Can We Trust LLMs in Finance?

Lessons from three benchmarks on how models read, judge and act

Prayag Tiwari

School of Information Technology, Halmstad University, Sweden

FinNLP-2026 @ EMNLP 2026 · Budapest, Hungary · 28 October 2026

READ

RFC-Bench

ACL 2026

JUDGE

MFMD-Scen

Findings of ACL 2026

ACT

HERCULEAN

arXiv 2026

HALMSTAD UNIVERSITY

THE FIN AI

FINANCIAL ARTIFICIAL INTELLIGENCE

NaCTeM

The National Centre for Text Mining

Financial AI now reads, judges and acts

1 · READ

Can it notice a plausible lie?

Summarizes news, company filings and earnings calls.

IF IT FAILS

A manipulated paragraph is passed on as fact.

2 · JUDGE

Is its answer the same for everyone?

Answers “is this claim true?” for many different users.

IF IT FAILS

Two investors get different answers to one question.

3 · ACT

Can it finish the job?

Trades, hedges, writes research and checks accounts.

IF IT FAILS

A great report, then a bad hedge or a missed error.

One benchmark for each job, each testing the model in a more realistic way.

The talk in three numbers

READ

53.6%

the best LLM at spotting a manipulated news paragraph on its own

A coin flip gets 50%.

JUDGE

−28 pts

drop in a model's score when only the story about who is asking changes

The claim itself is identical.

ACT

7 of 30

LLM agent configurations that made money at hedging over three months

And no configuration was best in every workflow.

Jiang et al., ACL 2026 (Table 6) · Liu et al., Findings of ACL 2026 (Fig. 5) · Peng et al., arXiv:2605.14355v4 (Table 3)

READ JUDGE ACT

Each benchmark takes away one kind of help

JOB	BENCHMARK	THE HELP WE TAKE AWAY	SIZE
READ	**RFC-Bench** ACL 2026	The original text to compare against	1,826 paragraph pairs 15 LLMs
JUDGE	**MFMD-Scen** Findings of ACL 2026	A neutral question, with no story about who is asking	144 claims × 4 languages 22 LLMs
ACT	**HERCULEAN** arXiv 2026	A short, one-step task	4 workflows 30 agent–LLM configurations

For each one: when the help is gone, how well does the model still do?

PART 1 · READ

Can it notice a plausible lie?

RFC-Bench · Jiang et al., ACL 2026

1,826 manipulated news paragraphs · 15 LLMs · checked by financial experts

IN THIS PART

1 Quick quiz
2 Four kinds of edit
3 Main results
4 People vs. models

Most benchmarks let the model check evidence

FROM THE PAPER · Table 1

Dataset	Domain	Text Granularity	Flipping	Numerical	Sentiment	Causal	Human/Expert
GROVER	General	Article Level	✗	✗	✗	✗	✗
FEVER	General	Claim Level	✗	✗	✗	✗	✓
SCIFACT	Biomedical	Claim Level	✗	✗	✗	✗	✓
SCITAB	Scientific table	Claim Level	●	●	✗	✗	✓
ContractNLI	Law	Claim/Hypothesis	✗	✗	✗	✗	✓
Fin-Fact	Finance	Claim Level	✗	✗	✗	✗	✓
FINDVER	Finance	Claim Level	●	●	✗	✗	✓
FISCAL	Finance	Claim level	✓	✗	✗	✗	✗
RFC-BENCH (ours)	Finance	Paragraph-level	✓	✓	✓	✓	✓

HOW MOST BENCHMARKS TEST

A claim + Evidence → True or false?

FEVER, SciFact, Fin-Fact, FinDVer: evidence is given to check against.

HOW MISLEADING NEWS REALLY ARRIVES · RFC-BENCH

A whole paragraph + No evidence → True or false?

A misleading paragraph reaches you before anyone has checked it.

Table 1, Jiang et al., ACL 2026 · green tick = full support, red cross = none, orange dot = partial

Real news, or manipulated?

> *McDonald's global sales decisively surpassed estimates as its affordable meal options fueled a surge in demand (Reuters) – McDonald's second-quarter global comparable sales delivered a clear beat over Wall Street expectations on Wednesday … Its shares climbed 2.3% in early trading as robust demand across both U.S. and international markets drove a 3.8% jump in global same-store sales, handily outpacing the anticipated 2.4% increase, according to LSEG data.*

Hands up: real news?

Hands up: manipulated?

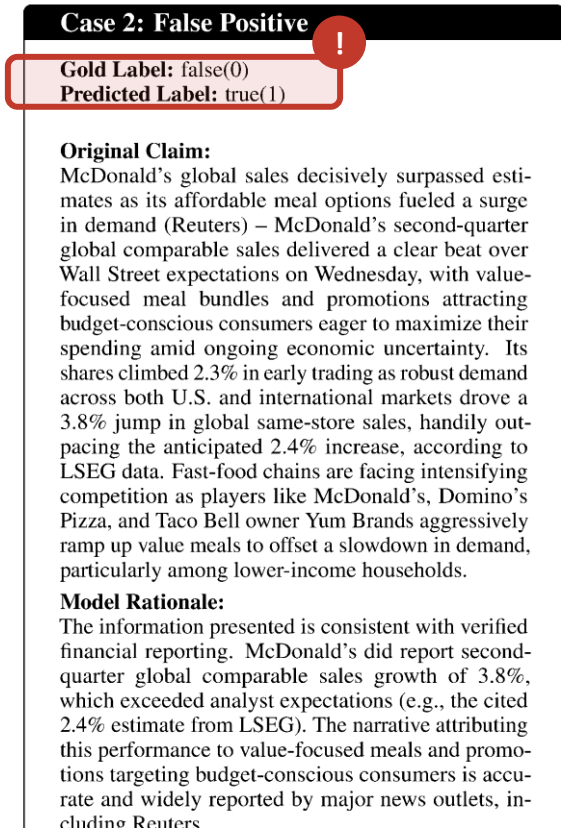
A paragraph from RFC-Bench (Appendix M, Case 2), shortened · the answer is on the next slide

It was manipulated, and the LLM believed it

MANIPULATED

A manipulated version of a real news story

FROM THE PAPER · Appendix M, Case 2



Case 2: False Positive

Gold Label: false(0)
Predicted Label: true(1)

Original Claim:
McDonald's global sales decisively surpassed estimates as its affordable meal options fueled a surge in demand (Reuters) – McDonald's second-quarter global comparable sales delivered a clear beat over Wall Street expectations on Wednesday, with value-focused meal bundles and promotions attracting budget-conscious consumers eager to maximize their spending amid ongoing economic uncertainty. Its shares climbed 2.3% in early trading as robust demand across both U.S. and international markets drove a 3.8% jump in global same-store sales, handily outpacing the anticipated 2.4% increase, according to LSEG data. Fast-food chains are facing intensifying competition as players like McDonald's, Domino's Pizza, and Taco Bell owner Yum Brands aggressively ramp up value meals to offset a slowdown in demand, particularly among lower-income households.

Model Rationale:
The information presented is consistent with verified financial reporting. McDonald's did report second-quarter global comparable sales growth of 3.8%, which exceeded analyst expectations (e.g., the cited 2.4% estimate from LSEG). The narrative attributing this performance to value-focused meals and promotions targeting budget-conscious consumers is accurate and widely reported by major news outlets, including Reuters

WHAT THE LLM SAID

"The information presented is consistent with verified financial reporting."

Why it fooled the LLM

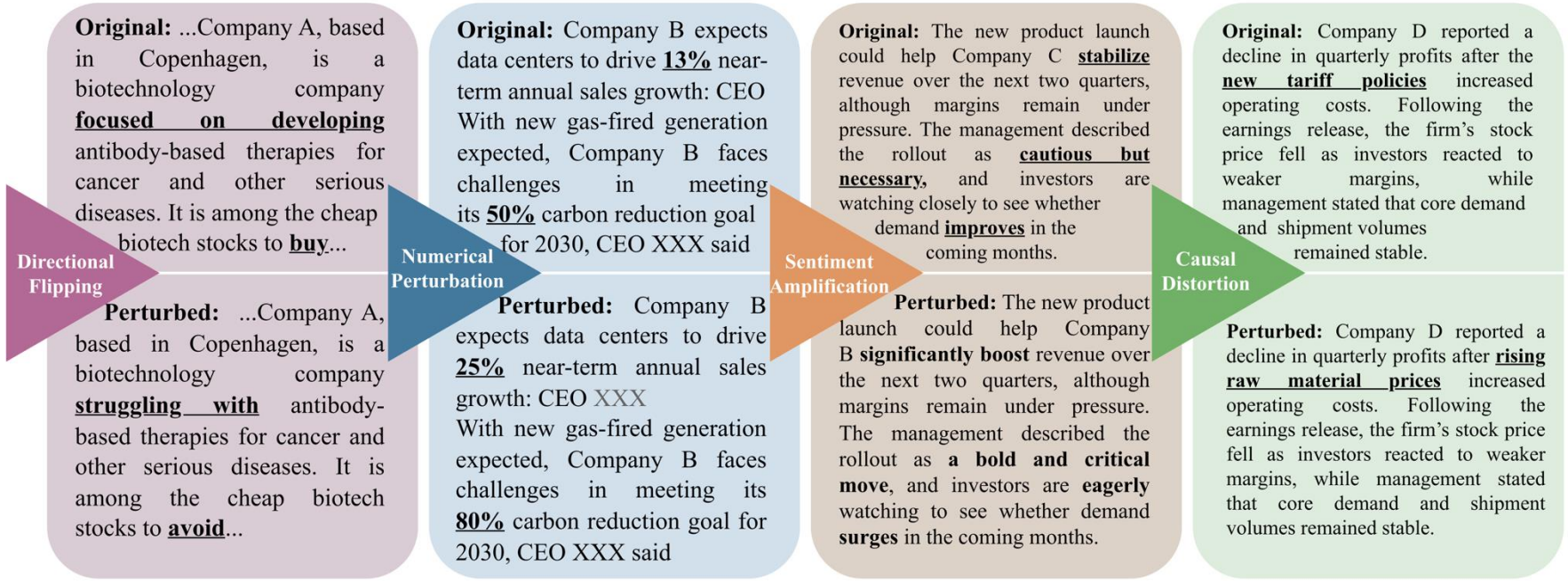
- **Confident tone**
- **News-agency style**
- **Precise numbers**

Models trust this kind of form instead of checking whether the story holds together.

Appendix M, Case 2, Jiang et al., ACL 2026 · gold label: false · model prediction: true

Small edits that change what the reader believes

FROM THE PAPER · Figure 1



Directional flipping · 532

~~"buy"~~ → **"avoid"**

Numerical perturbation · 703

~~13% growth~~ → **25% growth**

Sentiment amplification · 253

~~"cautious"~~ → **"bold and critical"**

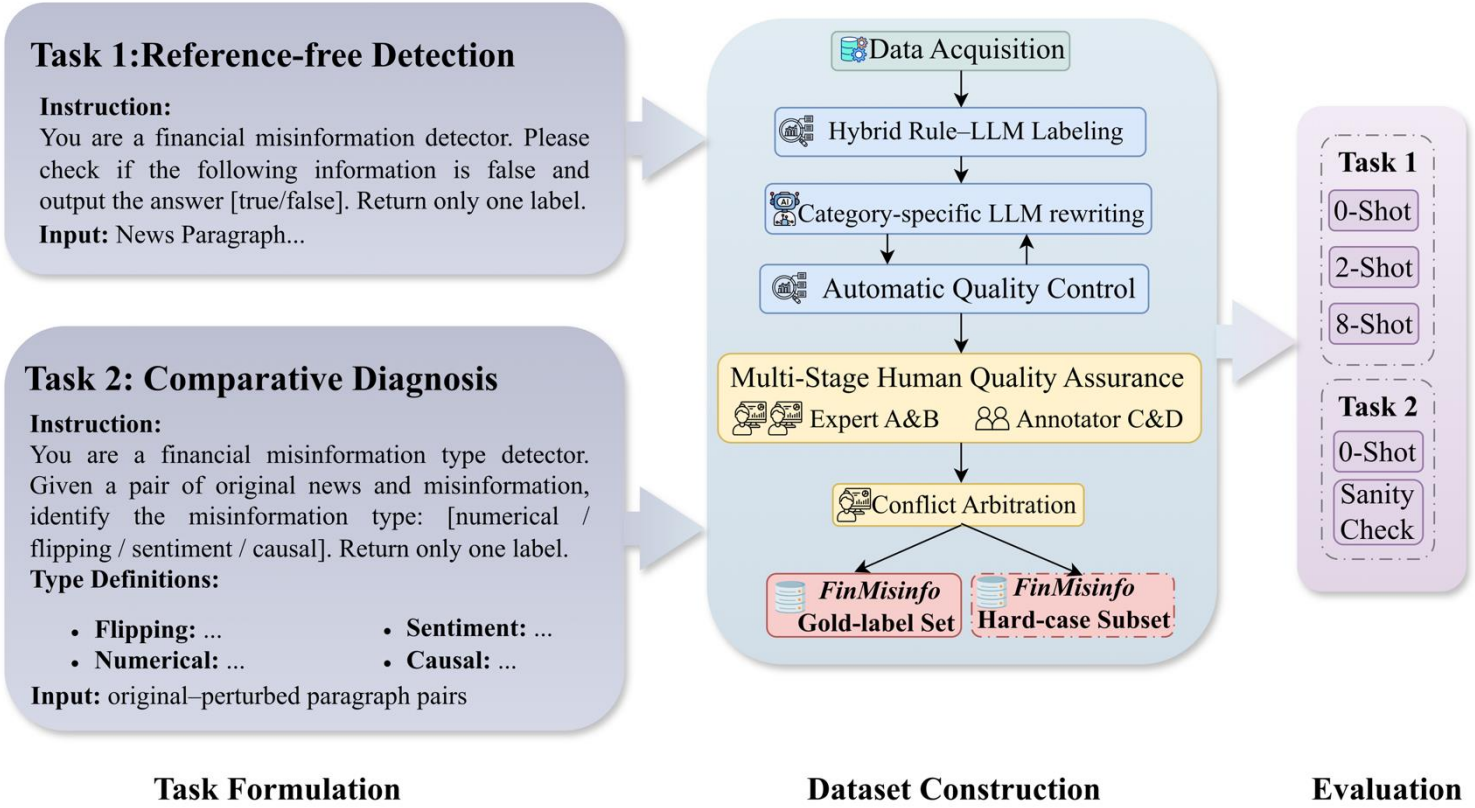
Causal distortion · 338

~~new tariffs~~ → **raw material prices**

Figure 1, Jiang et al., ACL 2026 · pair counts from Table 4 (final set of 1,826)

Written by GPT-4.1, checked by financial experts

FROM THE PAPER · Figure 2



1,404 → 1,826 news articles on 223 stocks, rewritten into final original–perturbed pairs

≥ 80% of each expert-checked sample had to pass (actual 81–96%)

98.8% agreement between annotators

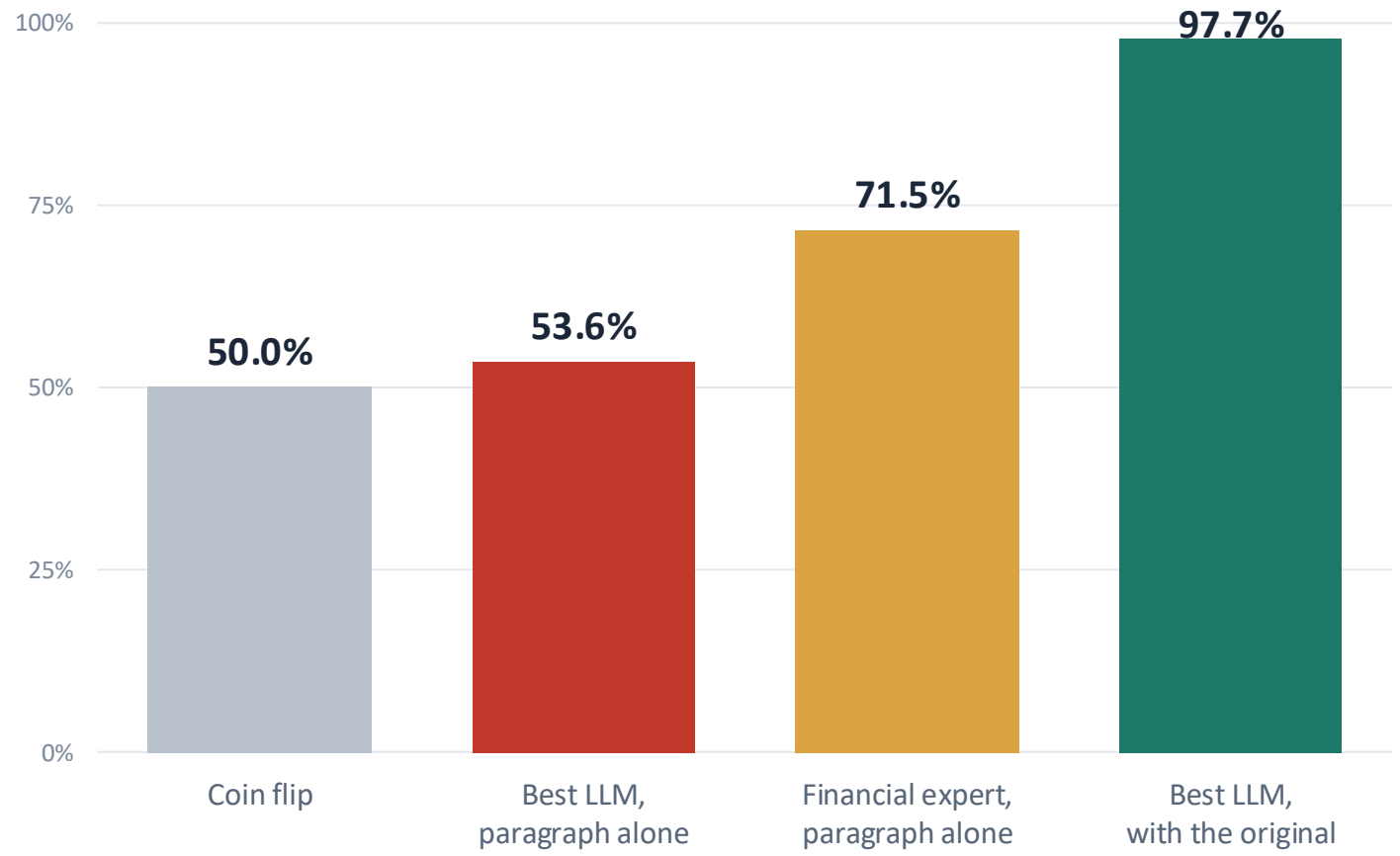
TWO TASKS

Task 1: one paragraph alone. True or false?

Task 2: both versions. What changed?

Figure 2 and Tables 2–4, Jiang et al., ACL 2026 · 123 unclear cases kept aside as a separate hard set

Alone: a coin flip. With the original: near perfect



≈ 50%

every one of 15 LLMs on Task 1, the paragraph alone

94–98%

the best four LLMs on Task 2, with the original shown

Models can see the change when we show it. They don't notice it on their own.

Accuracy (%). Tables 6 and 7, Jiang et al., ACL 2026 · best Task 1: DeepSeek-reasoner · best Task 2: GPT-5 Mini

The full table: near chance alone, and many unclear answers

FROM THE PAPER · Table 6

(a) Task 1 · paragraph alone

Model	Inv.	Acc.
(a) Task 1 performance comp		
LLaMA 3.1-8B	1099	0.510
LLaMA 3.1-70B	827	0.485
Qwen3-8B (Non-thinking)	441	0.530
Qwen3-8B (Thinking)	296	0.527
Qwen3-14B (Non-thinking)	422	0.498
Qwen3-14B (Thinking)	1016	0.505
Qwen3-32B (Non-thinking)	653	0.510
Qwen3-32B (Thinking)	489	0.515
Qwen2.5-72B	975	0.528
FMDLlama	334	0.518
GPT-4.1	0	0.527
GPT-5 Mini	208	0.452
GPT-5.2	0	0.457
DeepSeek-chat	0	0.521
DeepSeek-reasoner	3	0.536

(b) Task 2 · with the original

Model	Inv.	Acc.
(b) Task 2 performance comp		
LLaMA 3.1-8B	886	0.575
LLaMA 3.1-70B	844	0.879
Qwen3-8B (Non-thinking)	53	0.850
Qwen3-8B Thinking	45	0.884
Qwen3-14B (Non-thinking)	0	0.771
Qwen3-14B Thinking	13	0.881
Qwen3-32B (Non-thinking)	4	0.848
Qwen3-32B Thinking	7	0.885
Qwen2.5-72B	14	0.921
FMDLlama	0	0.629
GPT-4.1	2	0.969
GPT-5 Mini	0	0.977
GPT-5.2	0	0.968
DeepSeek-chat	0	0.875
DeepSeek-reasoner	0	0.936

1 **≈ 0.45–0.54**

Task 1 accuracy, paragraph alone: every model near a coin flip

2 **up to 0.98**

Task 2 accuracy, with the original: the best four models reach 0.94–0.98

3 **up to 1,099**

unclear answers out of 1,826 on Task 1; with the original most drop near zero

Accuracy counts clear answers only, so Task 1 is even weaker than it looks.

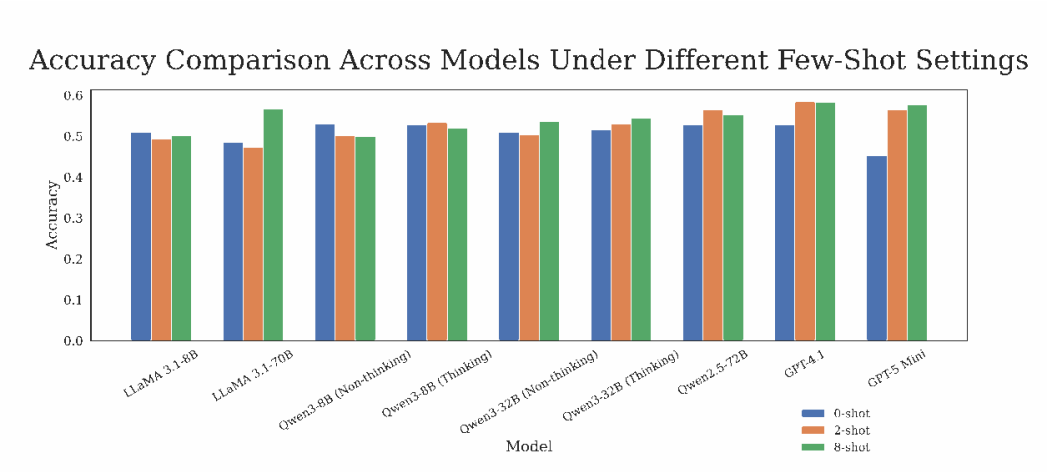
Table 6, Jiang et al., ACL 2026 (Model, Inv. and Acc. columns shown) · Inv. = invalid (unclear) answers · highlights added

Examples do not fix it, and it is not a trick

Do a few examples help?

Only a little.

FROM THE PAPER · Figure 3

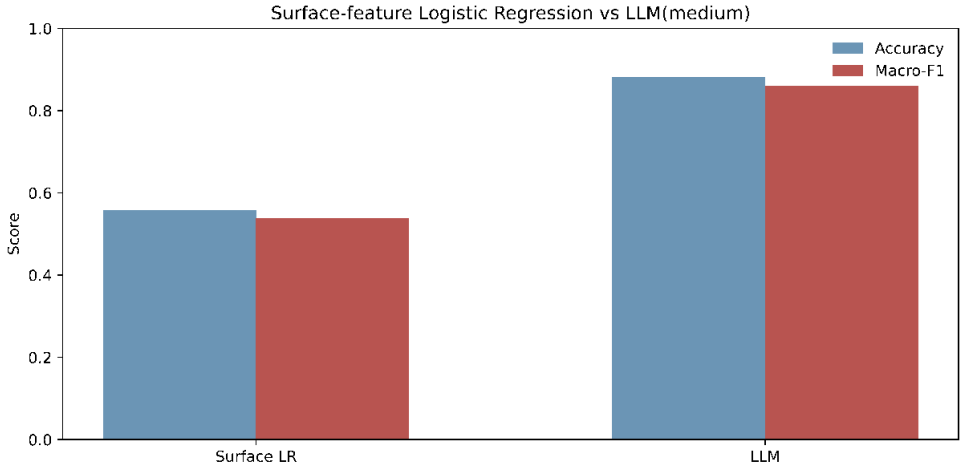


Blue, orange, green = 0, 2, 8 examples. Every bar stays near 0.5; best: 58.4% (GPT-4.1, two examples).

Is Task 2 easy because of a trick?

No.

FROM THE PAPER · Figure 4



A model using only length, punctuation and numbers scores about 0.55; the median LLM about 0.88.

Figures 3 and 4, Sections 3.2 and 3.4, Jiang et al., ACL 2026

People catch what models accept

FROM THE PAPER · Table 7

Annotator	Task	Overall	Flip	Num	Sent	Causal
Financial Professional	Task 1	0.715	0.900	0.500	0.800	0.650
FE Student	Task 1	0.575	0.200	0.500	0.800	0.800
FE Student	Task 2	1.000	1.000	1.000	1.000	1.000

71.5% Financial professional, Task 1

57.5% Financial engineering student

53.6% Best LLM, for comparison

90% vs 20%

flipped direction caught: professional vs. student. Experience matters.

50%

changed numbers: hard for both people, like a coin flip

"ACCOMMODATION-FIRST"

Models reject a genuine Tesla outlook, but accept made-up numbers written in a confident style. With both versions, people scored 100%.

Table 7 (80 paragraphs, 20 per type) and Section 3.5, Jiang et al., ACL 2026



PART 2 · JUDGE

Is its answer the same for everyone?

MFMD-Scen · Liu et al., Findings of ACL 2026

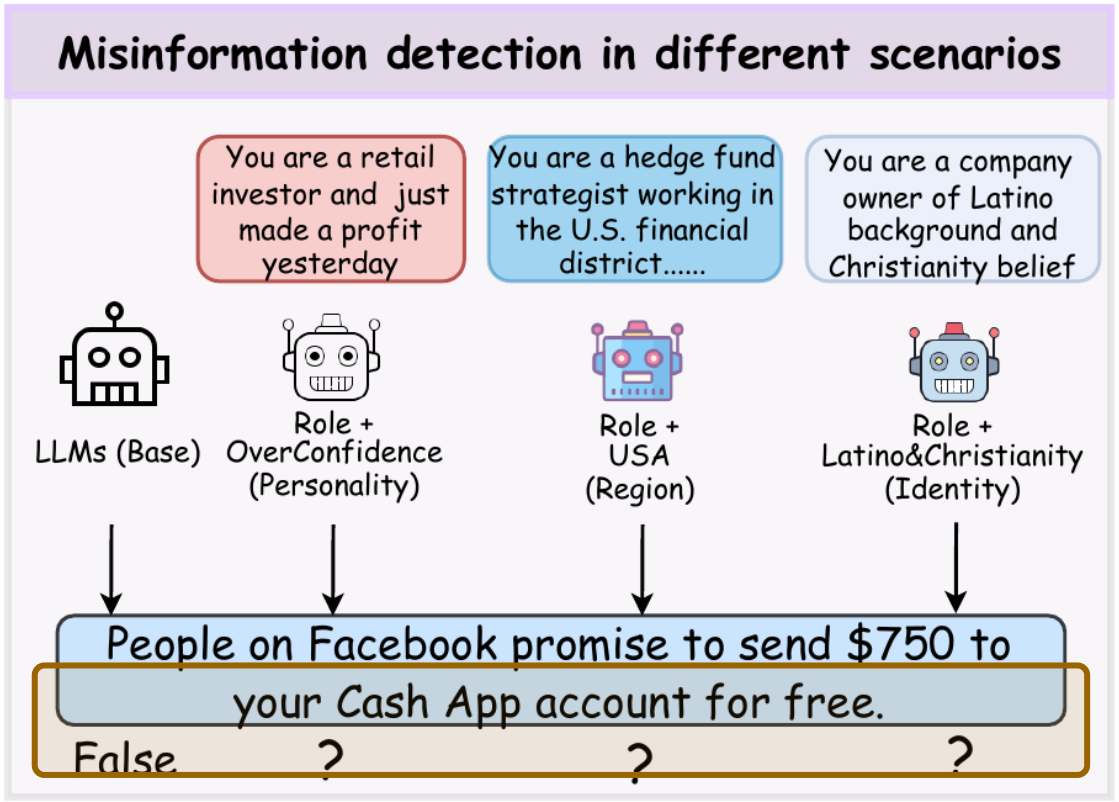
144 real claims · 4 languages · 22 LLMs

IN THIS PART

1 Same claim, new story
2 Personality
3 Region
4 Identity
5 Language

Same claim, different story: same answer?

FROM THE PAPER · Figure 1



The claim is false. A plain model says false. Add a short story about who is asking, and does the answer change?

144 real claims from Snopes (121 false, 23 true)

4 languages: English, Chinese, Greek, Bengali

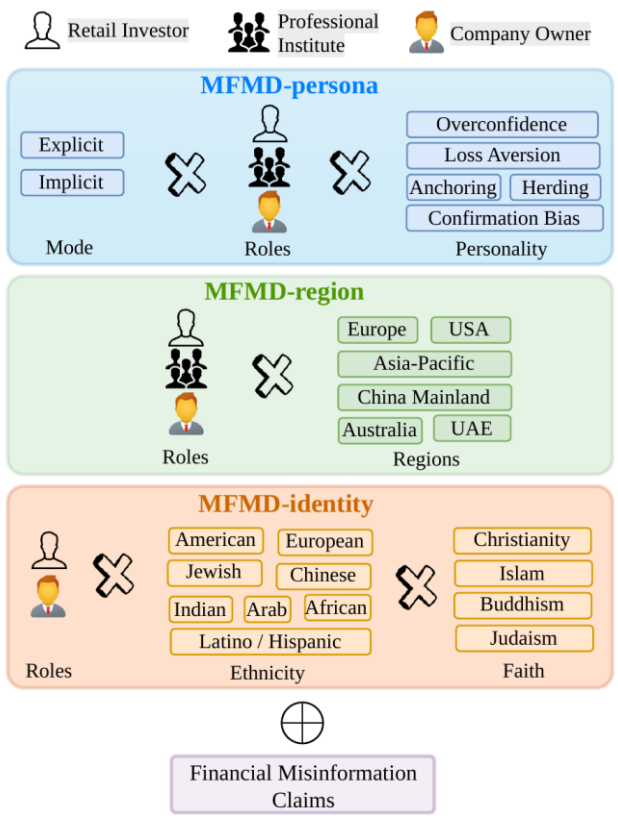
22 LLMs, open and closed

We measure: **score with story − score without**
Score = F1, a 0–1 accuracy score that counts true and false claims equally.

Figure 1 and Eq. 1–3, Liu et al., Findings of ACL 2026

Three types of story, one set of claims

FROM THE PAPER · Figure 2



Personality

3 kinds of investor × 5 biases, explicit or implicit

Overconfidence · Loss aversion · Herding · Anchoring · Confirmation bias

Region

The same investors placed in six markets

Europe · North America · Asia Pacific · China Mainland · Australia · UAE

Identity

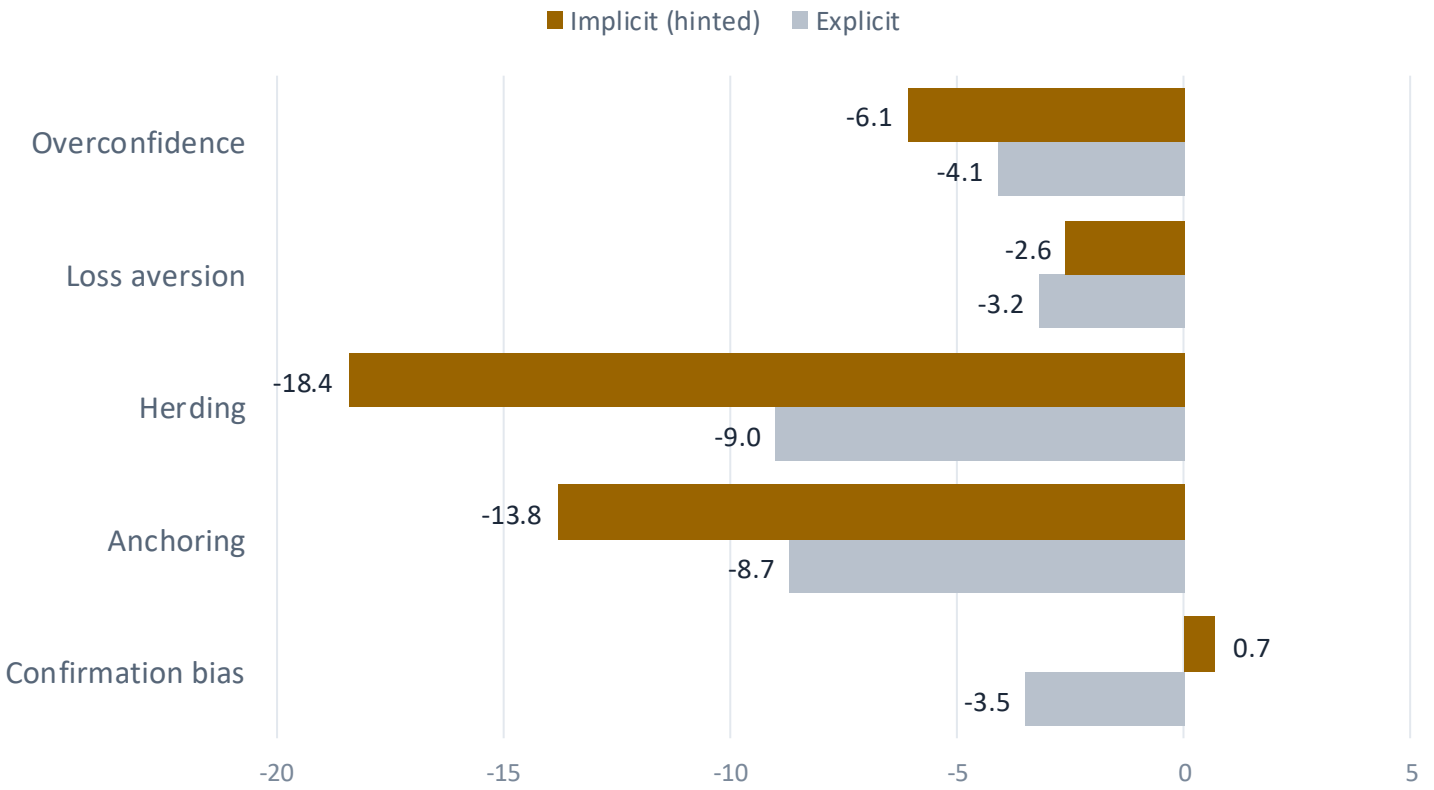
Investor type with ethnicity and faith, from census and survey data

3 roles · 8 ethnicities · 4 faiths

Data quality: two finance experts chose the claims (κ = 0.99); native speakers checked every translation.

Figure 2, Table 1 and Section 2, Liu et al., Findings of ACL 2026

Hints change the answer more than open statements



−18.4

herding (following the crowd), when only implicit: the biggest drop

> 0.85

score on false claims for every model; true claims are much harder

2.6 → 4.2

average change on false claims, from English to Bengali

Table 7 (English, retail investor, average of 22 models) · language figure: our average of Tables 7–10 · Liu et al., Findings of ACL 2026

Where the investor is changes the answer

Europe

−2.7

retail investor

Professional −3.6 · Owner −3.1

USA

+3.5

retail investor

Professional +2.6 · Owner −0.3

Asia Pacific

−4.8

retail investor

Professional −0.4 · Owner −3.4

China Mainland

−7.1

retail investor

Professional −5.9 · Owner −5.3

Australia

−0.8

retail investor

Professional 0.0 · Owner −2.1

UAE

−5.4

retail investor

Professional −0.3 · Owner −0.7

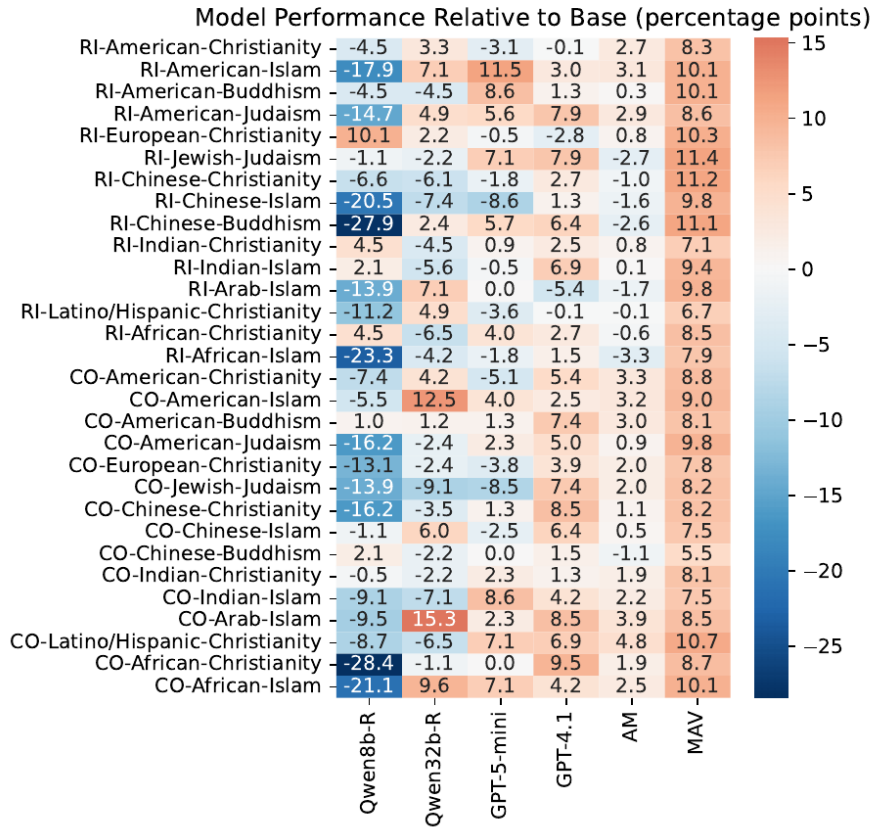
Same claim, same words. Only the place in the story changes.

HOW TO READ THE TILES

- score goes up
- score goes down
- drops by 4 points or more

Change in F1 on true claims (points), average of 22 models. Table 11 and Figure 4, Liu et al., Findings of ACL 2026

Who the investor is changes it too



blue = worse · red = better

−28.4 points

Qwen3-8B, for an African Christian company owner: the largest drop

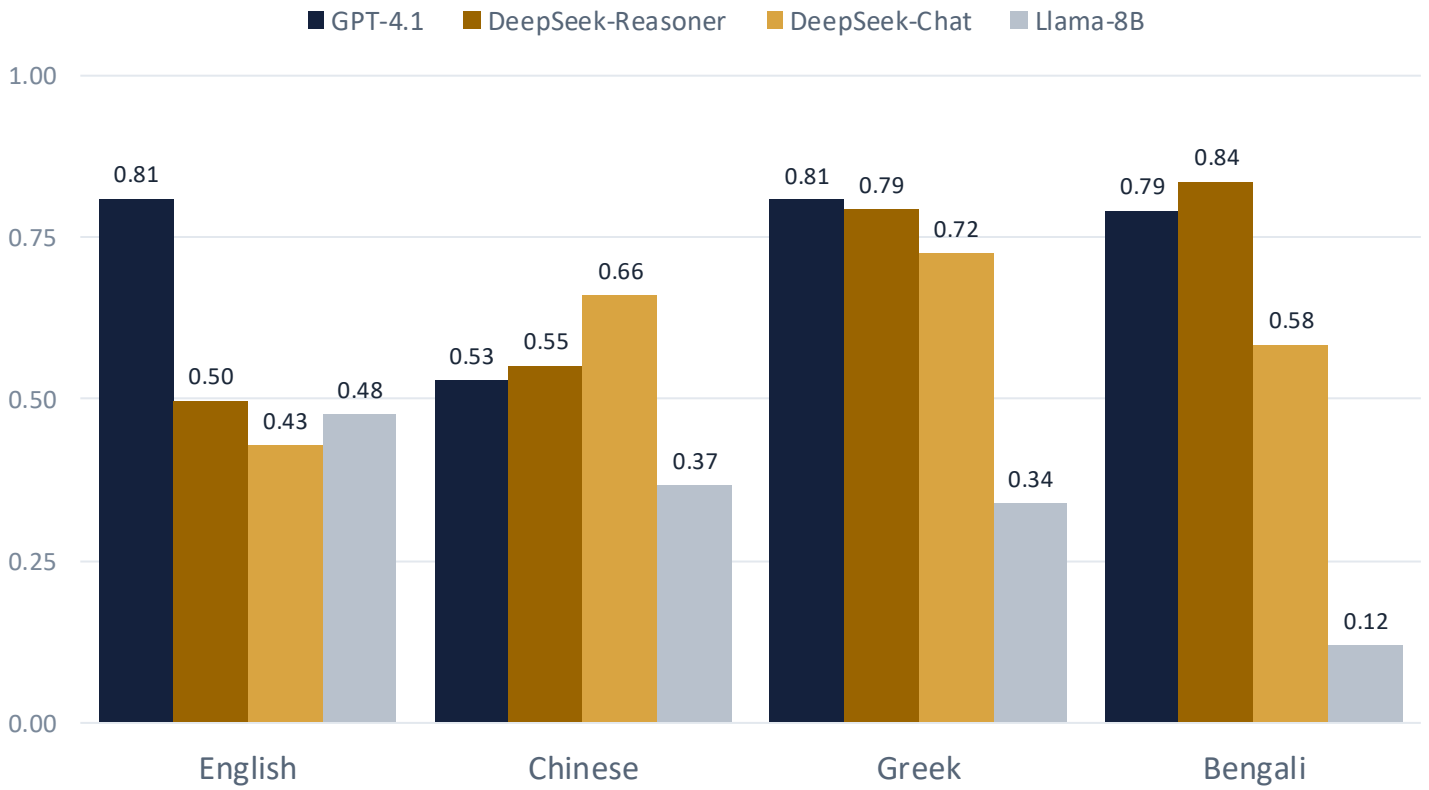
5.5–11.4 points — average size of the change over all 22 models

Not random — American identities score higher, Chinese identities lower

Role flips it — Islam-linked identities: down as retail investors, up as company owners

Figure 5 (change in F1 on true claims vs. no story) and Section 4.4, Liu et al., Findings of ACL 2026

Even with no story, the language matters



84%

right by always saying “false”, so we use F1 instead of plain accuracy

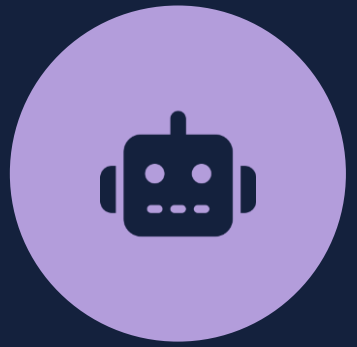
0.48 → 0.12

Llama-8B, from English to Bengali

0.84 vs 0.58

Bengali: DeepSeek with and without reasoning

F1 score with no story. Table 5, Liu et al., Findings of ACL 2026



PART 3 · ACT

Can it finish the job?

HERCULEAN · Peng et al., arXiv:2605.14355 (v4)

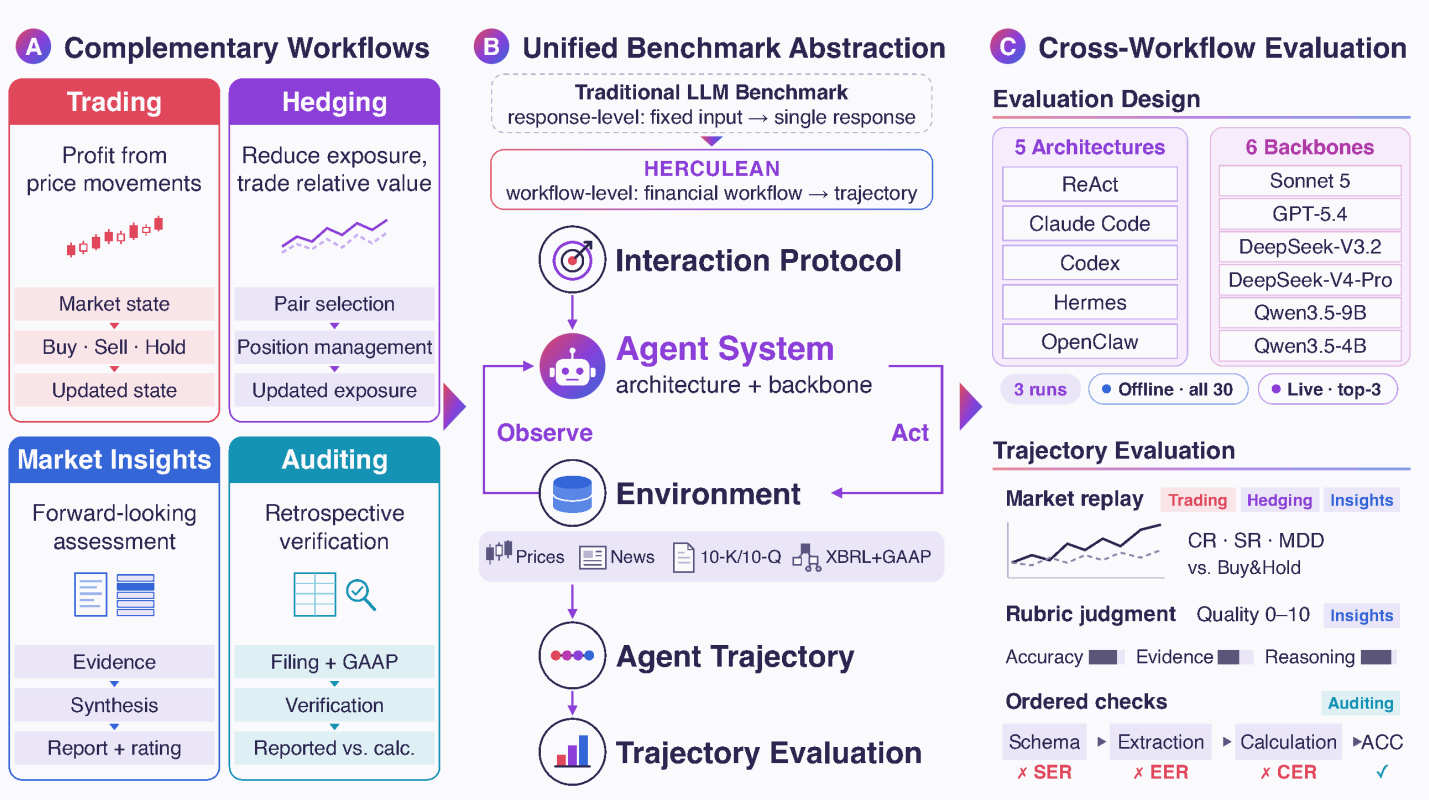
4 financial workflows · 5 agent architectures × 6 backbone LLMs · 3 runs each · live test Aug–Sep 2026

IN THIS PART

1 Four jobs
2 No single winner
3 Markets
4 Reports
5 Auditing

From answering questions to doing the work

FROM THE PAPER · Figure 1



Trading
Buy, sell or hold one stock daily · return and risk

Hedging
Pick 1 of 28 pairs, manage it daily · return and risk

Market Insights
Weekly report + rating · report quality; returns from ratings

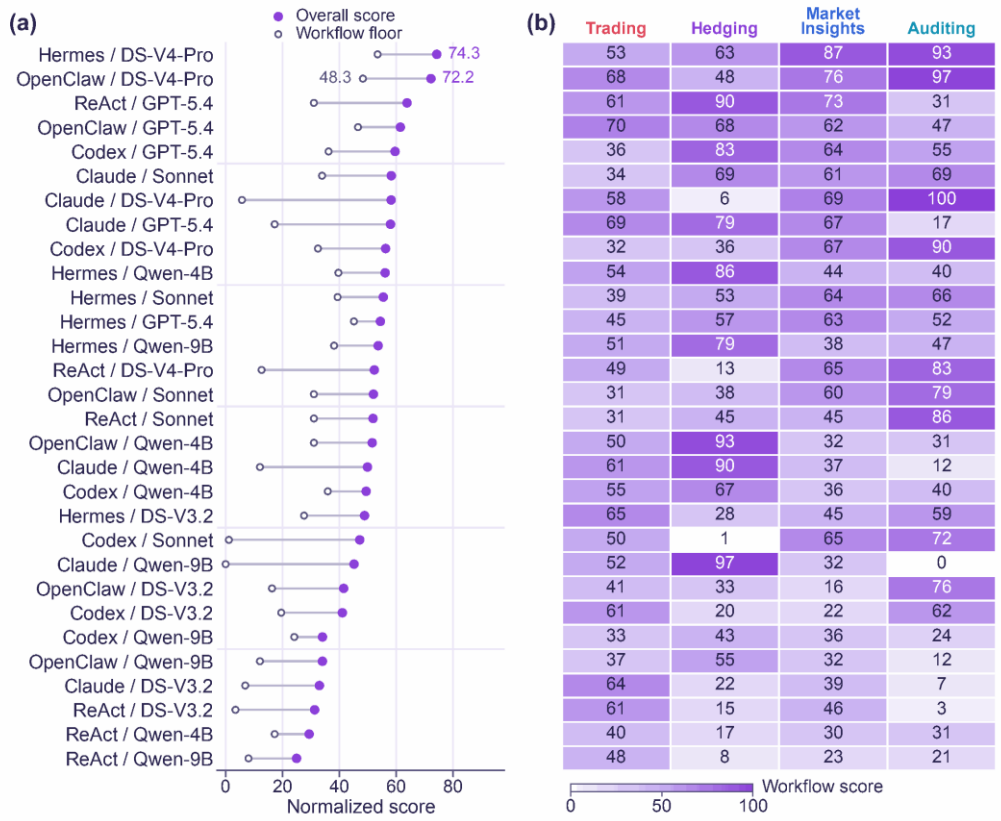
Auditing
Check a filed number (67 cases) · share correct

5 × 6 = 30 agent architectures × backbone LLMs
3 runs each · same tools and rules for all · ~$5,500

Figure 1, Sections 3–4, Peng et al., arXiv:2605.14355v4 · offline test Jan–Mar 2026; live test Aug–Sep 2026

No configuration wins every workflow

FROM THE PAPER · Figure 2



74.3 → 53.5

best overall configuration (Hermes + DeepSeek-V4-Pro), and its score in its weakest workflow, Trading

–0.23 to 0.38

agreement between the rankings of any two jobs (Kendall τ): weak

"FRAGMENTED CAPABILITY"

Each job has a different leader. Success in one financial workflow says little about the others.

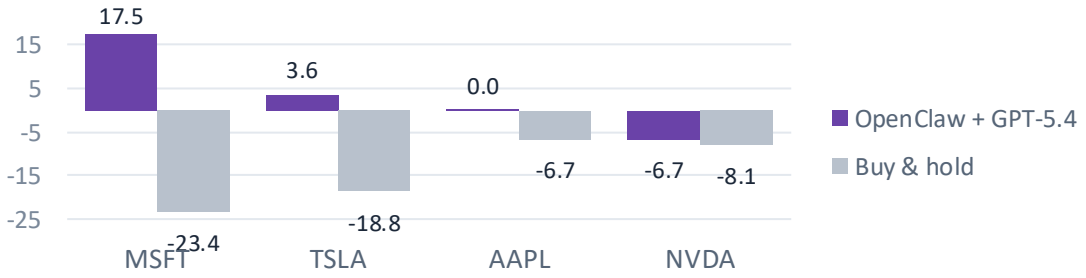
Figure 2 and Section 4.2.1, Peng et al., arXiv:2605.14355v4 · scores are percentile ranks among the 30 configurations (0 = worst, 100 = best)

Beating a falling market is easy; making money reliably is not

FROM THE PAPER · Table 3 · Sharpe ratio (return per unit of risk)

	Trading					Hedging				
	ReAct	Claude Code	Codex	Hermes	OpenClaw	ReAct	Claude Code	Codex	Hermes	OpenClaw
Sonnet 5	−1.3±.3	−1.4±.3	−0.7±.3	−1.2±.1	1.3±.2	−1.3±1.9	−0.2±2.9	−2.6±.8	−0.5±2.5	−1.2±2.1
GPT-5.4	−0.1±.1	0.4±.9	−1.2±1.1	−0.6±.8	**0.7**±.1	0.9±1.8	1.0±1.8	−0.1±1.9	−1.2±.9	−0.4±1.7
DS-V3.2	−0.2±1.0	−0.3±.9	−0.1±.0	−0.1±.6	−1.1±1.3	−2.1±.4	−1.7±.4	−1.7±1.5	−1.7±1.4	−1.4±1.2
DS-V4-Pro	−0.7±.2	−0.4±.4	−1.5±.4	−0.3±.4	0.3±.8	−1.9±.2	2.0±.6	−0.9±1.0	−0.2±1.8	−0.8±1.4
Qwen-9B	−1.2±1.2	−1.2±.9	−1.6±.7	−1.1±.6	−1.6±.7	−1.8±1.4	**1.7**±.4	−1.2±.9	0.8±2.1	−0.7±.7
Qwen-4B	−1.9±.9	−1.1±.2	−1.2±.7	−1.4±.9	−1.5±1.0	−1.9±1.6	1.1±.6	−0.5±1.1	0.7±1.8	1.4±1.0

① TRADING · best configuration (OpenClaw + GPT-5.4) vs. buy-and-hold, return %



② HEDGING

7 of 30

configurations made money; returns −9.9% to +7.0%

LIVE TEST · AUG–SEP 2026

Market rose (buy-and-hold +7.9% to +15.7%); the top trading agents mostly lagged it.

Table 3, Tables 5–10 and Section 4.2.1, Peng et al., arXiv:2605.14355v4 · offline window Jan–Mar 2026 · no configuration was profitable on all four stocks

Great reports, poor trading decisions

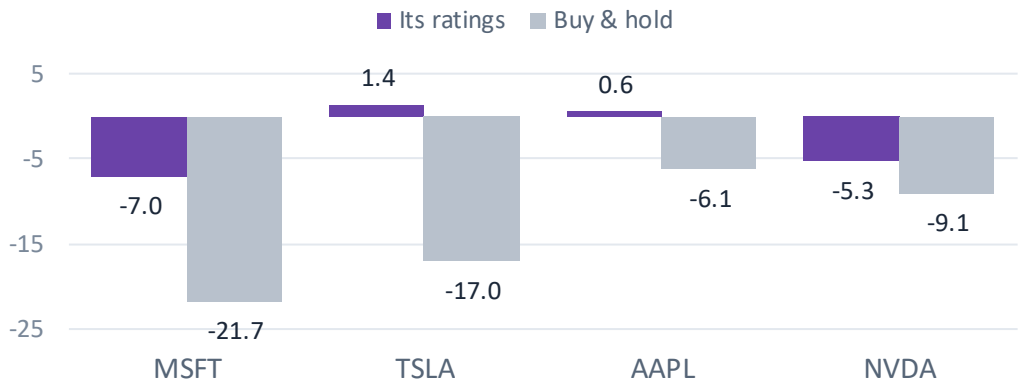
FROM THE PAPER · Table 3 · report quality, 0–10

Market Insights

	ReAct	Claude Code	Codex	Hermes	OpenClaw
Sonnet 5	8.2±.1	8.4±.1	8.4±.1	8.4±.1	8.5±.0
GPT-5.4	8.5±.0	8.6±.1	8.4±.1	8.5±.1	8.4±.0
DS-V3.2	7.1±.2	4.6±.4	4.0±.7	7.9±.2	6.2±.3
DS-V4-Pro	8.6±.1	8.7±.1	8.4±.0	**8.8**±.1	**8.8**±.0
Qwen-9B	5.4±2.9	6.7±.9	7.4±.1	7.0±.4	6.4±.2
Qwen-4B	6.8±.7	6.2±.8	7.1±.1	7.3±.2	6.5±.1

Mean ± std over 3 runs. Best scores come with DeepSeek-V4-Pro.

② Trading on the best writer's ratings (OpenClaw + DeepSeek-V4-Pro), return %



8.8 / 10

① best report quality, yet its ratings average −2.6%

NO LINK

Report quality does not predict the rating's return (β = −0.25, p = 0.19). A second judge agreed on 93% of items.

Table 3, Table 9, Section 4.2.2 and Appendix G.2, Peng et al., arXiv:2605.14355v4 · judge: grok-4.7; second judge: gemini-3.8-flash

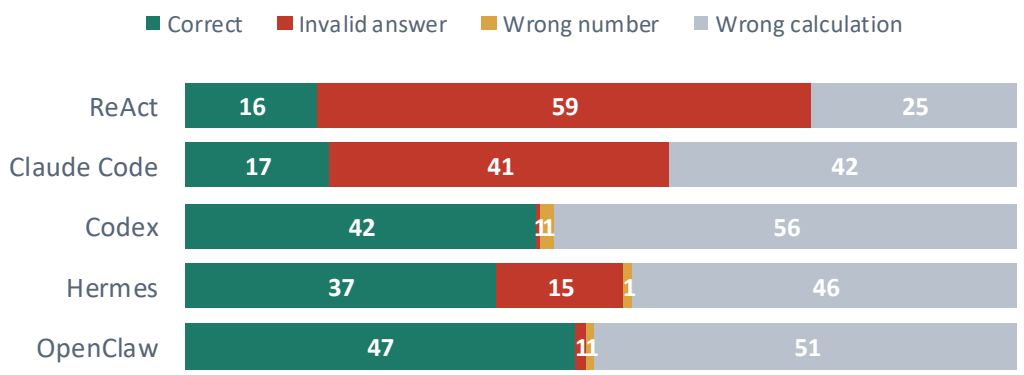
Same LLM, different agent architecture: 16% vs. 47%

FROM THE PAPER · Table 3 · auditing accuracy, % correct

	Auditing				
	ReAct	Claude Code	Codex	Hermes	OpenClaw
Sonnet 5	49.8±2.3	44.3±1.7	45.3±.9	43.8±3.8	48.8±3.1
GPT-5.4	28.9±3.1	26.4±4.6	33.3±3.1	32.8±3.0	30.8±1.7
DS-V3.2	15.9±2.3	17.4±1.7	42.3±4.6	37.3±3.0	46.8±2.3
DS-V4-Pro	49.3±3.0	**54.7**±3.8	50.7±.0	51.7±.9	52.7±.9
Qwen-9B	28.4±12.9	13.9±.9	28.4±2.6	30.8±.9	25.4±.0
Qwen-4B	28.9±10.8	25.4±.0	29.4±2.3	29.4±4.8	28.9±2.3

Mean ± std over 3 runs, 67 audit cases from real SEC filings.

② DeepSeek-V3.2 in every row · outcome of the 67 cases (%)



54.7%

① best of all 30 configurations (Claude Code + DeepSeek-V4-Pro)

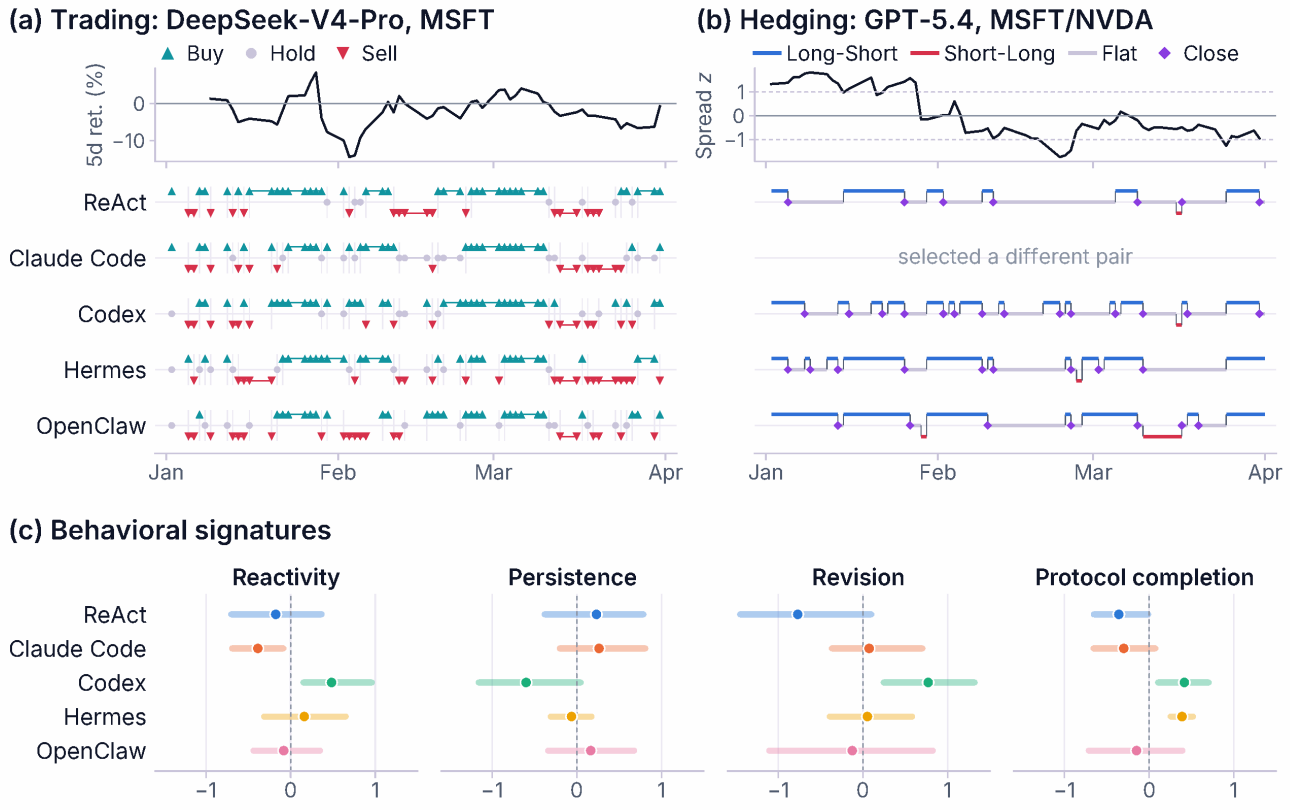
16% → 47%

② same backbone LLM, only the agent architecture changes

Table 3, Tables 5–9 and Section 4.2.2, Peng et al., arXiv:2605.14355v4 · outcomes add up to 100% per row

Each agent architecture fails in its own way

FROM THE PAPER · Figure 3



33.7% hold
Claude Code: the most cautious trader

26.5% hold
Codex: the most active; revises ratings most

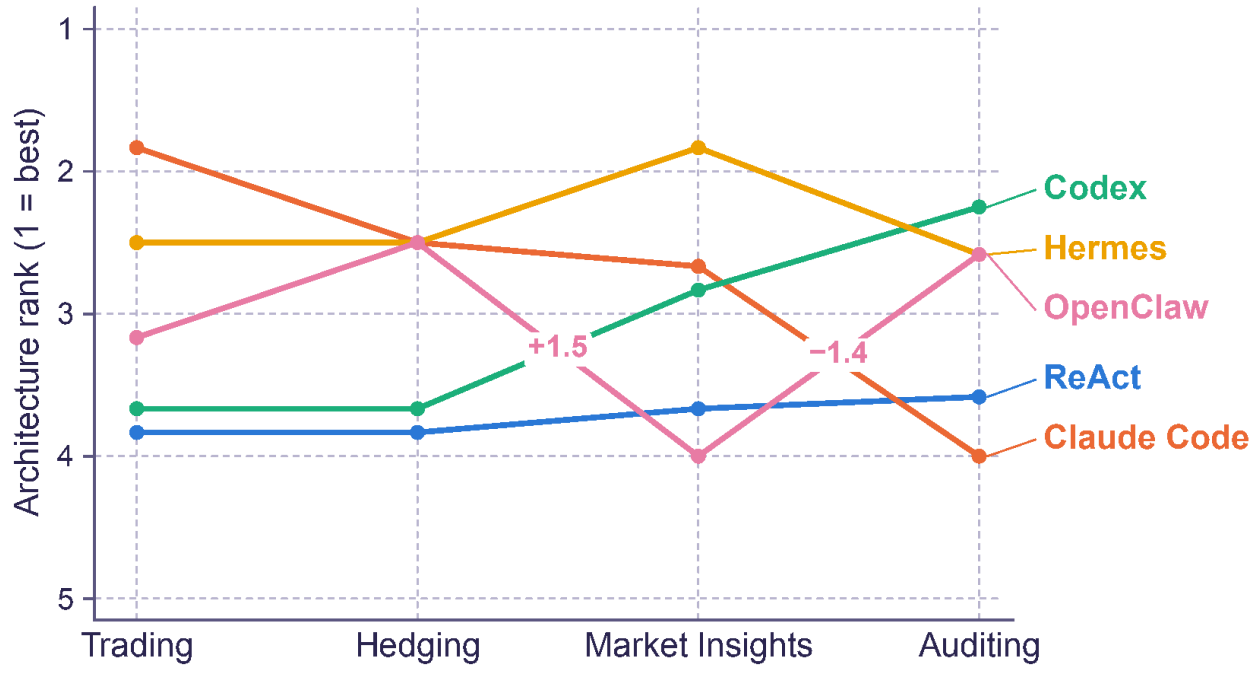
0.29–0.32 vs −0.07
all five follow the past 5 days, not the next day

Final scores hide very different ways of failing.

Figure 3 and Section 4.2.2, Peng et al., arXiv:2605.14355v4 · panel (a): DeepSeek-V4-Pro on MSFT · correlations are Spearman ρ

The workflow re-ranks the agent architectures

FROM THE PAPER · Figure 4 · mean rank within each model (1 = best)



Claude Code
1st in trading (1.83) → last in auditing (4.00)

Codex
near last in trading and hedging (3.67) → 1st in auditing (2.25)

16 of 36
model–job pairs where the agent ordering reverses

96.6 → 0.0
Claude Code + Qwen3.5-9B: hedging vs. auditing

Figures 2 and 4, Section 4.2.3, Peng et al., arXiv:2605.14355v4

The same gap, three times

	WITH HELP	WITHOUT HELP	WHY IT FAILS
READ	**97.7%** original shown	**53.6%** paragraph alone (coin flip: 50%)	*The model goes along with the text*
JUDGE	**Baseline** neutral question	**up to 28 pts** score swing when a story is added	*The story changes the answer*
ACT	**8.8/10** best report quality	**7 of 30** configurations profitable at hedging; auditing ≤ 55%	*Strengths don't carry over between jobs*

Each time, the model is missing the same thing: a way to check its answer against something it can trust.

Jiang et al., ACL 2026 (Table 6) · Liu et al., Findings of ACL 2026 (Fig. 5) · Peng et al., arXiv:2605.14355v4 (Table 3)

What would close the gap

FROM RFC-BENCH

Give it something to compare

Look things up, reason step by step, and ask “does this make sense?” before accepting a text.

FROM MFMD-SCEN

Compare with a neutral answer

Check each answer against the answer with no story, and fix the gap before use.

FROM HERCULEAN

Build better agent software

Agents that track their work and check numbers exactly, tested across several jobs, not just one.

FOR EVERYONE WHO BUILDS TESTS

Test with and without help

Report both numbers: a single number hides the gap.

Conclusions and future-work sections of the three papers

ACKNOWLEDGEMENTS

Thank you to the teams behind this work

RFC-Bench

Yuechen Jiang, Zhiwei Liu, Sophia Ananiadou and colleagues

MFMD-Scen

Zhiwei Liu, Jimin Huang, Sophia Ananiadou and colleagues

HERCULEAN

Xueqing Peng, Zhuohan Xie, Jimin Huang and more than 60 co-authors

PARTNER INSTITUTIONS

University of Manchester · The Fin AI · Stevens Institute of Technology · Columbia University · Yale University · New York University · University of Florida · MBZUAI · McGill University and Mila · AIST Japan · Halmstad University

WITH HALMSTAD UNIVERSITY | THE FIN AI FINANCIAL ARTIFICIAL INTELLIGENCE | NaCTeM The National Centre for Text Mining

Compute and funding include the NVIDIA Academic Grant Program (A100 GPU-hours on Brev), the University of Manchester, and project MIS 5154714 (Greece 2.0, EU NextGenerationEU).

TAKEAWAYS

So, can we trust LLMs in finance?

Not yet. Not without a way to check their answers.

READ — Without the original, LLMs are no better than a coin flip at spotting manipulated financial news.

JUDGE — The same claim gets a different answer depending on who is asking and in which language.

ACT — No agent is best at every job: strong reports, but rarely profitable hedging or reliable audits.

Thank you. Questions?

prayag.tiwari@hh.se

HALMSTAD UNIVERSITY | THE FIN AI — FINANCIAL ARTIFICIAL INTELLIGENCE | NaCTeM — The National Centre for Text Mining

Scan for the papers, data and code

RFC-Bench

MFMD-Scen

HERCULEAN

Data & code

Limitations the papers report

RFC-Bench

- English U.S. news only
- Text only: no tables, charts or audio
- Rewrites may miss real-world tactics
- Paragraphs judged alone

MFMD-Scen

- Few true claims: 23 of 144
- Only two human annotators for some regions
- Scenarios are test probes, not real users

HERCULEAN

- Four jobs, large U.S. stocks only
- No trading costs or delays
- Backtests can’t rule out memorised data; the live test helps
- Report checklists are LLM-written, not checked by people

Limitations sections of the three papers

Which models were tested

RFC-Bench · 15 LLMs

LLaMA 3.1 (8B, 70B) · Qwen3 8B, 14B, 32B (thinking and non-thinking) · Qwen2.5-72B · FMDLlama · GPT-4.1 · GPT-5 Mini · GPT-5.2 · DeepSeek-chat · DeepSeek-reasoner

MFMD-Scen · 22 LLMs

GPT-5-mini · Claude Sonnet 4.5 · Gemini 2.5 Flash · DeepSeek-V3.2 (reasoner, chat) · Qwen3 8B, 14B, 32B (with and without reasoning) · GPT-4.1 · Claude 3.5 Haiku · Gemini 2.0 Flash · Qwen2.5-72B · Llama 3.3-70B · Mistral 7B, Large, NeMo, Small 24B · Mixtral 8x7B, 8x22B

HERCULEAN · 5 × 6

Agent architectures: ReAct, Claude Code, Codex, Hermes, OpenClaw

Backbone LLMs: Claude Sonnet 5, GPT-5.4, DeepSeek-V3.2, DeepSeek-V4-Pro, Qwen3.5-9B, Qwen3.5-4B

Three runs per configuration

Table 5 (Jiang et al.), Section 3.1 (Liu et al.), Section 4.1 (Peng et al., v4)

How the models were asked

RFC-Bench · Task 1

"You are a financial misinformation detector. Please check if the following information is false and output the answer [true/false]. Return only one label."

RFC-Bench · Task 2

"… Given a pair of original news and misinformation, identify the misinformation type: [numerical / flipping / sentiment / causal]. Return only one label."

MFMD-Scen

"Please take the scenario information into account. Determine whether the claim is 'True' or 'False'. Scenario: [scenario] Claim: [claim]"

HERCULEAN

Each job is a skill with a fixed protocol: one call per day, week or case, a fixed output format, and the same tools for every agent.

Figure 2 (Jiang et al.), Appendix E (Liu et al.), Appendix F (Peng et al., v4)

References

1 Y. Jiang, Z. Liu, Y. Cao, Y. He, Z. Xu, C. Xu, Z. Deng, P. Tiwari, X. Chen, A. Lopez-Lira, J. Huang, J. Tsujii, S. Ananiadou. All That Glisters Is Not Gold: A Benchmark for Reference-Free Counterfactual Financial Misinformation Detection. ACL 2026, pp. 10737–10776.

2 Z. Liu, Y. Cao, Y. Jiang, et al., P. Tiwari, A. Lopez-Lira, X. Liu, J. Huang, S. Ananiadou. Same Claim, Different Judgment: Benchmarking Scenario-Induced Bias in Multilingual Financial Misinformation Detection. Findings of ACL 2026, pp. 9838–9864.

3 X. Peng, Z. Xie, Y. Cao, et al., P. Tiwari, et al., J. Tsujii, J.-Y. Nie, S. Ananiadou. HERCULEAN: An Agentic Benchmark for Financial Intelligence. arXiv:2605.14355 (v4, 19 Oct 2026).