiguity, but it also simplifies source-fact selection and removes planning—so the contrast pinned–unpinned measures the combined benefit of an explicit formula and explicit source specification, not a pure definitional effect. We observed this directly: a pinned prompt that spells the formula in generic English (“+ current debt”) without naming the exact tags underperforms tag-specific pinning, so the tags do real work. We therefore read the contrast as an upper bound on the definitional component and use two further instruments to sharpen it. For composites with multiple legitimate conventions (gearing as net-debt/equity, gross-debt/equity, or debt/(debt+equity)), an any-standard-definition oracle credits a match to any standard form, separating *valid alternative convention* from error; and the error-mode classifier below labels what remains. Separating formula-only from tags-only pinning is the natural next experiment and is not run here.

3.4 Error-mode classifier

Every *wrong* claim is labeled into mutually exclusive modes: different-concept (matches a sibling ratio of the same firm-year, e.g. EBITDA-margin reported for operating-margin); scale (off by a power of ten after normalization); valid convention (matches an alternative standard definition); undefined; and unexplained residual numerical error—the remainder once the rule-based explanations fail. Every mode is a deterministic, rule-based predicate computed against the oracle, so no learned or LLM judge is involved and the decomposition is itself reproducible. We deliberately do not call the residual “arithmetic error”: without reconstructing the model’s computation, a residual wrong value may come from a wrong fact or period, an omitted input, an unmodelled convention, or a genuine arithmetic slip, and the classifier cannot tell these apart. What it establishes is narrower and still useful: after crediting scale artifacts, sibling confusions and valid conventions, a measurable fraction of wrong composites has *no definitional explanation*. This converts the qualitative “why” into per-concept, per-model fractions, replacing a misleading single

median (uninformative for the bimodal error distributions we observe).

3.5 Cascade error-propagation

For each wrong descendant ratio we test whether recomputing it from the model’s *own claimed ancestor* plus the true remaining inputs reproduces the model’s wrong value. If so the error is explained (propagated); otherwise local. The propagation score is explained/(explained + local) over wrong descendants whose ancestor was claimed; ancestors not claimed form a separate undefined bucket. Results are in §5.5.

3.6 Reproducibility and governance

Every record carries provenance (source, URL, accession, time, SHA-256); the oracle carries cell-level lineage; every model call logs model id, provider, training cutoff, prompt-template hash and sampling. A cost ledger and budget circuit-breaker bound spend; eight quality gates gate the pipeline. Decoding is deterministic (temperature 0); because serverless *seed* is best-effort, reproducibility rests on the frozen response cache, archived so results survive model deprecation. A leakage gate requires every input filing to post-date the models’ training cutoffs; unlike a forecasting task, D1 has no future “label-year” filing, so only the analyzed filing’s date is gated.

Pre-registration. The measurement design was fixed before scoring: the materiality screen and in-scope universe, the fidelity-threshold sweep (0.1/1/5%) and magnitude-based normalization, the five-way error-mode taxonomy, the pinned/unpinned and any-standard-definition protocol, and the leakage gate. No concept, threshold, or model was added or dropped after inspecting outputs. The common-subset, per-model CoT/PoT and per-model propagation analyses added for the camera-ready were requested by reviewers and are recomputed from the frozen cache with the released evaluator; they add no new inference.

4 Experimental setup

Data. SEC EDGAR annual filings (10-K) filed on or after 1 February 2025, i.e. strictly after the latest *verified* model training cutoff (Gemini, January 2025). A materiality screen defines the in-scope universe: us-gaap filer, classified balance sheet, revenue present, total assets $\geq$ \$50M; financials and utilities (SIC 6000–6799, 4900–4949)

Model	Provider	cutoff	firm-years
Llama-3.1-8B	DeepInfra	2023-12	729
Qwen2.5-72B	DeepInfra	2024-09	729
Gemini-3.1-Flash-Lite	Google	2025-01	729
GPT-5-mini	OpenRouter	2024-05	729
DeepSeek-V3.2	SiliconFlow	2025-12 [†]	345
Grok-4.3	OpenRouter	2025-12 [†]	345

Table 1: Six-model panel. The training cutoff gates eligibility: four models cover all 729 firm-years; DeepSeek and Grok cover the 345 firm-years filed after their ([†]provider-reported, unverified) cutoff. Those 345 firm-years form the common subset used for all cross-model comparisons.

are excluded. The frame is 400 firms / 729 firm-years: because the window spans roughly a year of filings, most firms contribute both their FY2024 report (filed February–April 2025) and their FY2025 report (filed late 2025 or early 2026); by XBRL fiscal-year label the 729 firm-years split into 307 FY2024, 378 FY2025, 43 FY2026 (non-calendar fiscal years) and one FY2023 late filer. Out-of-scope firms are logged, not counted against coverage.

Models and samples. A panel spanning capability tiers and providers, each with a documented training cutoff used for leakage control (Table 1). Four models (Llama-3.1-8B, Qwen2.5-72B, Gemini-3.1-Flash-Lite, GPT-5-mini) have verified cutoffs before the window and are evaluated on all 729 firm-years. Two models (DeepSeek-V3.2, Grok-4.3) have provider-reported cutoffs of December 2025, so the leakage gate retains for them only the 345 firm-years filed after 4 December 2025 (302 FY2025 + 43 FY2026 labels, 345 distinct firms). We call these 345 firm-years the common subset: all six models were run on it, and every cross-model comparison below is reported on it as well as on each model’s full sample. Each firm-year context is built once and fanned out to the D1 variants {unpinned, pinned} $\times$ {CoT, PoT}; decoding is temperature 0. The full six-model run cost well under a \$50 cap.

5 Results

5.1 Fidelity separates the panel by capability

Per-concept fidelity forms a clean ladder (Fig. 1): Llama-3.1-8B is a floor (many composites near zero), Qwen-72B a middle tier, and the four strong models (Gemini, DeepSeek, Grok, GPT-5-mini) are high across most concepts. Simple two-fact

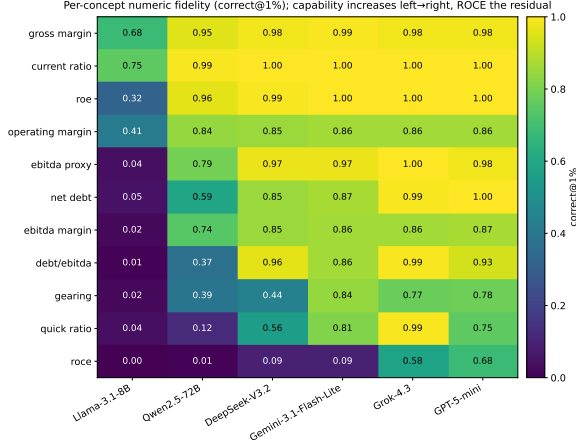


Figure 1: Per-concept fidelity (correct@1%), 11 concepts $\times$ 6 models, pooled over conditions. Capability increases left $\rightarrow$ right; ROCE is the residual hard concept.

ratios (current ratio, gross margin, ROE) are computed faithfully by everyone except Llama. Multi-input composites separate the panel: quick ratio, gearing, and especially ROCE—a multi-step aggregate ($\text{EBIT} \times (1 - \tau)$ over capital employed)—collapse for weaker models and are the residual hard concepts. ROCE is the sharpest separator: pooling conditions, GPT-5-mini reaches 0.68 correct@1% and Grok 0.58, every other model ≤ 0.09 ; *pinned*, GPT-5-mini reaches 1.00 and Grok 0.84, while Gemini and DeepSeek stay near 0.17. With six models this is a description of the panel, not a law; we use “capability-dependent” rather than “capability-gated” in what follows.

5.2 Pinning helps capable models; the “computational limit” is a weak-model artifact

Supplying the exact formula and source tags (pinning) raises overall fidelity for five of six models; the lone exception is Llama (Fig. 2, $\Delta = -0.028$), which cannot reliably execute even a fully-specified formula (e.g. current ratio -0.14). For capable models pinning is neutral-to-strongly-helpful, and it is largest exactly on the hard composites: GPT-5-mini ROCE $+0.749$ and quick ratio $+0.543$; DeepSeek gearing $+0.790$; Gemini debt/EBITDA $+0.221$; GPT-5-mini debt/EBITDA $+0.144$. An earlier four-model reading (dominated by weaker models) had suggested the EBITDA chain was a *robust* “pinned $<$ unpinned computational limit”; the six-model panel falsifies that—capability *inverts* the effect. The correct statement is: the hardest

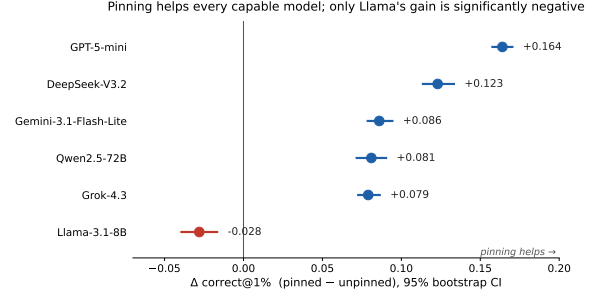


Figure 2: Overall $\Delta \text{correct@1\%}$ (pinned – unpinned) per model, each on its full sample. Pinning helps every capable model; only the weakest (Llama) is hurt.

Model	unpin.	pin.	Δ [95% CI]	attrib.
Llama-3.1-8B	0.315	0.291	$-0.024[-.041, -.006]$	0.684
Qwen2.5-72B	0.618	0.697	$+0.079[.064, .093]$	0.707
Gemini-3.1-F-L	0.793	0.882	$+0.089[.076, .100]$	0.049
GPT-5-mini	0.807	0.969	$+0.163[.153, .173]$	0.826
Grok-4.3	0.872	0.951	$+0.079[.071, .087]$	0.929
DeepSeek-V3.2	0.719	0.842	$+0.123[.112, .135]$	0.828

Table 2: Common subset (345 firm-years shared by all six models): unpinned and pinned correct@1%, pinned–unpinned gain with 95% firm-level bootstrap CI (2,000 resamples, seed fixed), and attribution rate.

composites are computable *if the model can follow the formula and the named inputs*, which is itself a capability threshold. Because pinning is compound (§3.3), these gains bound the definitional component from above; the any-standard oracle shows how much of it is convention proper (gearing: strict 0.54 $\rightarrow$ any-standard 0.66 over 7,128 claims).

5.3 The ranking holds on the common subset

Table 2 re-scores all six models on the 345 firm-years they share, with firm-level bootstrap intervals. Every number moves by less than one point relative to the full sample (Table 4); the capability ordering, the sign and size of every pinning gain, and Gemini’s attribution collapse are unchanged. The four-model versus two-model sample difference therefore does not drive any conclusion.

5.4 Why wrong numbers are wrong

The error-mode classifier attributes each incorrect claim. For composites the dominant mode is the unexplained residual (gearing $\approx 85\%$ residual, $\approx 10\%$ valid gross-debt/equity convention; quick ratio $\approx 77\%$ residual, $\approx 16\%$ reporting the current ratio instead; ROCE 70% residual, 4% ROE reported instead); margins are dominated by a

percent $\leftrightarrow$ decimal scale artifact ($\approx 12\text{--}20\% \times 100$). This is sharper than “LLMs use different definitions”: after crediting valid conventions via the any-standard oracle, most residual error on composites has no definitional explanation—and it shrinks with capability. Three cases from the frozen cache illustrate the modes. *Different-concept*: Llama, unpinned, reports quick ratio = 1.083 for a firm-year whose oracle quick ratio is 0.188 but whose current ratio is 1.077—the model computed the sibling and cited `AssetsCurrent`. *Scale*: Llama, pinned, reports operating margin = 0.0091 against an oracle 0.0908—off by exactly one power of ten after normalization (the unit field was blank, so magnitude-based normalization left the value alone and the classifier caught the $\times 10$ gap). *Unexplained residual*: Qwen, unpinned, reports ROCE = -0.121 for a firm-year whose ROCE is $+0.327$ and whose ROE is -0.489 , citing `NetIncomeLoss`; the sign and the cited fact suggest net income was used where EBIT belongs, but the value matches neither ROE nor any standard convention, so the classifier correctly declines to name a cause.

5.5 Cascade propagation and CoT vs. PoT

Propagation. Over all wrong descendants with a claimed ancestor, the propagation score is 0.15 unpinned (570 explained / 3,184 local) and 0.32 pinned (1,124 / 2,365): most wrong composites are *local* errors, not inherited from a wrong intermediate. Per model (Table 3) the picture is capability-dependent in an informative way. Strong models rarely propagate (GPT-5-mini 0.03/0.02, Grok 0.01/0.03 unpinned/pinned): when they get a descendant wrong, the ancestor was usually right, so the error is introduced at the last step. Weak and middle models propagate more (Llama 0.18/0.32, Qwen 0.16/0.38), and pinning *raises* their propagation score—once the formula is fixed, a wrong ancestor is carried faithfully into the descendant. DeepSeek pinned is the outlier (0.52, 78/149): a good executor of a supplied formula that inherits its own EBITDA-proxy errors. The metric therefore separates “cannot execute the last step” from “executes correctly from a wrong input”, which a per-concept rate alone cannot.

CoT vs. PoT. Program-of-thought prompting does not improve fidelity for any model: the lift is within ± 1.7 points everywhere (pooled: CoT 0.721, PoT 0.715). It does change *attribution* for one model: GPT-5-mini cites a valid source fact 91.9% of the time under CoT but 73.0% under PoT (-18.9

Model	Propagation		PoT – CoT	
	unpin.	pin.	fidelity	attrib.
Llama-3.1-8B	0.18	0.32	-0.001	-0.024
Qwen2.5-72B	0.16	0.38	$+0.005$	-0.004
Gemini-3.1-F-L	0.36	0.14	-0.002	$+0.020$
GPT-5-mini	0.03	0.02	$+0.017$	-0.189
Grok-4.3	0.01	0.03	-0.008	-0.011
DeepSeek-V3.2	0.04	0.52	-0.002	-0.004

Table 3: Left: cascade propagation score (explained/(explained+local)) per model and condition, each on its full sample. Right: program-of-thought minus chain-of-thought, for correct@1% and attribution.

Model	<i>n</i>	attrib.	pin.	Δ [95% CI]
Llama-3.1-8B	729	0.687	0.293	$-0.028[-.040, -.016]$
Qwen2.5-72B	729	0.715	0.696	$+0.081[.071, .091]$
Gemini-3.1-F-L	729	0.057	0.878	$+0.086[.078, .095]$
GPT-5-mini	729	0.824	0.971	$+0.164[.157, .171]$
Grok-4.3	345	0.929	0.951	$+0.079[.072, .087]$
DeepSeek-V3.2	345	0.828	0.842	$+0.123[.113, .134]$

Table 4: Full sample per model (*n* firm-years): attribution rate, pinned correct@1%, and pinned–unpinned gain Δ with 95% firm-level bootstrap CIs (2,000 resamples). Every gain CI excludes zero; Llama’s is significantly negative.

points), i.e. writing the computation as code makes the strongest reasoner *less* likely to name the facts it used. Since PoT is often recommended for numeric tasks, this is a practical caveat: on this task it buys no fidelity and can cost auditability.

5.6 Attribution and fidelity are distinct and can decouple

Whether a model *cites its source facts* is a separate axis from whether it computes correctly (Fig. 3, Tables 4 and 2). Five models pair strong fidelity with strong attribution (Grok 0.93, GPT-5-mini 0.82, DeepSeek 0.83, Llama 0.69, Qwen 0.72). Gemini is the outlier: it computes well (most concepts >0.84) yet cites almost nothing (attribution 0.057) and has the worst schema-adherence (163 parse-failures vs. $\leq 0.4\%$ elsewhere)—a “right answer, no receipts” model whose outputs are unusable under a traceability or regulatory requirement (European Union, 2024). GPT-5-mini, restored from a truncation failure by a reasoning-budget fix, becomes the strongest pinned reasoner (pinned fidelity 0.97) with high attribution. We stress that our attribution metric is partial (one cited input in lineage); it is sufficient to expose Gemini’s collapse, but it overstates the auditability of the others relative to a full-lineage standard.

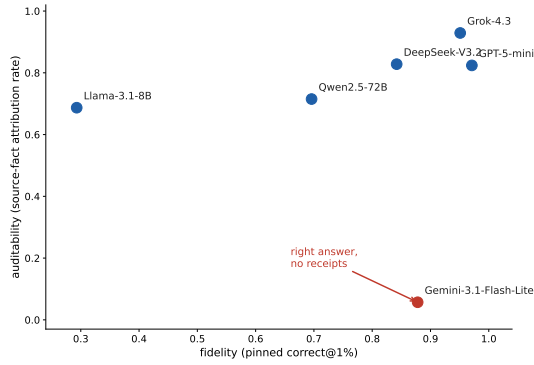


Figure 3: Attribution (source-fact citation) vs. fidelity. Gemini computes well but cites nothing.

6 Discussion

The oracle is the instrument: by recomputing the cascade deterministically, it converts an opaque “wrong number” into an attributable cause, and it exposes four structure-level facts a single hallucination rate would hide. First, fidelity is capability-dependent: simple ratios are universal, but multi-input composites—and ROCE above all—are computed reliably only by capable models, so an aggregate “accuracy” conflates concept difficulty with model strength. Second, disclosing the definition and the inputs helps whoever can execute them: pinning inverts the apparent “computational limit” seen on weak models, which cautions against generalizing prompt-intervention effects across a narrow model pool. Third, errors are mostly local, and code does not help: wrong composites are rarely inherited from a wrong intermediate, and program-of-thought neither improves fidelity nor preserves attribution. Fourth, attribution is a distinct axis: a model can be accurate yet uncitable (Gemini), which matters precisely where LLM financial outputs would be deployed—under traceability or regulatory obligations. Practically: pin definitions *and* name the inputs, prefer capable models for multi-step aggregates, verify ROCE-class concepts regardless, evaluate source-fact citation separately from accuracy, and do not assume code-style prompting is free.

7 Conclusion

LLM financial-fidelity failures are not monolithic. Using a deterministic XBRL cascade oracle over six models, on a full sample and a common subset, we show that composite fidelity is capability-dependent (ROCE the residual hard concept), that pinning the definition and inputs helps every capa-

ble model—inverting an apparent computational limit that was a weak-model artifact—that wrong composites are mostly local rather than propagated errors and that program-of-thought buys no fidelity, and that source-fact attribution is distinct from accuracy and can decouple, with one model computing correctly while citing nothing.

Limitations

Scope is US-GAAP XBRL filers; continental SIG aggregates use documented proxies ($\text{EBITDA} \approx \text{EBE}$), and financials/utilities are excluded by design. The oracle fixes conventions (current debt $\rightarrow 0$ when untagged) and selects tags by a first-match rule; determinism does not guarantee that every reference value is the accounting ground truth an analyst would choose. The any-standard oracle mitigates but does not eliminate convention dependence, and we have not reconciled the oracle against published ratios or an accounting-expert audit of a random firm-year subset; that audit is the most valuable next step and we release the lineage needed to run it. The pinned/unpinned contrast is a compound intervention (formula + tags + conventions) and is read as an upper bound on the definitional component; formula-only and tags-only arms were not run. The “unexplained residual” error mode is a residual category and may contain wrong-fact, wrong-period or omitted-input errors alongside genuine arithmetic slips. The attribution metric is partial (one cited input in lineage), not full-lineage coverage. Two models (DeepSeek, Grok) have provider-reported late-2025 cutoffs that are not independently verifiable (with conflicting secondary sources); for them the leakage gate is enforced on filing dates with a three-day margin (earliest retained filing 4 December 2025), so we treat their results as leakage-controlled only to the resolution of a self-reported cutoff, and every conclusion is also verified on the four cleanly-gated models and on the common subset. The panel is six models with one prompt family; “capability-dependent” describes this panel, not a general law. Values are self-reported; we score outputs and treat refusals and parse-failures as distinct outcomes, never imputed. Serverless models drift; we archive the frozen response cache so results remain exactly reproducible. Bootstrap 95% confidence intervals over firm-years are reported for the pinning gain (Tables 4 and 2); per-concept and attribution-vs-fidelity CIs are released with the artifacts.

Acknowledgements

We thank the three anonymous FinNLP reviewers, whose requests for a common-subset comparison, the propagation and CoT/PoT results, and a more careful reading of the pinning contrast shaped this version. This work is part of the AI-reliability research programme of Bulb’s Head SAS.

Reproducibility statement

Code, frozen model responses, run manifests with content hashes, pinned dataset checksums, quality-gate logs and the camera-ready analysis script are archived at [10.5281/zenodo.21112768](https://doi.org/10.5281/zenodo.21112768) (concept DOI, always-latest version). Third-party datasets are not redistributed (download scripts + checksums only). One command rebuilds the result tables bit-identically from the frozen cache; the common-subset, propagation and CoT/PoT tables in this version are produced by the same evaluator on the same cache. This satisfies the FinNLP open-research policy (released prompts, inference configurations, generated outputs, and evaluation artifacts).

References

- Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research (TMLR)*. ArXiv:2211.12588.
- 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. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of EMNLP*. ArXiv:2109.00122.
- Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Con-vFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of EMNLP*. ArXiv:2210.03849.
- European Union. 2024. Regulation (EU) 2024/1689 (Artificial Intelligence Act). Official Journal of the European Union. Annex III: credit scoring classified as high-risk.
- Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided language models. In *Proceedings of ICML*. ArXiv:2211.10435.
- Dongxin Guo, Jikun Wu, and Siu Ming Yiu. 2026. Fin-Ground: Detecting and grounding financial hallucinations via atomic claim verification. *arXiv preprint arXiv:2604.23588*.
- Shijie Han et al. 2024. **XBRL agent: Leveraging large language models for financial report analysis**. In *Proceedings of the 5th ACM International Conference on AI in Finance (ICAIF)*.
- Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. FinanceBench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12).
- Alex G. Kim, Maximilian Muhn, and Valeri V. Nikolaev. 2024. Financial statement analysis with large language models. *arXiv preprint arXiv:2407.17866*.
- Haohang Li et al. 2026. AuditFlow: Executable symbolic environments for structured financial reporting verification. *arXiv preprint arXiv:2606.03031*.
- Yitao Long, Tiansheng Hu, Yilun Zhao, Arman Cohan, and Chen Zhao. 2025. FinLFQA: Evaluating attributed text generation of LLMs in financial long-form question answering. *arXiv preprint arXiv:2510.06426*.
- Stephen H. Penman. 2013. *Financial Statement Analysis and Security Valuation*, 5 edition. McGraw-Hill.
- Varshini Reddy et al. 2024. DocFinQA: A long-context financial reasoning dataset. *arXiv preprint arXiv:2401.06915*.
- G. Bennett Stewart. 1991. *The Quest for Value: A Guide for Senior Managers*. HarperBusiness.
- Zichen Tang, Haihong E, Ziyang Ma, Haoyang He, Jiacheng Liu, et al. 2025. FinanceReasoning: Benchmarking financial numerical reasoning more credible, comprehensive and challenging. *arXiv preprint arXiv:2506.05828*.
- Yan Wang, Lingfei Qian, Xueqing Peng, Yang Ren, Keyi Wang, et al. 2025. FinTagging: An LLM-ready benchmark for extracting and structuring financial information. *arXiv preprint arXiv:2505.20650*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*. ArXiv:2201.11903.
- Qianqian Xie et al. 2024. FinBen: A holistic financial benchmark for large language models. In *NeurIPS Datasets and Benchmarks Track*. ArXiv:2402.12659.

- Hongyang Yang et al. 2024. FinRobot: An open-source AI agent platform for financial applications using large language models. *arXiv preprint arXiv:2405.14767*.
- Mengao Zhang, Jiayu Fu, Tanya Warrier, Yuwen Wang, Tianhui Tan, and Ke-wei Huang. 2025. FAITH: A framework for assessing intrinsic tabular hallucinations in finance. *arXiv preprint arXiv:2508.05201*.
- Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. In *Proceedings of ACL*.
- Yilun Zhao et al. 2024. FinDVer: Explainable claim verification over long and hybrid-content financial documents. In *Proceedings of EMNLP*. ArXiv:2411.05764.
- Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of ACL*.