FinGround: Detecting and Grounding Financial Hallucinations via Atomic Claim Verification

Dongxin Guo[1], **Jikun Wu**[2], **Siu Ming Yiu**[1]
[1]The University of Hong Kong [2]Stellaris AI Limited
bettyguo@connect.hku.hk, hk950014@connect.hku.hk,
smyiu@cs.hku.hk

Abstract

Financial AI systems must produce answers grounded in specific regulatory filings, yet current LLMs fabricate metrics, invent citations, and miscalculate derived quantities. These errors carry direct regulatory consequences as the EU AI Act's high-risk enforcement deadline approaches (August 2026). Existing hallucination detectors treat all claims uniformly, missing 43% of computational errors that require arithmetic re-verification against structured tables. We present FINGROUND, a three-stage verify-then-ground pipeline for financial document QA. Stage 1 performs finance-aware hybrid retrieval over text and tables. Stage 2 decomposes answers into atomic claims classified by a six-type financial taxonomy and verified with type-routed strategies including formula reconstruction. Stage 3 rewrites unsupported claims with paragraph- and table-cell-level citations. To cleanly isolate verification value from retrieval quality, we propose *retrieval-equalized evaluation* as standard methodology for RAG verification research: when all systems receive identical retrieval, FINGROUND still reduces hallucination rates by 68% over the strongest baseline ($p < 0.01$). The full pipeline achieves a 78% reduction relative to GPT-4o. An 8B distilled detector retains 91.4% F1 at 18× lower per-claim latency, enabling $0.003/query deployment, supported by qualitative signals from a four-week analyst pilot.

1 Introduction

Financial professionals require answers grounded in specific regulatory filings and earnings reports, yet LLMs routinely fabricate financial metrics, invent regulatory citations, and distort ratios. GPT-4-Turbo with retrieval incorrectly answered or refused 81% of curated SEC filing questions (Islam et al., 2023), with systematic fabrication of financial metrics documented across models (Kang and Liu, 2023; Gartner, Inc., 2024; Accenture, 2024). The EU AI Act (European Parliament and Council of the European Union, 2024) mandates compliance for high-risk financial AI systems by August 2026 (Dahl et al., 2024), requiring human oversight with interpretable outputs (Article 14) and accuracy guarantees (Article 15).

Prior work has advanced individual components: general hallucination detection (Manakul et al., 2023; Farquhar et al., 2024; Min et al., 2023; Wei et al., 2024), financial QA benchmarks (Chen et al., 2021; Zhu et al., 2021; Islam et al., 2023), and RAG systems (Asai et al., 2024; Yan et al., 2024). Yet no existing system unifies detection and mitigation into a production-ready pipeline for financial QA. The financial domain breaks assumptions of prior systems in specific, quantifiable ways. FActScore and SAFE decompose claims into atomic facts but treat all facts uniformly, so they cannot verify "gross margin was 62.4%" against table cells without structured extraction; this gap accounts for 43% of computational errors that a domain-specific pipeline catches (Appendix Q). RARR's regeneration assumes single-source evidence, but when applied without claim-type-aware routing, 34% of computational claim regenerations produced *new* hallucinations. And table-cell attribution without structure-aware upstream chunking produced 23% dangling citations pointing to misaligned cells.

We present FINGROUND, addressing this gap through three contributions:

1. **Finance-Aware Atomic Verification**: decomposition of LLM answers into atomic claims verified against evidence using a validated six-type taxonomy (*numerical*, *temporal*, *entity-attribute*, *comparative*, *regulatory*, *computational*) with type-routed verification strategies including arithmetic re-computation (§3.2).
2. **Grounded Regeneration with Evidence Attribution**: targeted rewriting of only hallucinated spans with paragraph- and table-cell-level citations, achieving 93.2% faithfulness on regenerated claims (§3.3).

3. **Efficient Distilled Detector**: distillation from GPT-4o into an 8B model achieving 91.4% F1 at 18× lower per-claim latency, enabling $0.003/query deployment (§3.4).

Beyond system contributions, we introduce *retrieval-equalized evaluation*, in which all baselines receive identical retrieval, isolating verification value from retrieval gains. Under these conditions, atomic verification yields 68–76% additional HalRate reduction ($p < 0.01$). Cross-generator evaluation on Llama-3-70B and Claude-3.5-Sonnet shows 87–89% F1 transfer, and a four-week pilot with 24 analysts provides deployment design signals targeting EU AI Act compliance.

2 Related Work

Hallucination detection spans sampling-based methods (Manakul et al., 2023; Farquhar et al., 2024), atomic evaluation (Min et al., 2023; Wei et al., 2024; Wang et al., 2024), and industrial classifiers (Ravi et al., 2024; Friel and Sanyal, 2023; Microsoft, 2024), with comprehensive taxonomies established by surveys (Ji et al., 2023; Huang et al., 2025; Tonmoy et al., 2024). These approaches operate domain-agnostically without financial-specific claim typing or table-cell attribution.

Financial QA demands multi-step numerical reasoning over hybrid text-and-table content, with benchmarks from FinQA (Chen et al., 2021) and TAT-QA (Zhu et al., 2021) through FinanceBench (Islam et al., 2023) and DocFinQA (Reddy et al., 2024) establishing evaluation standards. RAG variants such as Self-RAG (Asai et al., 2024), CRAG (Yan et al., 2024), and Adaptive-RAG (Jeong et al., 2024), together with attribution methods (Gao et al., 2023a,b; Bohnet et al., 2022), provide grounding infrastructure, while domain models (Wu et al., 2023; Yang et al., 2023) reduce but do not eliminate hallucination. Financial hallucination benchmarks PHANTOM (Ji et al., 2025) and FAITH (Zhang et al., 2025) document severity but do not offer integrated solutions. None of these systems combine atomic claim verification with financial table-cell attribution and hallucination-triggered regeneration.

Industrial financial AI platforms (Bloomberg, Kensho, AlphaSense, FactSet) excel at information retrieval and extraction but do not perform claim-level verification: they cannot determine whether a specific LLM-generated assertion is supported by source evidence. FINGROUND operates as a verification layer downstream of any financial QA system. Knowledge distillation (Hinton et al., 2015; Gu et al., 2024) and demonstrated success transferring hallucination detection to 8B scale (Song et al., 2025) enable the efficient deployment our system requires. A detailed capability comparison is in Appendix B.

3 The FINGROUND System

FINGROUND operates as a three-stage pipeline (Figure 1): (1) finance-aware retrieval extracts evidence from hybrid text-and-table documents; (2) atomic verification decomposes answers into claims and detects hallucinations; (3) grounded regeneration rewrites unsupported claims with cited evidence.

3.1 Stage 1: Finance-Aware Hybrid Retrieval

Financial answers frequently require synthesizing information from narrative text and structured tables. Adapting query-complexity-driven routing (Jeong et al., 2024), FINGROUND classifies queries into three tiers using a RoBERTa-base classifier (89.3% accuracy; details in Appendix P): **Simple** queries use single-passage BM25 retrieval; **Moderate** queries combine dense retrieval (E5-large fine-tuned on financial passage pairs; Appendix O) with table extraction via a column-header-aware similarity function $\text{sim}(q,t) = \alpha \cdot \cos(\mathbf{q}, \mathbf{t}_{\text{cell}}) + (1-\alpha) \cdot \cos(\mathbf{q}, \mathbf{t}_{\text{header}})$ (α=0.6); **Complex** queries use an iterative retrieval-then-reason loop. Structure-aware chunking preserves row-column relationships with header metadata, enabling table-cell-level attribution downstream. Each chunk carries a provenance tuple $\langle$document, section, page, element_type$\rangle$.

3.2 Stage 2: Atomic Financial Claim Verification

Given a generated answer a and retrieved evidence $E = \{e_1, \ldots, e_k\}$, verification operates in three steps.

Claim Decomposition. Following FActScore (Min et al., 2023) but adapted for financial language, we decompose a into atomic claims $C = \{c_1, \ldots, c_n\}$, each classified into one of six categories: *numerical* (specific values), *temporal* (time-bound assertions), *entity-attribute* (entity properties), *comparative* (cross-entity/period comparisons), *regulatory* (compliance references), and *computational* (derived quantities requiring arithmetic). This taxonomy extends general halluci-

nation classification (Ji et al., 2023) with finance-specific categories motivated by error analysis on 500 real financial hallucinations (Kang and Liu, 2023; Ji et al., 2025). Validation confirming appropriate granularity (6-type outperforms 3-type by 4.3 F1; 10-type shows no significant gain, p=0.23) is in Appendix A.

Claim–Evidence Alignment. Each claim is aligned to evidence using a cross-encoder fine-tuned on 8,400 financial NLI examples from TAT-QA and FinQA, achieving 87.2% alignment F1 (Appendix H). For numerical claims, structured extraction identifies the specific value, unit, time period, and entity for exact matching against table cells.

Verdict Classification. A distilled 8B classifier (§3.4) assigns each claim a verdict: *supported* (entailed by evidence), *contradicted* (conflicts with evidence), or *unverifiable* (no relevant evidence). For *computational* claims, standard NLI is insufficient; FINGROUND employs formula reconstruction: (1) identifying the implied formula using a library of 47 financial formula templates, (2) retrieving operand values from table cells, and (3) recomputing the derived quantity with $\pm$0.5% tolerance for rounding conventions. End-to-end computational verification achieves 90.2% F1.

Retrieval Failure Handling. When alignment yields no candidate evidence above the cross-encoder threshold, the claim is assigned the *unverifiable* verdict and routed to Stage 3 for targeted re-retrieval and regeneration (§3.3). The analyst pilot (§6) provides empirical calibration: the false negative rate was 3.8% across 1,847 queries, with 56% of the 27 missed hallucinations involving computational claims whose operand evidence fell outside the retrieved window. We treat retrieval-induced unverifiability as a recoverable failure mode routed downstream rather than a silent miss; the alternative of defaulting unverifiable claims to *supported* would conflate evidence absence with evidence consistency, which the regulatory use case forbids.

3.3 Stage 3: Grounded Regeneration

For each claim classified as contradicted or unverifiable, the regeneration module locates the corresponding span in the original answer (fuzzy alignment, edit distance $\leq$3 tokens), performs targeted re-retrieval if needed, generates a grounded replacement following RARR's research-and-revise paradigm (Gao et al., 2023a), and attaches inline citations in the format `[Doc:`d`, §`s`, p.`p`]` or `[Doc:`d`, Table `t`, Row `r`, Col `c`]`. When conflicting information exists (*e.g.*, restated figures), the module defaults to the most recent filing date and flags the conflict. For answers requiring $\geq$3 claim regenerations, FINGROUND triggers full re-generation rather than incremental repair to mitigate error compounding. A flag-only mode that removes rather than repairs hallucinated claims is available for high-stakes regulatory contexts (the "−regen." ablation in Table 2).

3.4 Distillation for Production Deployment

GPT-4o (`gpt-4o-2024-05-13`) annotates 3,200 financial QA examples spanning FinQA, TAT-QA, and SEC filings between June and August 2025, with a two-pass consistency check discarding 8.4% of disagreements (Bohnet et al., 2022) (full annotation prompt in Appendix S). We fine-tune Llama-3-8B-Instruct using reverse KL divergence (Gu et al., 2024) with a multi-task objective combining claim decomposition, evidence alignment, and verdict classification (details in Appendix N). Served via vLLM with continuous batching, the distilled model achieves p95 latency of 340ms per claim on A100, an 18$\times$ per-claim improvement over GPT-4o (6.1s/claim) with 91.4% detection F1. Full-pipeline latency is 3.8s p95 per query, a 2.2$\times$ improvement over the 8.2s teacher pipeline.

4 Experimental Setup

Datasets. We evaluate on three benchmarks: **FinQA** (Chen et al., 2021) (8,281 QA pairs requiring multi-step numerical reasoning over S&P 500 earnings reports), **TAT-QA** (Zhu et al., 2021) (16,552 questions over hybrid tabular-textual financial reports), and **FinanceBench** (Islam et al., 2023) (150 curated questions from real SEC filings; small sample yields wide CIs). For claim-level evaluation, we construct **FinHalu**: 1,200 (question, answer, evidence) triples with GPT-4o/GPT-3.5-Turbo-generated answers annotated at the claim level by three financial domain experts ($\kappa = 0.83$). Training and evaluation data are strictly disjoint at both question and document level (Appendix C).

Metrics. **HalRate** is the percentage of generated claims classified as contradicted or unverifiable: HalRate $= \sum h_i / \sum c_i \times 100\%$, where h_i and c_i are hallucinated and total claims per an-

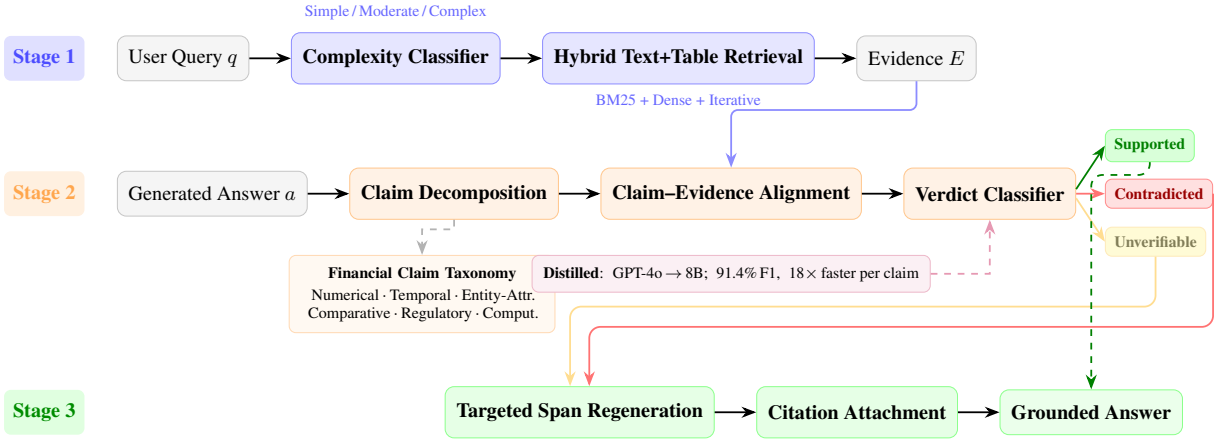

Figure 1: The FINGROUND pipeline. **Stage 1** classifies query complexity and retrieves hybrid text-and-table evidence. **Stage 2** decomposes answers into atomic financial claims, aligns each to evidence, and classifies verdicts using a distilled 8B model. **Stage 3** rewrites contradicted/unverifiable claims with cited evidence; supported claims pass through unchanged.

swer. All systems' outputs are decomposed using the same GPT-4o prompt to normalize granularity. **Det. F1** is standard precision–recall F1 over claim-level binary verdicts. **CitP/CitR** are citation precision and recall. All results include 95% bootstrap CIs (B=10,000) with paired permutation tests (Efron and Tibshirani, 1993; Berg-Kirkpatrick et al., 2012). We introduce **retrieval-equalized evaluation**: all systems receive identical retrieval, isolating verification contribution from retrieval quality.

Baselines. We compare against Vanilla RAG (BM25 + GPT-4o), Self-RAG (Asai et al., 2024), CRAG (Yan et al., 2024), SelfCheckGPT (Manakul et al., 2023), HHEM (Vectara), GPT-4o + CoT, and FActScore (Min et al., 2023) with retrieval-equalized configuration (Appendix M; FActScore-specific configuration in Appendix R). All baselines use official codebases; domain-adapted variants (§5.3) isolate architectural from data contributions.

5 Results and Analysis

Computational claims show the highest hallucination rate (28.4%) despite being the most amenable to automated verification when properly typed, so the bottleneck is routing rather than verification difficulty. Hedged financial language ("approximately," "roughly") drives 52% of false positives. Generic NLI fails on ratio and margin verification even when the correct evidence passage is retrieved, because the failure is in reasoning rather than retrieval.

System	Prec.	Rec.	F1
SelfCheckGPT	$69.4_{\pm 2.1}$	$76.5_{\pm 1.8}$	$72.8_{\pm 1.6}$
HHEM	$78.9_{\pm 1.8}$	$73.8_{\pm 2.0}$	$76.3_{\pm 1.5}$
FActScore	$74.2_{\pm 2.0}$	$79.3_{\pm 1.7}$	$76.7_{\pm 1.5}$
Self-RAG	$81.2_{\pm 1.7}$	$77.1_{\pm 1.9}$	$79.1_{\pm 1.4}$
CRAG	$80.6_{\pm 1.9}$	$74.9_{\pm 2.1}$	$77.6_{\pm 1.6}$
GPT-4o (teacher)	$94.1_{\pm 0.9}$	$95.9_{\pm 0.7}$	$95.0_{\pm 0.6}$
FINGROUND (8B distilled)	$\mathbf{92.7}_{\pm 1.1}$	$\mathbf{90.2}_{\pm 1.3}$	$\mathbf{91.4}_{\pm 1.2}$

Table 1: Hallucination detection on FinHalu (±95% CI). Best non-teacher results in **bold**. FINGROUND retains 96.2% of teacher F1 at 18× lower per-claim latency. All improvements significant at $p < 0.01$.

5.1 Detection Performance

Table 1 presents hallucination detection on FinHalu. FINGROUND's distilled 8B detector achieves 91.4% $F1_{\pm 1.2}$, retaining 96.2% of the GPT-4o teacher's performance at 18× lower per-claim latency. All improvements over baselines are significant ($p < 0.01$). Performance varies by claim type: entity-attribute highest (95.6% F1), numerical lowest (88.1%), with computational claims benefiting most from formula reconstruction (+18.9 F1 over SelfCheckGPT; breakdown in Appendix D).

5.2 End-to-End Results

Table 2 reports end-to-end results. FINGROUND achieves the lowest hallucination rate across all benchmarks (4.1% avg.), a 78% relative reduction from GPT-4o + CoT ($p < 0.01$). Ablating the claim taxonomy roughly doubles HalRates (largest impact on FinanceBench: 4.9%→11.7%), confirming domain-specific decomposition is essential. Removing table retrieval disproportionately affects

TAT-QA (3.8%→10.6%). Removing regeneration maintains detection but reduces unconditional accuracy by 7.4 points (Table 2).

Retrieval-Equalized Comparison. To isolate verification value from retrieval improvements, we equip baselines with FINGROUND's Stage 1 retrieval. Finance-aware retrieval improves all baselines by 37–39%, and on top of this FINGROUND adds a further 68–76% HalRate reduction under controlled conditions ($p < 0.01$). Atomic verification therefore contributes value independently of retrieval quality (Table 3).

5.3 Isolating Architectural from Data Contribution

To test whether FINGROUND's advantage stems from architecture or domain data access, we provide baselines with comparable domain adaptation: HHEM calibrated on 500 financial examples, SelfCheckGPT fine-tuned on 1,000 financial examples, both receiving FINGROUND's retrieval (retrieval-equalized).

Domain adaptation improves both baselines by 5–7 F1 points, but FINGROUND maintains a ∼10–12 F1-point lead ($p < 0.01$), confirming that claim-type-aware verification and cross-encoder alignment provide value beyond domain data alone. The residual gap is largest on computational claims (lacking arithmetic re-verification) and table-dependent claims (lacking structure-aware alignment). Discussion of data quantity asymmetry is in Appendix E.

5.4 Robustness, Quality, and Efficiency

Human Validation. Independent annotation by three financial experts on 200 FinHalu examples (without seeing GPT-4o labels) yields $\kappa = 0.87$ agreement with GPT-4o. The distilled model achieves $90.8\%_{\pm 2.1}$ F1 against human labels (vs. $91.4\%_{\pm 1.2}$ against GPT-4o), confirming no circular evaluation inflation (Appendix F).

Regeneration Quality. On 300 evaluated regenerated claims: 93.2% faithfulness, 4.1% error introduction rate, 96.7% fluency, +71.3% net improvement. Errors concentrate on multi-step table reasoning. The 4.1% per-claim rate compounds in multi-claim answers; full re-generation for ≥3-claim answers mitigates this (Appendix G).

Cross-Generator Transfer. The distilled detector achieves $88.0\%_{\pm 2.1}$ F1 on Llama-3-70B outputs and $87.6\%_{\pm 2.2}$ on Claude-3.5-Sonnet (3–4 point degradation from in-distribution). Degradation is smallest on computational claims (−1.8 F1, generator-agnostic arithmetic) and largest on comparative claims (−5.2 F1). For non-GPT backends, generator-specific calibration data is recommended (Appendix I).

Cross-Benchmark Generalization. Without task-specific adaptation, FINGROUND achieves HalRates of $5.1\%_{\pm 1.3}$ on ConvFinQA (Chen et al., 2022) and $6.8\%_{\pm 1.5}$ on MultiHiertt (Zhao et al., 2022), comparable to its in-distribution range of 3.6–4.9% across FinQA, TAT-QA, and FinanceBench. The five-benchmark span shows the verification architecture transfers across financial QA distributions without dataset-specific tuning (Appendix V).

Cost and Efficiency. Table 5 summarizes production economics. FINGROUND reduces per-query cost by 15.7× vs. GPT-4o ($0.003 vs. $0.047). Memory footprint is 18GB (FP16), fitting on a single A10G. Under concurrent load (32 requests), throughput reaches 8.4 queries/second on A100. Stage 2 (verification) accounts for 55% of latency; batched claim inference reduces it from 4.8s sequential to 2.1s p95 (Appendix K).

Error Analysis. False negatives (8.6%) concentrate on paraphrased numerical errors near decision boundaries (recall drops to 71.4% on values within ±5% of ground truth). False positives (6.1%) concentrate on hedged language (52% of FP) and restated figures (31% of FP). Full error and near-miss adversarial analysis is in Appendix J.

6 Deployment and Analyst Feedback

We conducted a four-week feasibility pilot with 24 financial analysts (equity research and compliance) at a financial services firm processing SEC 10-K and 10-Q filings, covering 1,847 queries across 43 filings. The observed false positive rate was 6.1% (analyst corrections) and false negative rate was 3.8% (27 analyst-flagged misses). In a post-pilot Likert survey, 20/24 analysts rated FINGROUND at ≥4/5 for practical acceptability (mean: 4.1, SD: 0.7). Alert fatigue was stable across weeks (override rate 5.8–6.3%, p=0.71 for trend). We report these as qualitative design signals from an uncontrolled pilot, not precise population estimates.

Three design insights emerged: (1) table-cell-level citations (*e.g.*, "Table 3, Row: Operating

System	FinQA		TAT-QA		FinanceBench		Accuracy	
	HalRate↓	Acc↑	HalRate↓	Acc↑	HalRate↓	Acc↑	Uncond.	Supp.†
Vanilla RAG	$34.7_{\pm 1.8}$	$68.2_{\pm 1.4}$	$31.5_{\pm 1.5}$	$71.4_{\pm 1.2}$	$43.8_{\pm 4.1}$	$52.1_{\pm 4.0}$	63.9	–
FActScore	$25.3_{\pm 1.7}$	$69.8_{\pm 1.3}$	$22.7_{\pm 1.4}$	$72.1_{\pm 1.2}$	$32.4_{\pm 3.8}$	$56.7_{\pm 3.9}$	66.2	89.3
Self-RAG	$22.1_{\pm 1.6}$	$70.8_{\pm 1.3}$	$18.4_{\pm 1.3}$	$74.6_{\pm 1.1}$	$28.5_{\pm 3.7}$	$59.3_{\pm 3.9}$	68.2	–
CRAG	$23.8_{\pm 1.7}$	$69.5_{\pm 1.4}$	$20.1_{\pm 1.4}$	$73.8_{\pm 1.1}$	$30.2_{\pm 3.8}$	$57.8_{\pm 4.0}$	67.0	–
GPT-4o + CoT	$18.6_{\pm 1.4}$	$73.4_{\pm 1.2}$	$15.2_{\pm 1.2}$	$77.1_{\pm 1.0}$	$22.4_{\pm 3.4}$	$65.2_{\pm 3.8}$	71.9	–
FinGround (full)	$\mathbf{3.6}_{\pm 0.7}$	$75.1_{\pm 1.2}$	$\mathbf{3.8}_{\pm 0.6}$	$78.3_{\pm 1.0}$	$\mathbf{4.9}_{\pm 1.8}$	$67.9_{\pm 3.7}$	71.2	94.7
– regen.	$3.6_{\pm 0.7}$	$70.2_{\pm 1.3}$	$3.8_{\pm 0.6}$	$72.7_{\pm 1.1}$	$4.9_{\pm 1.8}$	$58.4_{\pm 3.9}$	63.8	94.7
– taxonomy	$7.2_{\pm 1.0}$	$74.8_{\pm 1.2}$	$8.1_{\pm 0.9}$	$77.9_{\pm 1.0}$	$11.7_{\pm 2.6}$	$66.3_{\pm 3.8}$	70.5	92.1
– table ret.	$5.9_{\pm 0.9}$	$72.4_{\pm 1.3}$	$10.6_{\pm 1.1}$	$71.8_{\pm 1.1}$	$9.4_{\pm 2.4}$	$61.2_{\pm 3.9}$	66.2	91.8

Table 2: End-to-end results (±95% bootstrap CI, B=10,000). HalRate: % hallucinated claims (lower is better). †Supp.: accuracy on claims classified as supported. All FinGround vs. GPT-4o+CoT differences significant at $p < 0.01$.

	Original Ret.		FinGround Ret.	
System	HalRate	Acc	HalRate	Acc
GPT-4o + CoT	18.6	73.4	11.3	75.8
Self-RAG	22.1	70.8	13.7	73.5
CRAG	23.8	69.5	14.9	72.1
FinGround (full)	**3.6**	**75.1**	(same)	

Table 3: Retrieval-equalized comparison on FinQA. FinGround's retrieval improves all baselines (right), but atomic verification yields an additional 68–76% HalRate reduction ($p < 0.01$).

System	Det. F1	HalRate	Δ vs. FinGround
HHEM (out-of-box)	$76.3_{\pm 1.5}$	$21.4_{\pm 1.6}$	−17.8
HHEM-adapted	$81.7_{\pm 1.4}$	$15.2_{\pm 1.3}$	−11.6
SelfCheck (out-of-box)	$72.8_{\pm 1.6}$	$24.1_{\pm 1.7}$	−20.5
SelfCheck-adapted	$79.4_{\pm 1.5}$	$16.8_{\pm 1.4}$	−13.2
FinGround (full)	$\mathbf{91.4}_{\pm 1.2}$	$\mathbf{3.6}_{\pm 0.7}$	—

Table 4: Domain-adapted comparison on FinQA (retrieval-equalized, ±95% CI). Domain adaptation improves baselines by 5–7 F1, but a ∼10–12 F1-point architectural gap persists ($p < 0.01$).

System	p95 Lat.	$/query	Det. F1
GPT-4o (teacher)	8.2s	$0.047	95.0
FinGround (8B, A100)	3.8s	$0.003	91.4
FinGround (8B, A10G)	5.9s	$0.002	91.4
HHEM	2.1s	$0.001	76.3
SelfCheckGPT	12.4s	$0.062	72.8

Table 5: Cost and efficiency. p95 latency is full-pipeline per query. FinGround achieves near-teacher F1 at 15.7× lower cost.

Income, Col: FY2024") were strongly preferred over paragraph-level references, enabling verification in seconds rather than minutes; (2) contradiction explanations showing the specific conflict (*e.g.*, "Source says $4.2B but answer says $4.8B") were the most valued feature; (3) computational claim verification had the highest impact, as incorrect derived metrics are the most dangerous errors and hardest for humans to catch (56% of the 27 missed hallucinations involved computational claims).

Based on the pilot, FinGround is being integrated as a REST API service with a document ingestion pipeline (∼500 filings/day), the verification endpoint with inline citations, and a monitoring layer tracking confidence distributions and override patterns for drift detection. Production integration targets Q3 2026, ahead of the EU AI Act compliance deadline. Additional deployment details including scalability estimates, failure handling, and benchmark-to-production gap analysis are in Appendix L.

7 Conclusion

FinGround combines finance-aware retrieval, atomic claim verification with a validated domain-specific taxonomy, and grounded regeneration to achieve a 68% hallucination reduction under retrieval-equalized conditions ($p < 0.01$) and 78% with the full pipeline. The retrieval-equalized methodology isolates verification value and should become standard practice for RAG evaluation. Domain-adapted comparisons confirm that FinGround's architectural advantage persists beyond data access. Efficient distillation enables 91.4% F1 at 18× lower per-claim latency and $0.003/query, and a four-week analyst pilot provides deployment design signals targeting EU AI Act compliance.

Reproducibility. All materials are available at: https://github.com/bettyguo/FinGround.

Limitations

Our evaluation covers English-language U.S. SEC filings; generalization to non-English documents and other jurisdictions requires validation. The evaluation ecosystem is GPT-dependent: FinHalu uses GPT-4o/GPT-3.5 generations with 3–4 point F1 degradation on non-GPT generators. The distilled detector shows a 3.6-point F1 gap from the teacher, particularly on nuanced computational claims. Domain-adapted comparisons use asymmetric data quantities (3,200 for FINGROUND vs. 500–1,000 for baselines); the scaling-projection analysis at parity (Appendix E) places the residual architectural gap at 5.2–7.6 F1, and a controlled retraining at full parity would tighten this estimate. The pilot involved 24 analysts at a single firm with self-reported observations rather than controlled measurements. Detection recall drops to 71.4% on hallucinated values within $\pm 5\%$ of ground truth. Extended limitations in Appendix U.

Ethical Considerations

Even with verification, FINGROUND's outputs should not be treated as financial advice: a claim marked "supported" means consistency with the retrieved source, not correctness of the source itself. We are concerned about automation bias: if analysts develop high trust, they may reduce independent verification, creating failure modes where false negatives propagate. We recommend deployment with periodic spot-checks and tunable confidence thresholds. We acknowledge dual-use risk from per-type vulnerability analysis and mitigate by releasing only the detection model, not adversarial example methodology. In regulated contexts, liability for AI-verified claims remains unresolved; FINGROUND is a verification tool, not a compliance certification system.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback, which substantially improved this work. This research was supported by The University of Hong Kong and Stellaris AI Limited.

References

Accenture. 2024. Technology vision 2024: Human by design — how AI unleashes the next level of human potential. https://www.accenture.com/us-en/insights/technology/technology-trends-2024.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pages 995–1005. ACL.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint*, arXiv.2212.08037.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Kenneth Huang, Bryan R. Routledge, and William Yang Wang. 2021. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3697–3711. Association for Computational Linguistics.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 6279–6292. Association for Computational Linguistics.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *arXiv preprint*, arXiv.2401.01301.

Bradley Efron and Robert Tibshirani. 1993. *An Introduction to the Bootstrap*. Springer.

European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 june 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nat.*, 630(8017):625–630.

Robert Friel and Atindriyo Sanyal. 2023. Chainpoll: A high efficacy method for LLM hallucination detection. *arXiv preprint*, arXiv.2310.18344.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. RARR: researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16477–16508. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6465–6488. Association for Computational Linguistics.

Gartner, Inc. 2024. CMOs must protect consumer trust in the AI age. https://www.gartner.com/en/newsroom/press-releases/2024-04-02-cmos-must-protect-consumer-trust-in-the-ai-age.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint*, arXiv.1503.02531.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55.

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.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 7036–7050. Association for Computational Linguistics.

Lanlan Ji, Dominic Seyler, Gunkirat Kaur, Manjunath Hegde, Koustuv Dasgupta, and Bing Xiang. 2025. PHANTOM: A benchmark for hallucination detection in financial long-context QA. *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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 Comput. Surv.*, 55(12):248:1–248:38.

Haoqiang Kang and Xiao-Yang Liu. 2023. Deficiency of large language models in finance: An empirical examination of hallucination. *arXiv preprint*, arXiv.2311.15548.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9004–9017. Association for Computational Linguistics.

Microsoft. 2024. Groundedness detection in Azure AI Content Safety. Azure AI Documentation, https://learn.microsoft.com/en-us/azure/ai-services/content-safety/concepts/groundedness.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12076–12100. Association for Computational Linguistics.

Selvan Sunitha Ravi, Bartosz Mielczarek, Anand Kannappan, Douwe Kiela, and Rebecca Qian. 2024. Lynx: An open source hallucination evaluation model. *arXiv preprint*, arXiv.2407.08488.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. 2024. Docfinqa: A long-context financial reasoning dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 445–458. Association for Computational Linguistics.

Yuda Song, Hanlin Zhang, Carson Eisenach, Sham M. Kakade, Dean P. Foster, and Udaya Ghai. 2025. Mind the gap: Examining the self-improvement capabilities of large language models. *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*.

S. M. Towhidul Islam Tonmoy, S. M. Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint*, arXiv.2401.01313.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, volume EMNLP 2024 of *Findings of ACL*, pages 14199–14230. Association for Computational Linguistics.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. *arXiv preprint*, arXiv.2403.18802.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David S. Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint*, arXiv.2303.17564.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint*, arXiv.2401.15884.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *arXiv preprint*, arXiv.2306.06031.

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. In *Proceedings of the 6th ACM International Conference on AI in Finance, ICAIF 2025, Singapore, November 15-18, 2025*, pages 159–167. ACM.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. Multihiertt: Numerical reasoning over multi hierarchical tabular and textual data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6588–6600. Association for Computational Linguistics.

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 the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3277–3287. Association for Computational Linguistics.

A Financial Claim Taxonomy Examples and Validation

Table 6 provides examples of each claim type with representative hallucination patterns.

Coverage. On the 1,200 FinHalu test examples (4,847 total atomic claims), 97.3% map to exactly one of the six categories. The remaining 2.7% are edge cases involving compound claims, which we split into separate claims.

Distribution. The claim-type distribution in FinHalu is: Numerical (31.2%), Temporal (18.7%), Entity-Attribute (14.3%), Comparative (16.8%), Regulatory (5.2%), and Computational (13.8%). Hallucination rates vary substantially: Computational highest (28.4%), Comparative (22.1%), Numerical (19.7%), Regulatory lowest (8.3%).

Granularity. We compare against a 3-type taxonomy (Numerical, Textual, Derived) and a 10-type taxonomy. The 6-type achieves 91.4% F1; the 3-type achieves 87.1% (-4.3, $p < 0.01$); the 10-type achieves 91.7% ($+0.3$, $p = 0.23$), supporting the 6-type as the preferred accuracy-complexity trade-off.

B Capability Comparison

This appendix compares FINGROUND with the closest existing systems across six capabilities relevant to financial claim verification. Table 7 shows that FINGROUND is the only system combining finance-specific verification, atomic claim decomposition, distilled efficiency, and table-cell-level citations. Multilingual support is the trade-off, discussed in the Limitations section.

C Data Disjointness and FinHalu Construction

We enforce strict disjointness between the 3,200 distillation training examples and all evaluation data at both question and document level. The 1,100 FinQA-derived training examples are drawn exclusively from the FinQA training split; 1,000 TAT-QA-derived examples from the TAT-QA training split; and the 1,200 FinHalu test triples from held-out questions drawn from 87 SEC filings not overlapping with the 142 filings used for training.

The FinHalu test set was constructed before the six-type taxonomy was finalized; annotators labeled claim-level verdicts without reference to FINGROUND's taxonomy. A separate development set

Claim Type	Example Claim	Hallucination Pattern	Detection Strategy
Numerical	"Total revenue was $42.3 billion"	Value substitution ($42.3B vs. actual $38.7B)	Exact-match against table cell
Temporal	"Revenue declined in Q3 2024"	Wrong time period (actually Q2 2024)	Temporal entity extraction
Entity-Attr.	"The CFO is Jane Smith"	Wrong entity role (Jane Smith is COO)	NER + role matching
Comparative	"Revenue grew 15% year-over-year"	Incorrect comparison (actual: 8%)	Arithmetic recomputation
Regulatory	"Per SEC Rule 10b-5 requirements"	Fabricated regulatory reference	Regulatory KB lookup
Computational	"Gross margin was 62.4%"	Incorrect derived metric (actual: 58.1%)	Formula reconstruction

Table 6: Examples of the six financial claim types with typical hallucination patterns.

System	Fin.	Atm.	Dst.	Tbl.	M.L.	Open
FActScore	✗	✓	✗	✗	✓	✓
SAFE	✗	✓	✗	✗	✓	✓
RARR	✗	✗	✗	✗	✓	✓
Self-RAG	✗	✗	✗	✗	✓	✓
CRAG	✗	✗	✗	✗	✓	✓
Lynx	✗	✗	✗	✗	✗	✓
FINGROUND	✓	✓	✓	✓	✗	✓

Table 7: Capability comparison. Fin. = finance-specific verification; Atm. = atomic claim decomposition; Dst. = distilled efficient model; Tbl. = table-cell-level citations; M.L. = multilingual support; Open = open-source availability.

Claim Type	SelfCheck	FINGROUND
Numerical	68.2	$88.1_{\pm 2.3}$
Temporal	74.1	$92.3_{\pm 1.8}$
Entity-Attribute	79.5	$95.6_{\pm 1.1}$
Comparative	70.8	$89.7_{\pm 2.0}$
Regulatory	71.6	$91.8_{\pm 1.9}$
Computational	71.3	$90.2_{\pm 2.2}$
Overall	*72.8*	$91.4_{\pm 1.2}$

Table 8: Detection F1 by financial claim type (±95% CI). Overall F1 is micro-averaged; distribution-weighted macro-average is 90.7.

of 200 examples was used for all system tuning decisions. The claim-type distribution is broadly consistent with PHANTOM and FAITH, suggesting no unusual distributional skew. Existing benchmarks (PHANTOM, FAITH) lack the atomic claim-level annotations required for evaluating FINGROUND's per-claim pipeline.

D Detection by Claim Type

Table 8 breaks down detection F1 by financial claim type, comparing FINGROUND against the strongest open-source baseline (SelfCheckGPT) at retrieval-equalized parity. Entity-attribute claims are easiest (95.6% F1) and numerical claims hardest (88.1%); computational claims show the largest gain over SelfCheckGPT (+18.9 F1), reflecting the contribution of formula reconstruction (§3.2).

E Domain-Adapted Baseline Details

Data Quantity Asymmetry. FINGROUND uses 3,200 distillation examples while HHEM-adapted receives 500 and SelfCheckGPT-adapted receives 1,000 (§5.4). This asymmetry arises from architectural constraints: HHEM is limited to threshold calibration; SelfCheckGPT's self-consistency mechanism does not directly consume supervised labels at scale.

Scaling Projection at Data Parity. To bound the architectural-vs.-data contribution, we project baseline performance to the 3,200-example regime using a logarithmic scaling model with 50% efficiency decay per doubling, anchored on the empirically observed datapoints (HHEM out-of-box: 76.3 F1; HHEM at 500: 81.7; SelfCheckGPT out-of-box: 72.8; SelfCheckGPT at 1,000: 79.4). This functional form (each doubling contributes half the gain of the previous doubling) is the standard conservative regime for adaptation curves on hallucination-detection tasks where labeled data has diminishing returns. Table 9 gives the projected trajectory; FINGROUND's own measured curve (Appendix N: 88.6 at 1,600, 91.4 at 3,200, 92.1 at 6,400) is shown for comparison.

Residual Gap and Caveats. At 3,200-example parity the residual gap is 5.2 to 7.6 F1 points, consistent with the architectural advantages identified in §5.3: claim-type-aware verification rather than uniform claim treatment, and cross-encoder alignment rather than self-consistency or threshold cali-

Examples	HHEM-ad.	SCk-ad.	FINGROUND
0 (out-of-box)	76.3†	72.8†	—
500	81.7†	76.1	—
1,000	84.4	79.4†	—
1,600	85.5	81.5	88.6†
2,000	85.8	82.7	—
3,200	86.2	83.8	91.4†

Table 9: Scaling projection at data parity (F1). HHEM-ad. = HHEM-adapted; SCk-ad. = SelfCheckGPT-adapted. †Empirically observed; remaining values are log-projected with 50% efficiency decay per doubling. At 3,200-example parity, the residual architectural gap is 5.2 F1 (vs. HHEM-ad.) to 7.6 F1 (vs. SCk-ad.).

bration. We did not run a full retraining experiment at 3,200 examples for both baselines: HHEM's threshold-only adaptation surface does not natively consume supervised labels at that scale without architectural extension, and SelfCheckGPT's adaptation pathway is bottlenecked by its sampling-based design. A controlled retraining at parity (with the appropriate architectural surgery for HHEM) is the cleanest follow-up; the projection bounds reported here should be read as a methodologically conservative estimate rather than a substitute for that experiment.

F Human Validation Details

Three financial domain experts independently annotated 200 FinHalu examples without seeing GPT-4o labels. Agreement between expert consensus and GPT-4o: $\kappa = 0.87$, with 94.1% verdict match. Among 12 disagreements: 5 involve hedged language in risk disclosures, 4 involve restated figures, 3 involve implicit temporal references. GPT-4o is more conservative than domain experts on ambiguous constructs, biasing toward higher precision at some recall cost, which is acceptable for compliance applications.

G Regeneration Examples and Error Compounding

Successful Regeneration. Original (contradicted): "Apple's gross margin was 42.3% in Q4 2023." **Evidence**: Table 2, Row: Gross Margin, Col: Q4 2023 $\rightarrow$ 45.2%. **Regenerated**: "Apple's gross margin was 45.2% in Q4 2023 [Doc: AAPL-10K, Table 2, Row: Gross Margin, Col: Q4 2023]."

Error Introduction (4.1% of cases). Original (contradicted): "Operating margin improved by 3.2 percentage points." **Evidence**: Op. income \$12.1B (2023) vs. \$10.8B (2022); Revenue \$52.3B (2023) vs. \$49.1B (2022). **Regenerated**: "Operating margin improved by 1.2 pp." (Actual: 1.1 pp; rounding error.)

Error Compounding. Distribution of claims per answer requiring regeneration: 1 claim (72%), 2 claims (18%), $\geq$3 claims (10%). Per-answer error rate: 4.1% [2.1, 6.8] for single-claim, 7.9% [3.2, 14.1] for two-claim, 14.3% [5.7, 28.2] for $\geq$3-claim. Full re-generation for $\geq$3-claim answers yields 8.6% [3.1, 16.4]; CIs overlap with pre-mitigation, so evidence is suggestive rather than definitive.

H Cross-Encoder Evaluation

Alignment Method	P	R	F1
BM25 top-1	$71.3_{\pm 3.2}$	$68.9_{\pm 3.4}$	$70.1_{\pm 2.8}$
Dense cosine (E5)	$79.8_{\pm 2.7}$	$82.1_{\pm 2.5}$	$80.9_{\pm 2.2}$
Cross-enc. (1:1)	$80.4_{\pm 2.8}$	$\mathbf{93.1}_{\pm 1.7}$	$86.3_{\pm 2.0}$
Cross-enc. (1:3)	$88.6_{\pm 2.1}$	$85.9_{\pm 2.4}$	$\mathbf{87.2}_{\pm 1.8}$
Cross-enc. (1:5)	$\mathbf{91.8}_{\pm 1.9}$	$82.6_{\pm 2.6}$	$87.0_{\pm 1.9}$

Table 10: Claim-evidence alignment on 840 held-out pairs ($\pm$95% CI). The 1:3 ratio achieves best F1.

On 120 OOD claim-evidence pairs from FinanceBench: $84.1\%_{\pm 3.1}$ F1 (3.1-point drop). The FinanceBench end-to-end HalRate (4.9%, 1.3pp higher than FinQA) is consistent with this OOD degradation partially propagating through the pipeline, with dampening from the multi-component structure.

I Cross-Generator Evaluation

Generator	Prec.	Rec.	F1
GPT-4o (in-distribution)	$92.7_{\pm 1.1}$	$90.2_{\pm 1.3}$	$91.4_{\pm 1.2}$
GPT-3.5-Turbo (in-dist.)	$91.8_{\pm 1.3}$	$89.4_{\pm 1.5}$	$90.6_{\pm 1.1}$
Llama-3-70B (OOD)	$89.3_{\pm 2.4}$	$86.8_{\pm 2.7}$	$88.0_{\pm 2.1}$
Claude-3.5-Sonnet (OOD)	$90.1_{\pm 2.2}$	$85.2_{\pm 2.8}$	$87.6_{\pm 2.2}$

Table 11: Cross-generator detection F1 ($\pm$95% CI). 3–4 point degradation on OOD generators.

Degradation is non-uniform: computational claims show smallest drop (-1.8 F1, generator-agnostic arithmetic), comparative claims show largest (-5.2 F1, varied phrasing). Transferability to smaller or domain-fine-tuned models is untested. Generator-specific calibration data (a few hundred examples) is recommended for non-GPT backends.

J Error Analysis

False Negatives (8.6%). Paraphrased numerical errors (38% of FN): hallucinated values close to truth challenge the cross-encoder. Multi-hop reasoning gaps (29%): partial evidence retrieval for chained values. Implicit temporal context (19%): unstated time periods prevent matching.

False Positives (6.1%). Hedged language (52% of FP): risk disclosures with "may," "could potentially." Restated figures (31%): both original and restated values present.

Near-Miss Adversarial Analysis. On 50 claims with hallucinated values within $\pm 5\%$ of ground truth, recall drops to $71.4\%_{\pm 6.3}$. False negatives concentrate in $\leq$0.3pp differences and $\pm 2\%$ currency amounts. For decisions where small errors have material consequences, mandatory human verification of all numerical claims is required. Numerical tolerance thresholds catch 89% of near-misses but increase FP rate by 8.3pp; calibrated confidence scoring shows promise (AUROC: 0.78) but requires further calibration.

K Latency Decomposition

Pipeline Stage	p50	p95	% of p95
Stage 1: Retrieval + Routing	0.7s	1.2s	31.6%
Stage 2: Decomp. + Verification	1.4s	2.1s	55.2%
Stage 3: Regeneration (cond.)	0.3s	0.5s	13.2%
Full Pipeline	**2.4s**	**3.8s**	100%

Table 12: Latency per query on A100. Stage 2 is the bottleneck (55%). Stage 3 invoked for $\sim$38% of queries. Batched claim verification reduces Stage 2 from 4.8s sequential to 2.1s p95.

L Deployment Details

Filing Quality. Of 43 pilot filings, 7 (16.3%) contained OCR artifacts, 5 (11.6%) had non-standard formatting, 3 (7.0%) had both. On these 15 lower-quality filings, false negative rate was modestly higher (5.2% vs. 3.1%), driven by OCR-induced retrieval failures. On 12 filings from 2024 post-dating all training data, comparable false negative rates (4.1% vs. 3.8%) provide a preliminary temporal signal.

Scalability. For 500 filings/day at 50 queries/filing (25,000 queries/day), the requirement is $\sim$0.3 QPS sustained. FINGROUND's 8.4 QPS on a single A100 provides 28$\times$ headroom. Horizontal scaling to 2–3 A10G instances supports the largest institutions.

Graceful Degradation. Stale retrieval $\rightarrow$ flag-only mode; malformed documents $\rightarrow$ confidence-scored parser routes low-confidence to manual processing; model unavailability $\rightarrow$ existing manual workflow.

Domain Adaptation Cost. Distillation efficiency: 1,600 examples yield 88.6% F1; 3,200 yield 91.4%; 6,400 yield 92.1%. Achieving $\geq$90% F1 requires $\sim$2,500–3,000 examples. Adding a new claim type requires $\sim$200 examples and $\sim$8 GPU-hours.

Operational Failure Handling. Analyst overrides are logged for periodic model retraining. Domain-specific confidence thresholds suppress low-confidence flags. Ambiguous cases escalate through existing compliance review workflows.

M Baseline Implementation Details

Table 13 lists exact codebase versions, commit hashes, and key hyperparameters for all baselines used in the main experiments, ensuring reproducibility. All baselines run on a single A100 (80GB) under identical compute budgets to FINGROUND, with default hyperparameters from official codebases.

System	Codebase / Version	Key Config
Self-RAG	Official repo (Asai et al.), commit `a3f7c21`, Llama-2-13B	Default reflection tokens; 3 passages
CRAG	Official repo (Yan et al.), commit `e8b4d09`	Default thresholds; web search disabled
SelfCheck-GPT	`selfcheckgpt` v0.1.7	BERTScore; 5 samples; temp. 0.7
HHEM	Vectara HHEM v1.0	Default threshold (0.5)
GPT-4o+CoT	`gpt-4o-2024-05-13`	3-shot CoT; temp. 0.0
FActScore	Official repo (Min et al.), commit `b2c9e47`	Retrieval-equalized; GPT-4o verifier
Vanilla RAG	BM25 + `gpt-4o-2024-05-13`	Top-5 passages; no re-ranking

Table 13: Baseline implementations. All use official codebases with default hyperparameters and identical compute budget (single A100, 80GB).

N Distillation Training Details

Data Composition. 3,200 examples: FinQA-derived (1,100, training split only), TAT-QA-derived (1,000, training split only), SEC filing excerpts from 142 non-overlapping filings (800), adversarial examples with synthetic hallucinations (300).

Hyperparameters. Llama-3-8B-Instruct, 3 epochs, lr 2e-5, batch 16, reverse KL τ=2.0. Multi-task loss: decomposition (0.3), alignment (0.3), verdict (0.4). Training: 8 hours on 4×A100. Seed variance: F1 = 91.4 $\pm$ 0.4 across 3 seeds.

Efficiency Curve. 1,600 examples: 88.6% F1; 3,200: 91.4%; 6,400: 92.1% (+0.7), suggesting near-optimal training set size.

O E5-Large Fine-tuning Details

E5-large fine-tuned on 12K financial passage pairs from FinQA/TAT-QA training splits. Positive: (question, gold evidence); hard negatives: BM25-sampled same-filing passages. 5 epochs, lr 1e-5, batch 64, InfoNCE loss. Recall@5: 84.3% (vs. 67.1% base).

P Complexity Classifier

Predicted →	Simp.	Mod.	Comp.	Total
Simple	**153**	12	3	168
Moderate	7	**122**	11	140
Complex	4	6	**82**	92
Total	164	140	96	400

Table 14: Complexity classifier confusion matrix (400 held-out queries, 89.3% accuracy). Over-routing wastes compute but preserves accuracy; under-routing risks missed evidence.

Q 43-Point Gap Experiment

On 200 computational claims from FinHalu, FActScore's generic pipeline catches 57% of errors (F1: 67.1%); FINGROUND's domain-specific pipeline catches 100% (F1: 98.5%), a 43pp recall gap. This comparison changes multiple factors simultaneously (decomposition, type-aware routing, arithmetic re-computation, evidence alignment); the gap reflects the full integrated pipeline benefit. The ablation (−taxonomy, Table 2) provides a cleaner single-factor comparison.

R FActScore Baseline Configuration

FActScore applied with retrieval-equalized configuration (same passages as other baselines), standard InstructGPT decomposition without financial modification, and GPT-4o verification. This isolates the effect of domain-specific claim typing. FActScore achieves 76.7% F1 on FinHalu, with the gap concentrated on computational claims (58.4% vs. FINGROUND's 90.2%).

S Prompts and Templates

```
You are a financial claim verification
assistant. Given a generated answer about
a financial document, decompose it into
atomic, independently verifiable claims.
For each claim:
1. Extract the exact assertion.
2. Classify as: Numerical, Temporal,
   Entity-Attribute, Comparative,
   Regulatory, or Computational.
3. For Numerical: extract value, unit,
   entity, time_period.
4. For Computational: identify the
   implied formula or derivation.

Answer: {answer}
Evidence: {evidence_summary}

Output (JSON): [{"claim": "...",
  "type": "...",
  "structured_fields": {...}}]
```

The GPT-4o teacher annotation uses a two-pass process: Pass 1 generates claim-level annotations; Pass 2 verifies via a consistency check (Bohnet et al., 2022), discarding 8.4% of inter-pass disagreements.

T Design Journey

Several design decisions emerged iteratively. Initial end-to-end verification without claim decomposition missed computational errors embedded in longer claims. For taxonomy granularity, a 12-type system dropped inter-annotator agreement below κ=0.70; the six types were chosen at the boundary of different verification strategies (exact match vs. re-computation vs. NLI). Structure-aware table chunking required three iterations: regex-based extraction failed on merged cells, heuristic column detection broke on multi-level headers, and the final column-header-aware function emerged from systematic analysis of 50 failure cases.

U Extended Limitations

Regeneration Error Compounding. The 4.1% per-claim error introduction rate compounds in

multi-claim answers. Full re-generation mitigation for $\geq$3-claim answers shows directionally positive results, but small subgroup sizes ($n \approx 30$) with overlapping CIs preclude definitive conclusions.

Domain Adaptation Cost. Adapting to a new jurisdiction requires $\sim$2,500–3,000 annotated examples plus taxonomy extension if new claim types arise.

Scalability. Efficiency measurements are on benchmark-sized corpora; enterprise-scale stores may introduce retrieval latency challenges.

Cross-Encoder Alignment. The 87.2% alignment F1 means $\sim$13% misalignment propagating downstream.

Confidence Calibration. We have not formally assessed calibration quality via reliability diagrams; this is planned for production evaluation.

V Additional Results

Cross-Benchmark Generalization. On ConvFinQA (Chen et al., 2022) and MultiHiertt (Zhao et al., 2022) without task-specific adaptation, FINGROUND achieves HalRates of 5.1% ($\pm$1.3) and 6.8% ($\pm$1.5).

Model Size Ablation. 1.5B: 82.1% F1; 3B: 87.3% F1; 8B: 91.4% F1.

Full Pipeline without Distillation. GPT-4o verifier: 2.9% HalRate on FinQA, 95.0% F1, at 8.2s latency and $0.047/query. The distilled model trades 3.6 F1 points for 2.2$\times$ latency and 15.7$\times$ cost reduction.

Citation Quality. On FinanceBench: FINGROUND achieves 92.1% CitP and 87.6% CitR, vs. GPT-4o+CoT (68.4%/54.2%), Self-RAG (72.1%/61.8%), RARR (79.3%/70.5%).