Today's Number, Yesterday's Claim: How Tool Design Shapes LLM Checks of Revised Statistics

Suprabhat Rapolu SUPRABHAT@DIRMACS.COM
Dirmacs Private Limited (dirmacs.com)

Abstract

Official statistics are revised after publication, so a claim that matched the record when made can contradict today's figures. We test how the time index of a data tool shapes LLM claim checking. On 322 news claims about US jobs, earnings and GDP that were accurate when published but conflict with the latest data, Llama-3.1-8B with a tool returning figures as of the claim's date beat DeepSeek-V4-Pro without tools (64.5% vs 51.4%), whose reported training cutoff predates every claim (preregistered; Holm $p = 0.0009$). Without data, the open models answered CONTRADICTED rather than UNVERIFIABLE; given only today's figures, models rejected accurate claims more often still, though for Llama this depended on the serving host. A frontier model trained into the period does better without tools only before its knowledge cutoff (exploratory). With the as-of tool, most one-vintage claims check correctly; most remaining errors involve revisions, which need two vintages. Agents that check data need time-indexed tools.

Keywords: tool use; small language models; fact verification; temporal grounding; data revisions; agent evaluation

1. Introduction

Agents that check claims against data usually fetch *the* value of a statistic. Official statistics, however, are revised after publication (Croushore and Stark, 2001; Orphanides, 2001). A claim that reported a figure correctly can therefore disagree with the figure as it stands today. A preregistered companion study of 19 US releases (Rapolu, 2026) found that 332 of 1,862 checkable news claims supported by the record at publication (17.8%) are contradicted by the latest vintage; 43.3% for payroll changes and 0% for CPI.

Which vintage a claim should be judged against is a property of the verification setup rather than obviously of model scale, which makes it a clean test of the workshop's question: *does a small open-weight model with a tool that returns figures as published on a given date check such claims more accurately than a frontier-scale model without tools*, and what does a tool that returns only today's figures do instead? Prior work on automated fact-checking (Thorne et al., 2018; Guo et al., 2022) and on numerical and time-series claims (V et al., 2024; Fu et al., 2024; Strong and Vlachos, 2025) evaluates against one current record; temporal verification orders evidence in time (Barik et al., 2024) and finance work targets look-ahead bias (Benhenda, 2026; Lopez-Lira et al., 2025), but none treats revision of the record, and tool-use and small-agent work (Schick et al., 2023; Yao et al., 2023; Belcak et al., 2025) has not examined a tool's time index.

© 2026 S. Rapolu.

Contributions. (1) A preregistered census evaluation: 2,132 claims (pilot excluded), three tool conditions, four models (open, 8B to 1.6T; one closed). (2) With a tool returning figures as published on the claim date, Llama-3.1-8B beats DeepSeek-V4-Pro without tools, whose reported cutoff predates the claims (64.5% vs 51.4%, Holm $p = 0.0009$); without data the open models answer CONTRADICTED rather than UNVERIFIABLE, and a tool returning only today's figures makes models reject accurate reporting more often (for Llama, depending on the serving host); a frontier model trained into the period does better without tools only before its cutoff (exploratory). (3) Remaining errors are mostly revision claims, which need two vintages: a gap in our tool interface with a direct fix.

2. Task, data and setup

Claims and record. We reuse the companion study's claims (Rapolu, 2026): 4,669 claims from 334 news documents on the 19 preregistered releases of January to August 2026 (every Employment Situation and CPI release and three GDP advance estimates), each linked by the audit's pipeline to a series, reference period and measure. The record is the ALFRED vintage archive (Federal Reserve Bank of St. Louis, 2026) for eight series.

Populations and reference labels. **FLIP** (332 claims): every claim the record supported on its publication date and contradicts in its latest vintage. **STABLE** (1,529): supported on both dates. **HOLDOUT** (291): every checkable claim of two releases held out from the audit's rule development, labelled by blind (LLM) adjudication; 75 of its labels are CONTRADICTED or UNVERIFIABLE, so it penalises a verifier that agrees with everything. The question is whether a claim was accurate by the record *available when it was made*, so the FLIP and STABLE reference is SUPPORTED. It rests on the audit's automated *supported* verdict, which agreed with blind adjudication on 142 of 153 unseen-release claims (0.928; 95% CI 0.876–0.959); its *contradicted* verdict (precision 0.167) is never used.

Conditions. Each model receives the claim, its publication date, series, reference period, measure and, for claims about forecasts, the consensus figure, and answers SUPPORTED, CONTRADICTED or UNVERIFIABLE with one sentence. **C0**: no tool. **C1**: `get_values(series, start, end)` returns the latest vintage. **C2**: the same with `as_of`, returning the vintage in force on that date; the prompt says to use the publication date. Both tools read only the local archive and return the level and the period, percent and year-on-year changes within one vintage; at most three calls. The prompt states the scoring rule: compare at the claim's precision and match direction words to the sign (Appendix A).

Models. **A**: Llama-3.1-8B-Instruct (8B; knowledge cutoff Dec. 2023). **B**: gpt-oss-20b (21B total, 3.6B active; Jun. 2024). **F**: DeepSeek-V4-Pro (1.6T total, 49B active; cutoff not stated by the developer, May 2025 in third-party listings). **S**: Claude Sonnet 5.5 (closed; Jun. 2026, inside the claim window). A, B and F run the full design; S runs FLIP and HOLDOUT. Calls go through OpenRouter with native function calling at temperature 0 and seed 2027, except that S accepts neither and F's pinned hosts accept no seed. F ran on two pinned fp8 hosts with reasoning off, B on one, and A across hosts (Appendix B).

Protocol. Populations, hypotheses and tests were frozen, with a hash of the claim list, before any model call. A 20-claim pilot then fixed harness bugs and added one line to the

Table 1: Main results (pilot excluded; %). FCR, UNV: share of FLIP claims (reference SUPPORTED) labelled CONTRADICTED, UNVERIFIABLE; brackets: 95% CI. $/100: US$ per 100 FLIP claims. –: not run; †: over 5% of episodes missing. First row: post-hoc baseline (D6); 100 on FLIP and STABLE by definition.

Model	Cond.	FLIP acc.	FLIP FCR	FLIP UNV	STABLE acc.	HOLDOUT acc.	$/100
Majority baseline	always SUPPORTED	100	0	0	100	74.2	0
Llama-3.1-8B	C0 no tool	46.6	49.4 [44.4, 54.5]	2.8	40.9	37.5	0.002
	C1 latest	29.8	66.8 [61.8, 71.8]	2.8	53.3	49.0	0.028
	C2 as-of	64.5	30.2 [25.2, 35.1]	3.7	55.0	45.5	0.030
gpt-oss-20b	C0 no tool	31.7	46.9 [40.5, 53.0]	21.4	25.5	30.9	0.002
	C1 latest	2.2	92.9 [89.4, 95.8]	5.0	67.2	58.1	0.004
	C2 as-of	75.8	19.6 [15.4, 23.7]	4.7	69.5	61.9	0.005
DeepSeek-V4-Pro	C0 no tool	51.6	38.8 [33.8, 44.2]	9.6	49.9	41.6	0.007
	C1 latest	0.9	59.3 [54.5, 64.2]	39.1	70.2	62.5	0.032
	C2 as-of	87.0	8.1 [5.0, 11.2]	3.4	76.1	63.6	0.029
Claude Sonnet 5.5	C0 no tool	77.6	2.5 [0.9, 4.3]	19.3	–	47.1	0.526
	C1 latest	46.9	17.4 [13.4, 21.6]	32.3	–	66.0	1.287
	C2 as-of	95.3	3.7 [1.8, 5.8]	0.6	–	69.1	0.995

prompt (keep reasoning brief); pilot claims are excluded from all results, and every later change is logged (Appendix B). **H1**: on FLIP, A-C2 is more accurate than F-C0. **H2**: for each model, C1 gives more false contradictions on FLIP than C0. **H3**: C2 gives fewer than C1. **H4**: on HOLDOUT, C2 is no less accurate than C0. Tests: exact McNemar on paired claims, Holm-corrected over H1–H3 for A and F; 95% CIs: bootstrap over source documents. A missing verdict is scored as wrong.

3. Results

H1 (confirmatory): supported, with a caveat. (Pilot excluded: FLIP $n = 322$; 3 episodes missing after retries, Appendix C.) On FLIP, Llama-3.1-8B with the as-of tool (A-C2) was correct on 64.5% of claims, against 51.4% for DeepSeek-V4-Pro without tools (F-C0; 51.6% on all 322 claims, Table 1; $n = 321$ pairs, Holm-adjusted exact McNemar $p = 0.0009$). F's reported training cutoff predates every release studied, so F-C0 cannot know these figures; it nonetheless labels 38.8% of FLIP claims CONTRADICTED and only 9.6% UNVERIFIABLE. Including pilot claims: 64.7% vs 50.5% ($n = 331$, $p = 0.0003$).

H2 and H3 (confirmatory). By construction a latest-value tool cannot support FLIP claims; the question is whether a model abstains or rejects. Both mostly reject: CONTRADICTED rises from 49.4% to 66.8% for A and from 38.8% to 59.3% for F (H2; Holm $p < 10^{-4}$ and $p < 10^{-4}$), with UNVERIFIABLE at 2.8% and 39.1%. The as-of tool cuts false contradictions to 30.2% and 8.1% (H3; Holm $p < 10^{-4}$ and $p < 10^{-4}$). gpt-oss-20b (B; 3.6B active parameters, secondary) follows the same pattern: 46.9% CONTRADICTED with no tool, 92.9% with C1 and 19.6% with C2, where its accuracy is 75.8%.

H4: C2 is no less accurate than C0. On HOLDOUT, whose labels are mixed, C2 minus C0 (paired, 95% document-bootstrap CI) is +22.0 points for F (CI [+13.0, +29.3];

63.6% vs 41.6%), S (+22.0 [+9.7, +33.3]) and B (+30.9 [+21.1, +38.1]); for A, C2 is +7.9 points higher ([-0.3, +16.2]; $p = 0.056$), not significant. No model reaches the 74.2% that always answering SUPPORTED scores on HOLDOUT (post-hoc; Appendix J). On STABLE, C2 reaches 55.0% (A) and 76.1% (F), against 40.9% and 49.9% without tools.

Exploratory: a frontier model's memory. Claude Sonnet 5.5 (S), whose stated cutoff is inside the claim window, scores 77.6% on FLIP without tools, above A-C2 ($p = 0.0004$). By release date, S-C0 is correct on 89.6% of claims up to its cutoff ($n = 249$) and 37.0% after it ($n = 73$), where A-C2 is ahead (61.6% vs 37.0%, $p = 0.0096$); the split is descriptive (the sides also differ in month and series mix) but matches evidence of recall up to, not after, a cutoff (Lopez-Lira et al., 2025). After its cutoff S mostly abstains (0.0% CONTRADICTED). C1 raises S's false contradictions from 2.5% to 17.4% ($p < 10^{-4}$).

Where the as-of tool fails: revision claims. 82 FLIP claims report a revision ("December was revised down by 2,000"), which needs two vintages. Under C2 they score 43.9% (A), 65.9% (F) and 90.2% (S), against 71.5%, 94.2% and 97.1% on other claims. Our prompt says to query the publication date; F and S still queried two or more vintages on 53.7% and 57.3% of revision claims, A on 0.0%. Among 20 random C2 FLIP failures each for A and F (Appendix H), the leading cause is a revision claim checked against one vintage (A 10, F 4). 1 of the 40 is a plausible label error and 2 are ambiguous: label error lowers measured C2 accuracy by an estimated 1.8 points (A) and 0.0–1.3 (F).

Rerun of A on one host (post-hoc). A was routed across hosts, so after review we reran it on one host (Appendix E). H1 holds (67.4% vs 51.6%, unadjusted $p < 10^{-4}$) and so does H3 (58.4% to 28.0%), but H2 does not for A: there it already contradicts 55.9% of FLIP claims with no tool, against 58.4% with C1 ($p = 0.551$). The October runs already differed by host: episodes served only by Groq contradicted 50.6% of FLIP claims with no tool ($n = 81$) and 75.4% with C1 ($n = 134$); hosts were not randomly assigned.

Per 100 claims, A-C2 cost US$0.030 and F-C2 $0.029, against $0.526 for S-C0 and $0.995 for S-C2.

4. Discussion and limitations

The tool decides most of the outcome here: without figures, models reject what they cannot check (F) or are right mainly where training covers the period (S); with today's figures, models reject accurate reporting more often; with the figures as published, an 8B model with the right data beats a 1.6T model without it. For tool builders: make the vintage a first-class tool argument and return revision history for revision claims.

Limitations. (1) One tool/prompt/model/serving setup on three US release families; not a general design-beats-scale result. (2) The prompt pre-links each claim to its series, period and measure, testing grounded-metadata verification, not end-to-end parsing; FLIP/STABLE labels come from the audit's pipeline (0.928 aggregate validation), not human-coded. (3) The as-of tool reads the archive that produced the FLIP labels, so H1–H3 are partly by construction; HOLDOUT does not show that models separate accurate from inaccurate claims (Appendix J). (4) A's results depend on its serving host (Appendix E).

References

Anab Maulana Barik, Wynne Hsu, and Mong Li Lee. Evidence-based temporal fact verification. *arXiv preprint arXiv:2407.15291*, 2024.

Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. Small language models are the future of agentic AI. *arXiv preprint arXiv:2506.02153*, 2025.

Mostapha Benhenda. Look-ahead-bench: a standardized benchmark of look-ahead bias in point-in-time LLMs for finance. *arXiv preprint arXiv:2601.13770*, 2026.

Dean Croushore and Tom Stark. A real-time data set for macroeconomists. *Journal of Econometrics*, 105(1):111–130, 2001.

Federal Reserve Bank of St. Louis. ALFRED: ArchivaL Federal Reserve Economic Data. https://alfred.stlouisfed.org/, 2026.

Yu Fu, Shunan Guo, Jane Hoffswell, Victor S. Bursztyn, Ryan Rossi, and John Stasko. "The data says otherwise" – towards automated fact-checking and communication of data claims. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*, 2024.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022.

Alejandro Lopez-Lira, Yuehua Tang, and Mingyin Zhu. The memorization problem: Can we trust LLMs' economic forecasts? *arXiv preprint arXiv:2504.14765*, 2025.

Athanasios Orphanides. Monetary policy rules based on real-time data. *American Economic Review*, 91(4):964–985, 2001.

Suprabhat Rapolu. Misinformation about economic numbers is rare; mis-verification is common, 2026. Working paper.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, 2023.

Marek Strong and Andreas Vlachos. TSVer: A benchmark for fact verification against time-series evidence. In *Proceedings of EMNLP*, pages 29906–29926, 2025.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of NAACL-HLT*, pages 809–819, 2018.

Venktesh V, Abhijit Anand, Avishek Anand, and Vinay Setty. QuanTemp: A real-world open-domain benchmark for fact-checking numerical claims. In *Proceedings of SIGIR*, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Appendix A. Prompts and tool schema

The system prompt is shared by all conditions; `{CONDITION}` is replaced by one of the three condition texts below. Files are reproduced verbatim (prompt version v2, frozen after the pilot).

System prompt.

```
You check claims that news coverage made about US official statistics (the BLS Employment Situation,
the BLS Consumer Price Index, and the BEA GDP release).

The question is always: was this claim accurate according to the official record available when it
was made? Judge it against the figures as they were published on or before the claim's publication
date, not against figures published later.

Scoring rule:
- Compare at the precision the claim states. Round the official figure to that precision before
comparing. Example: "115,000 jobs" is compared to the nearest thousand; "0.3%" is compared at one
decimal place.
- Words such as "about", "nearly" or "roughly" allow the difference that the stated rounding implies.
- Direction words ("rose", "fell", "slowed", "accelerated", "beat", "missed") must match the sign or
the comparison in the record.
- Answer SUPPORTED if the record available at publication matches the claim, CONTRADICTED if it does
not, and UNVERIFIABLE only if that record cannot settle the claim.

{CONDITION}

Keep your reasoning brief: a few short sentences at most.

Finish with exactly one line of JSON and nothing after it:
{"verdict": "SUPPORTED" | "CONTRADICTED" | "UNVERIFIABLE", "reason": "<one sentence>"}
```

Condition text, C0 (no tool).

```
You have no tools. Answer from your own knowledge of the official figures.
```

Condition text, C1 (latest-value tool).

```
You have one tool, get_values(series, start_period, end_period), which returns the latest available
official figures for a series from an archive of the official record. You may call it at most 3 times
.
```

Condition text, C2 (as-of tool).

```
You have one tool, get_values(series, start_period, end_period, as_of), which returns official
figures for a series exactly as they were published on the date as_of, from an archive of the
official record. Set as_of to the claim's publication date. You may call it at most 3 times.
```

User message. Fields in braces are filled from the claim record; the consensus line appears only for claims about forecasts.

```
Claim: "{claim_text}"
Publication date: {as_of}
Series: {series} ({series_name}; units: {units})
```

```
Reference period: {ref}
Measure: {measure}{consensus_line}
```

Tool schema (C2). The C1 schema is identical without `as_of`, and describes the figures as the latest available.

```
{
 "type": "function",
 "function": {
  "name": "get_values",
  "description": "Official figures for a US statistical series exactly as published on a given date (
  the vintage in force on as_of). Returns, for each period from start_period to end_period: the value
  , the change from the prior period, the percent change from the prior month (monthly series) or the
   annualized growth rate (GDPC1), and the percent change from a year earlier, all as published on
  as_of.",
  "parameters": {
   "type": "object",
   "properties": {
    "series": {
     "type": "string",
     "enum": [
      "PAYEMS",
      "UNRATE",
      "CES0500000003",
      "CPIAUCSL",
      "CPIAUCNS",
      "CPILFESL",
      "CPILFENS",
      "GDPC1"
     ],
     "description": "Series code."
    },
    "start_period": {
     "type": "string",
     "description": "First period: YYYY-MM for monthly series, YYYYQn for GDPC1."
    },
    "end_period": {
     "type": "string",
     "description": "Last period, same format. Omit for a single period."
    },
    "as_of": {
     "type": "string",
     "description": "Date YYYY-MM-DD; the figures returned are those published on or before this date
     ."
    }
   },
   "required": [
    "series",
    "start_period",
    "as_of"
   ]
  }
 }
}
```

Appendix B. Code, protocol and models

Code, the frozen protocol (populations, hypotheses and tests, with the claim-list hash), the log of every later change and the aggregate results are at

https://github.com/suprabhatrapolu/revised-stats-tool-eval

Claim texts and labels come from the companion study and will be released with it.

Model (OpenRouter ID)	Knowledge cutoff (source)	Serving
`meta-llama/llama-3.1-8b-instruct`	Dec. 2023 (Meta model card)	routed across hosts; logged per request
`openai/gpt-oss-20b`	Jun. 2024 (OpenAI documentation)	one pinned host, fp8
`deepseek/deepseek-v4-pro`	not stated; May 2025 (third-party listings)	two pinned hosts, fp8; reasoning off
`anthropic/claude-sonnet-5.5`	Jun. 2026 (Anthropic)	default sampling (no temperature or seed)

Total spend for the October runs, including the pilot and superseded episodes: US$21.23. The post-review rerun (Appendix E) cost a further US$5.24, including stopped and failed requests.

Appendix C. Run integrity, missing episodes and truncation

All metrics are computed from per-episode result files; every cell's expected, completed and missing counts are in the `integrity.csv` file in the repository (Appendix B). Episodes completed per cell (pilot claims excluded), then episodes without a verdict:

Model	Cond.	FLIP	STABLE	HOLDOUT
Llama-3.1-8B	C0 no tool	322/322	1519/1519	291/291
Llama-3.1-8B	C1 latest	322/322	1519/1519	290/291
Llama-3.1-8B	C2 as-of	321/322	1519/1519	290/291
gpt-oss-20b	C0 no tool	322/322	1519/1519	291/291
gpt-oss-20b	C1 latest	322/322	1519/1519	291/291
gpt-oss-20b	C2 as-of	322/322	1519/1519	291/291
DeepSeek-V4-Pro	C0 no tool	322/322	1519/1519	291/291
DeepSeek-V4-Pro	C1 latest	322/322	1519/1519	291/291
DeepSeek-V4-Pro	C2 as-of	322/322	1519/1519	291/291
Claude Sonnet 5.5	C0 no tool	322/322	not run	291/291
Claude Sonnet 5.5	C1 latest	322/322	not run	291/291
Claude Sonnet 5.5	C2 as-of	322/322	not run	291/291

Model	Cond.	Subset	n	no verdict % (truncated %)	acc. %: scored / excl. truncated
Llama-3.1-8B	C0 no tool	FLIP	322	1.2 (0.3)	46.6 / 46.7
Llama-3.1-8B	C0 no tool	HOLDOUT	291	4.8 (0.3)	37.5 / 37.6
Llama-3.1-8B	C0 no tool	STABLE	1519	2.9 (0.4)	40.9 / 41.1
Llama-3.1-8B	C1 latest	FLIP	322	0.6 (0.0)	29.8 / 29.8
Llama-3.1-8B	C1 latest	HOLDOUT	290	1.0 (0.0)	49.0 / 49.0
Llama-3.1-8B	C1 latest	STABLE	1519	1.8 (0.1)	53.3 / 53.4
Llama-3.1-8B	C2 as-of	FLIP	321	1.6 (0.3)	64.5 / 64.7
Llama-3.1-8B	C2 as-of	HOLDOUT	290	3.4 (0.7)	45.5 / 45.8
Llama-3.1-8B	C2 as-of	STABLE	1519	1.6 (0.1)	55.0 / 55.0
gpt-oss-20b	C1 latest	STABLE	1519	0.5 (0.0)	67.2 / 67.2
gpt-oss-20b	C2 as-of	STABLE	1519	0.3 (0.0)	69.5 / 69.5
DeepSeek-V4-Pro	C0 no tool	STABLE	1519	0.1 (0.0)	49.9 / 49.9
DeepSeek-V4-Pro	C1 latest	FLIP	322	0.6 (0.6)	0.9 / 0.9
DeepSeek-V4-Pro	C1 latest	HOLDOUT	291	2.4 (0.3)	62.5 / 62.8
DeepSeek-V4-Pro	C1 latest	STABLE	1519	2.5 (0.4)	70.2 / 70.5
DeepSeek-V4-Pro	C2 as-of	FLIP	322	1.6 (0.0)	87.0 / 87.0
DeepSeek-V4-Pro	C2 as-of	HOLDOUT	291	2.1 (0.0)	63.6 / 63.6
DeepSeek-V4-Pro	C2 as-of	STABLE	1519	2.8 (0.1)	76.1 / 76.2
Claude Sonnet 5.5	C0 no tool	FLIP	322	0.6 (0.6)	77.6 / 78.1
Claude Sonnet 5.5	C0 no tool	HOLDOUT	291	0.7 (0.7)	47.1 / 47.4
Claude Sonnet 5.5	C1 latest	FLIP	322	3.4 (3.4)	46.9 / 48.6
Claude Sonnet 5.5	C1 latest	HOLDOUT	291	1.7 (1.4)	66.0 / 66.9
Claude Sonnet 5.5	C2 as-of	FLIP	322	0.3 (0.3)	95.3 / 95.6
Claude Sonnet 5.5	C2 as-of	HOLDOUT	291	0.7 (0.3)	69.1 / 69.3

A missing verdict is scored as wrong, as preregistered, and never as UNVERIFIABLE.

Appendix D. Tool-call reliability

Model	Cond.	calls	calls/claim	malformed %	error %	as_of = pub. %
Llama-3.1-8B	C1 latest	4,694	2.31	30.5	12.0	–
Llama-3.1-8B	C2 as-of	3,877	2.06	47.4	4.6	99.2
gpt-oss-20b	C1 latest	1,961	1.07	0.0	2.2	–
gpt-oss-20b	C2 as-of	1,970	1.07	0.0	1.3	99.8
DeepSeek-V4-Pro	C1 latest	2,158	1.16	0.0	4.2	–
DeepSeek-V4-Pro	C2 as-of	2,267	1.22	0.0	3.9	92.0
Claude Sonnet 5.5	C1 latest	328	1.02	0.0	0.0	–
Claude Sonnet 5.5	C2 as-of	372	1.16	0.0	0.0	86.8

Malformed: arguments not exactly one JSON object. For Llama-3.1-8B these are all exact duplicates (two identical objects concatenated; 3,269 of 3,269) and come from one serving host: 1,262 of 1,263 calls in episodes served only by that host, against 0 of 4,160 on the other. They are a serving artifact, not model behaviour; the harness used the first object, so no tool result changed. Error: the tool returned an error (bad series, period or date). as_of: share of C2 calls whose `as_of` equals the publication date. FLIP and STABLE, pilot excluded.

Appendix E. Llama-3.1-8B by serving host

Episodes served entirely by one host; hosts with $n \geq 30$. Hosts were not assigned at random, so differences may also reflect claim mix and time of request. Accuracy differs materially between hosts, so A's results average over hosts.

Cond.	Subset	Host	n	acc. %	contradicted %	unverifiable %
C0 no tool	FLIP	DeepInfra	120	61.7	37.5	0.0
C0 no tool	FLIP	Groq	81	44.4	50.6	3.7
C0 no tool	FLIP	Novita	118	32.2	61.0	5.1
C0 no tool	HOLDOUT	DeepInfra	118	45.8	41.5	4.2
C0 no tool	HOLDOUT	Groq	34	38.2	55.9	5.9
C0 no tool	HOLDOUT	Novita	117	31.6	58.1	4.3
C0 no tool	STABLE	DeepInfra	728	48.9	47.9	2.1
C0 no tool	STABLE	Groq	229	37.1	49.3	12.2
C0 no tool	STABLE	Novita	553	32.2	56.1	5.8
C1 latest	FLIP	CoreWeave	125	38.4	56.8	4.8
C1 latest	FLIP	Groq	134	21.6	75.4	1.5
C1 latest	HOLDOUT	CoreWeave	62	45.2	35.5	14.5
C1 latest	HOLDOUT	Groq	147	51.7	27.9	14.3
C1 latest	STABLE	CoreWeave	426	45.5	37.6	16.9
C1 latest	STABLE	Groq	683	57.7	37.5	4.2
C2 as-of	FLIP	CoreWeave	119	69.7	25.2	5.0
C2 as-of	FLIP	Groq	135	62.2	31.9	2.2
C2 as-of	HOLDOUT	CoreWeave	109	50.5	26.6	31.2
C2 as-of	HOLDOUT	Groq	61	45.9	37.7	9.8
C2 as-of	STABLE	CoreWeave	558	46.6	25.8	26.9
C2 as-of	STABLE	Groq	421	61.3	34.7	2.4
Episodes served by more than one host (excluded above): 1281						

Post-hoc note (D6): in the preregistered runs B and F, each on pinned fp8 hosts, show the same direction as A in every condition (C1 raises false contradictions, C2 lowers them, C2 beats C0).

Rerun of Llama-3.1-8B on one pinned host (post-hoc, D7). After review we reran A on every claim (three conditions; FLIP, STABLE and HOLDOUT; 6,396 episodes after excluding pilot claims) on one host with no fallback: CoreWeave, which serves unquantized (bf16) weights. Prompt, tool, seed, call limit and scoring are unchanged; 0 episodes are missing. The preregistered results above are not replaced.

Cond.	Run	FLIP acc.	FLIP FCR	STABLE acc.	HOLDOUT acc.	no verdict
C0 no tool	October, mixed hosts	46.6	49.4	40.9	37.5	2.9
	rerun, one host	32.9	55.9	28.2	27.8	7.8
C1 latest	October, mixed hosts	29.8	66.8	53.3	49.0	1.5
	rerun, one host	36.6	58.4	40.0	42.6	3.0
C2 as-of	October, mixed hosts	64.5	30.2	55.0	45.5	1.9
	rerun, one host	67.4	28.0	37.4	42.6	1.8

H1 recomputed with the rerun against the October F-C0 episodes: 67.4% vs 51.6% ($n = 322$, difference +15.8 points, 95% CI [+8.6, +23.4]; exact McNemar $p < 10^{-4}$, unadjusted). H3: false contradictions fall from 58.4% (C1) to 28.0% (C2), $p < 10^{-4}$. H4: C2 minus C0 on HOLDOUT is +14.8 points ([+7.5, +21.6]; $p = 0.0002$). H2 is not reproduced: 55.9% CONTRADICTED with no tool and 58.4% with C1 (difference +2.5 points, [-4.9, +9.9]; $p = 0.551$), because on this host A already rejects more claims without a tool. Levels differ between the two runs, most on STABLE, and the same claim gets the same verdict in both runs in 53.3% (C0), 67.0% (C1) and 69.0% (C2) of cases, so A's point estimates are host-dependent. Two things differ from October besides the host. We first pinned another host, which rejects tool calls that do not match the schema instead of returning them; that run was stopped unscored. CoreWeave does not accept `tool_choice: none`, so the final turn after three calls is sent without the tool definition. Both are logged (D7).

Appendix F. False contradictions on FLIP by series

Model	Cond.	payroll	avg. hourly earnings	GDP
Llama-3.1-8B	C0 no tool	48.9 (n=266)	46.4 (n=28)	57.1 (n=28)
Llama-3.1-8B	C1 latest	71.4 (n=266)	28.6 (n=28)	60.7 (n=28)
Llama-3.1-8B	C2 as-of	31.7 (n=265)	17.9 (n=28)	28.6 (n=28)
gpt-oss-20b	C0 no tool	51.1 (n=266)	35.7 (n=28)	17.9 (n=28)
gpt-oss-20b	C1 latest	94.0 (n=266)	92.9 (n=28)	82.1 (n=28)
gpt-oss-20b	C2 as-of	21.8 (n=266)	14.3 (n=28)	3.6 (n=28)
DeepSeek-V4-Pro	C0 no tool	38.0 (n=266)	17.9 (n=28)	67.9 (n=28)
DeepSeek-V4-Pro	C1 latest	56.4 (n=266)	100.0 (n=28)	46.4 (n=28)
DeepSeek-V4-Pro	C2 as-of	8.6 (n=266)	7.1 (n=28)	3.6 (n=28)
Claude Sonnet 5.5	C0 no tool	1.9 (n=266)	10.7 (n=28)	0.0 (n=28)
Claude Sonnet 5.5	C1 latest	13.2 (n=266)	71.4 (n=28)	3.6 (n=28)
Claude Sonnet 5.5	C2 as-of	3.0 (n=266)	10.7 (n=28)	3.6 (n=28)

Appendix G. Results by knowledge cutoff (exploratory)

Model	Cond.	cutoff (end)	releases	n	acc. %	FCR %
Llama-3.1-8B	C0 no tool	2023-12-31	after cutoff	322	46.6 [41.6, 51.1]	49.4
Llama-3.1-8B	C1 latest	2023-12-31	after cutoff	322	29.8 [25.1, 34.7]	66.8
Llama-3.1-8B	C2 as-of	2023-12-31	after cutoff	321	64.5 [59.5, 69.5]	30.2
gpt-oss-20b	C0 no tool	2024-06-30	after cutoff	322	31.7 [26.3, 37.5]	46.9
gpt-oss-20b	C1 latest	2024-06-30	after cutoff	322	2.2 [0.7, 3.9]	92.9
gpt-oss-20b	C2 as-of	2024-06-30	after cutoff	322	75.8 [71.3, 80.3]	19.6
DeepSeek-V4-Pro	C0 no tool	2025-05-31	after cutoff	322	51.6 [45.5, 57.3]	38.8
DeepSeek-V4-Pro	C1 latest	2025-05-31	after cutoff	322	0.9 [0.0, 2.4]	59.3
DeepSeek-V4-Pro	C2 as-of	2025-05-31	after cutoff	322	87.0 [83.0, 90.9]	8.1
Claude Sonnet 5.5	C0 no tool	2026-06-30	on/before cutoff	249	89.6 [84.3, 94.1]	3.2
Claude Sonnet 5.5	C0 no tool	2026-06-30	after cutoff	73	37.0 [25.4, 48.0]	0.0
Claude Sonnet 5.5	C1 latest	2026-06-30	on/before cutoff	249	59.8 [51.8, 67.5]	15.7
Claude Sonnet 5.5	C1 latest	2026-06-30	after cutoff	73	2.7 [0.0, 8.2]	23.3
Claude Sonnet 5.5	C2 as-of	2026-06-30	on/before cutoff	249	95.6 [93.1, 97.9]	4.0
Claude Sonnet 5.5	C2 as-of	2026-06-30	after cutoff	73	94.5 [88.9, 98.7]	2.7

Appendix H. C2 error analysis

Twenty C2 FLIP failures were sampled each for Llama-3.1-8B and DeepSeek-V4-Pro (seed 2027). Mechanical categories (no verdict, abstention, missing or failed tool calls, wrong series or date) were assigned by script; the remainder were coded by an LLM assistant reading the full transcript and tool output. Code counts are in the repository; the per-claim codes and notes quote claim text and will be released with the companion data.

Cause (20 random C2 FLIP failures each)	Llama-3.1-8B	DeepSeek-V4-Pro
revision claim: earlier print not consulted	10	4
abstained (UNVERIFIABLE)	3	6
aggregation, arithmetic or units	5	1
no verdict: tried another tool call after the 3-call budget	0	4
precision rule applied over-strictly	0	3
relative-period confusion	0	2
misread tool output	1	0
possible reference-label error	1	0

Appendix I. Use of LLMs

LLMs are the object of study. An LLM-based coding assistant was also used to write the evaluation harness and analysis scripts, to draft text, and to code the error analysis above; all numbers are generated by scripts from the saved run outputs, and the author reviewed the design, code and text.

Appendix J. HOLDOUT baseline and per-class accuracy (post-hoc)

This section is a post-hoc addition (DEVIATIONS.md, D6); it does not change any pre-registered hypothesis or result. Accuracy is per-population agreement with the reference label: FLIP and STABLE reference SUPPORTED; HOLDOUT reference is the mixed blind label, 216 SUPPORTED and 75 not. A verifier that answers SUPPORTED to every claim therefore scores the majority baseline 74.2% under every condition; no model in Table 1 reaches it (best is S-C2 at 69.1%). To test whether the as-of tool *separates* accurate claims from false ones we report accuracy on only the 75 non-SUPPORTED HOLDOUT claims. Table 2 gives that accuracy without a tool and with the as-of tool. With the tool it is higher for Llama-3.1-8B, about the same for gpt-oss-20b and DeepSeek-V4-Pro, and lower for Claude Sonnet 5.5 (29.3% against 44.0%), which answers SUPPORTED to more of these claims once it has the tool. The as-of tool fixes the FLIP failure and helps on STABLE, but this study does not show that it helps a model separate accurate claims from inaccurate ones. 69 of the 75 claims are UNVERIFIABLE and only 6 are CONTRADICTED, so this is mostly a test of abstention.

Table 2: Accuracy (%) on the 75 non-SUPPORTED HOLDOUT claims, and share of them answered SUPPORTED, without a tool (C0) and with the as-of tool (C2). Post-hoc (D6). Always answering SUPPORTED scores 0 here. Llama-3.1-8B's C2 figures are out of 74 (1 episode missing after retries); all other cells are out of 75.

	Accuracy		Said SUPPORTED	
Model	C0	C2	C0	C2
Llama-3.1-8B	20.0	31.1	30.7	23.0
gpt-oss-20b	57.3	58.7	14.7	13.3
DeepSeek-V4-Pro	38.7	41.3	41.3	33.3
Claude Sonnet 5.5	44.0	29.3	37.3	53.3