BEYOND BENCHMARKS: DISAGREEMENT AMONG FRONTIER LLMS ON REAL-WORLD FACT-CHECKS

A PREPRINT

Kosta Jordanov
Lenz, CEO
Sofia, Bulgaria
kosta@lenz.io
0009-0008-5711-2085

David Yordanov
Bocconi University, student
Milan, Italy
david.yordanov@lenz.io
0009-0007-3979-4193

Yana Jordanova
American College of Sofia, student
Sofia, Bulgaria
yana.jordanova@lenz.io
0009-0008-4941-6759

Research version v1.1 — August 7, 2026

ABSTRACT

Frontier large language models achieve comparable accuracy on public benchmarks, a parity often read as evidence that they are interchangeable as assessors of factual claims. We test that assumption directly, on claims not drawn from any public benchmark. Five frontier models were each asked to adjudicate 1,000 real-world claims submitted by users to a fact-checking platform. They were tasked with assigning every claim a verdict on a five-point scale from True to False, as well as reporting their confidence in each answer. On the 997 claims where all five models returned a usable verdict, they fail to reach consensus on 63%, and on 23% the two furthest-apart verdicts differ by at least two verdict categories. The ordinal Krippendorff's α of 0.77 reflects structured but far from interchangeable judgement. We find that disagreement concentrates in the intermediate verdicts, where claims resist clean adjudication: the definitive poles are unanimous about half the time, against roughly one in ten for the intermediate verdicts. In relation to disagreement, our results show that model confidence is not a reliable indicator of whether the panel will agree. The models are highly confident almost everywhere, rating 76% of answers 9 or 10 on a 1–10 confidence scale, yet they agree with one another on confidence markedly less than on the verdicts themselves ($\alpha = 0.44$). The verdict that is given to an assertion will then depend heavily on which model one consults, which matters a great deal as more people turn to such models to verify information.

1 Introduction

Large language models are increasingly consulted as arbiters of factual questions, by individual users and by automated fact-checking systems alike. Since the frontier models score similarly on public benchmarks, it is easy to assume that it makes little difference which one you ask. Benchmark parity, however, says nothing about whether the models reach the same judgement on the same claim. If the perceived validity of a claim depends on which model a person or platform happens to use, model disagreement becomes a source of inaccuracy independent of any individual model's benchmark standing. To measure this disagreement, we presented five frontier models with 1,000 recent user-submitted claims under a controlled, forced-choice prompt. Each model was required to classify every claim on a verdict scale: True, Mostly True, Mixed, Mostly False, or False and give a score $(1 - 10)$ on how confident it is in its assessment. All five models returned a usable verdict on 997 of those claims, and every figure reported below is computed on that cohort.

We find substantial disagreement: on 63% of claims (632 / 997; 95% CI 60–66%) at least one model dissents from the panel majority or no majority forms at all, and on 23% (232 / 997; 95% CI 21–26%) at least two models' answers differ by two or more verdict options. This study extends an earlier version of our work [1] under a revised methodology (§1.1); despite these methodological improvements, significant disagreement remains. We also find that a model's self-reported confidence does little to indicate this disagreement. The models report high confidence on the great majority of their answers yet disagree no less often, and the confidence scores themselves show weaker inter-rater reliability than the verdicts they accompany. Only in aggregate does confidence carry a signal, since the panel agrees on the

verdict far more often on claims where every model is confident than on those where at least one is not. In practice, the answer a user obtains can therefore hinge on the model they happen to ask, a dependency that is important to further study.

1.1 Previous Works

There is an assortment of literature regarding accuracy rates of different models and methods to improve them, but there are few studies particularly relevant to our work. Sahitaj et al. [2] establish LLM baselines for automated fact-checking across binary, three-class, and five-class labelling schemes, finding that larger LLMs outperform smaller ones in both accuracy and reasoning quality on real-world claims from PolitiFact. Turning to inter-model disagreement specifically, Yang and Wang [3] compare model predictions on two reasoning benchmarks, MMLU-Pro and GPQA, and find that even leading models of comparable accuracy give different answers on 16–38% of items. For real-world claim verification with human annotation, AVeriTeC [4] remains the canonical corpus; however, having been publicly available since 2023, it has likely entered the training data of current models, allowing them to recall published answers rather than assess claims independently. Regarding AI models' confidence scores, Kadavath et al. [5] find that larger language models are reasonably well-calibrated on multiple-choice and true/false questions, with self-reported confidence tracking actual correctness, though this calibration degrades on novel task formats and only partially generalises across tasks.

Relation to Prior Version. This study extends an earlier version of our work [1] by improving the methodology used. The scale has been expanded to five symmetric verdict categories (True, Mostly True, Mixed, Mostly False, False), each accompanied by a brief definition in the prompt so as to avoid differing interpretations. Models are also asked to provide reasoning before reaching a verdict and then give a confidence score. There are now a total of five, instead of the initial four models, each in its most recent publicly available version at the time of writing. They are configured with web retrieval, while excluding lenz.io as a source where the provider's API permitted it. We have run the prompts with an entirely fresh corpus of claims (§3.1).

1.2 Contributions

Whereas much prior work has centred on measuring and improving accuracy, our study takes inter-model disagreement itself as the object of analysis. Rather than drawing on benchmarks or other public datasets, whose items and reference answers are well documented online and plausibly present in training data and during retrieval, we evaluate recent, real-world claims whose verdicts the models should not have encountered before prompting and are purposely excluded (*as much as possible*) in their retrieval. In addition, we examine the confidence level each model reports on every claim and use that to present how it differs based on the specific model and type of claim, as well as to show how the level of disagreement relates to the confidence reported. We evaluate five of the most widely used frontier systems: Claude Fable 5, GPT-5.6-Sol, Gemini 3.1 Pro + Search, Sonar Deep Research, and Grok 4.5, in their most recent publicly available versions at the time of writing, so that observed disagreement reflects the verdicts users currently receive.

2 Research Focus

This study focuses on the extent to which frontier LLMs agree with one another when judging real-world claims, and on where that disagreement concentrates. We examine inter-model disagreement rather than accuracy. Because the corpus carries no human-labelled *correct* verdict for any claim, accuracy cannot be assessed, and we make no claim as to which verdicts are right. Instead, we examine how consistently five leading models reach the same judgement on the same claim, and what that consistency might depend on.

We approach this through two aspects, the models' verdict agreement and their self-reported confidence. Verdict agreement is measured with the following:

- **Panel disagreement rate.** The share of claims on which at least one model dissents from the majority, or no majority forms. A majority is a verdict held by at least three of the five models.
- **Maximum verdict distance.** How far apart the two most-divergent verdicts fall on the ordinal five-point scale.
- **Pairwise agreement.** How often each pair of models assigns the same verdict across the corpus.
- **Per-model majority alignment.** How often each model matches the strict majority of the other four, where such a majority exists.
- **Krippendorff's ordinal α.** Panel-level inter-rater reliability.

Each of these is also broken down by verdict category and by claim domain, to locate where disagreement concentrates. For confidence we examine the following:

- **Confidence distribution.** How the reported scores are spread across the panel and its individual models.
- **Confidence against disagreement.** Whether reported confidence corresponds to inter-model disagreement at the level of the individual claim.
- **Inter-model confidence agreement.** Whether the models agree with one another on how confident to be, quantified with Krippendorff's α on the confidence scale.

Note: A majority formed between models is not considered truth. It is used only as a reference point from which to measure disagreement, not as a stand-in for correctness. A high confidence score is likewise not treated as evidence that a verdict is correct.

3 Methodology

3.1 Corpus

The corpus used consists of the 1,000 most recent real-world claims submitted by users to the fact-checking platform Lenz that satisfy the eligibility criteria in §3.1.1, with no claim submission predating May 1, 2026. We constructed this corpus rather than adopting an existing dataset in order to minimise the risk that the claims, or verdicts concerning them, appear in the models' training data. User-submitted claims also better reflect the types of questions AI models and fact-checking systems are being used to answer.

Claim standardisation. Models were presented with a distilled formulation of each claim rather than the user's raw text. This standardisation removes emotionally charged language and ambiguous phrasing, with the goal of ensuring the claims are reasonably verifiable. This process is performed within the Lenz platform itself, and all five models receive the identical standardised text, so no part of the measured disagreement can come from differences in how the claim was worded.

3.1.1 Exclusions

The corpus excludes claims marked private; claims from staff, internal, or API submissions; claims with editorial status pending or hidden (including PII-flagged claims); near-duplicate claims (collapsed to a single canonical row by cosine distance on OpenAI `text-embedding-3-small`); and claims where at least one model failed to produce a parseable verdict after one retry (3 residual cell-level parse errors).

Note: Out of the 1,000 total claims the models were presented with, Claude Fable 5 did not return a valid output, even after the retry, to 145 claims. After rerunning those claims with Opus 4.8 as a fallback, 3 claims remained rejected (listed in Appendix B). For that reason, the results given are all using a total of 997 claims.

3.2 Model Selection

We selected one frontier AI from five of the most popular AI companies: Claude Fable 5 (Anthropic), GPT-5.6-Sol (OpenAI), Gemini 3.1 Pro + Search (Google), Sonar Deep Research (Perplexity), and Grok 4.5 (xAI). Each of the five was set up to search the web, so the verdicts they return are representative of those a user querying them would receive.

3.3 Excluding Lenz as a Source in Retrieval

All five models had retrieval enabled and could look up sources at inference time. An explicit lenz.io exclusion was passed to four of the five providers; Gemini's search-grounding tool does not accept a domain-exclusion parameter under a standard Developer API key (only under Vertex AI's Enterprise Agent Platform mode, which this study does not use), so Gemini 3.1 Pro + Search ran without this guard and we looked over the sources it cited. Despite the restrictions, in 39 of the total 5,000 runs lenz.io appeared in their sources and in 10 of those the page cited was Lenz's own verdict on the very claim being evaluated. For example, one of the claims was directed towards inter-model disagreement and the examining model had sourced the previous version of our paper as one out of the many used resources. While it is impossible to completely eliminate contamination from our study, we believe that we have reduced it as much as possible, so that our statistics are still relevant and representative.

3.4 Prompt

Every model receives the same system prompt and user prompt template, reproduced in full below. The `{date}` anchor is the date on which the claim was submitted by a user to Lenz for verification; it pins each model to the claim's submission-time epistemic frame.

System Prompt.

```
You are a fact-checker.  Evaluate the truthfulness of the claim as of the stated
date.  Respond with JSON only.
```

User Prompt Template.

```
Evaluate this claim as of {date}:
"{claim}"

Choose exactly one verdict from the scale below:
    "True" - the claim is accurate
    "Mostly True" - the claim is largely accurate with minor caveats or omissions
    "Mixed" - the claim has both accurate and inaccurate elements
    "Mostly False" - the claim is largely inaccurate with some basis in fact
    "False" - the claim is inaccurate

Respond with a JSON object containing exactly these fields:
    "reasoning":  2-4 sentences of claim analysis and verdict justification
    "verdict":  one of the five labels above
    "confidence_level":  your level of certainty in the verdict on a 1 to 10 integer
scale (1 = completely uncertain, 10 = fully certain)
```

Note: No "Abstain" category is provided for. The models vary in their willingness to abstain. In the initial testing phase, some of the frontier models refused to judge certain claims whereas other models were always willing to make a judgement call, so an abstention is a model-specific action: one model's refusal to answer cannot be compared with another model's confidently wrong verdict, despite both of them having epistemic weight. Forcing the same five options on every model captures the disagreement between the judgements, not between the decisions. Abstaining is also not a result which a user can use. If a user is trying to check a claim and finds out that the model refuses to make a judgement about it, they still have not got any result, and our task is to identify the disagreement between the results that users actually receive. Nor is a forced option a forced certainty — each judgement carries a confidence level from 1 to 10, which is where a model registers doubt. Unparseable outputs are not reclassified into a verdict category; claims with any parse error are excluded from the complete-claim cohort (§3.1.1).

3.5 Call Configuration

All five models run at a medium reasoning tier. Extended thinking is enabled at each provider's medium depth setting (Anthropic `effort=medium`; OpenAI and xAI `reasoning_effort=medium`; Gemini `thinking_budget=16384`; Sonar Deep Research's multi-step research is always on, at Perplexity's default `reasoning_effort` of `medium`). Retrieval/web search is enabled on every model (OpenAI at `search_context_size=medium`). Responses were constrained to a fixed JSON schema (`reasoning`, `verdict`, `confidence_level`) enforced natively at the API level, forcing each model to emit strict, machine-readable JSON. Retrieval was constrained to exclude lenz.io as a source wherever the provider's API permitted it, with the single exception and its consequences detailed in §3.3. Deterministic decoding was requested where supported (`temperature=0.0`), and output caps were raised on reasoning models so extended thinking cannot truncate the verdict. The per-model settings summarised here are recorded in full in `model_settings.yaml` in the harness repository (Appendix C), which is the authoritative record of what each provider was sent.

3.6 Grading

No LLM grader. All measurements derive from direct parsed-label equality between the five frontier verdicts on the same claim. No reference label or ground truth is used.

3.7 Statistical Analysis

Sampling frame. The corpus used consists of the 997 most recent eligible claims from a single platform; it is neither a probability sample from any larger population nor a complete enumeration. The reported Wilson 95% CIs [6] are nominal binomial intervals, based on a model in which each claim is an independent draw from a hypothetical stream of similar eligible submissions; they are not coverage statements about "all real-world fact-checks."

Note: Lenz claims are not independent: users cluster submissions around news events and often submit multiple related claims in a session. True sampling variability under a cluster model would likely exceed what Wilson reports; we surface CIs as a minimum precision floor, not a guaranteed coverage interval.

Inter-rater Reliability. The verdict scale is ordinal, so we score agreement with Krippendorff's α at the ordinal level of measurement [7] rather than Fleiss' κ, which would treat the categories as nominal and understate agreement. The confidence level scale is also ordinal and can be scored similarly.

4 Results by Verdict

4.1 Panel Disagreement Rate

For each claim, we begin by determining whether a strict majority has formed. When such a majority exists, we examine how many of the remaining models have dissented. In the remaining cases, where no single verdict is agreed upon by (at least) three of the five models, the claim is counted towards the "*models split, no majority*", as indicated in Table 1.

Frontier verdict pattern	Claims	Share
All 5 agreed (unanimity)	365	37% (34–40%)
1 of 5 dissented	246	25% (22–27%)
2 of 5 dissented	274	27% (25–30%)
Models split, no majority	112	11% (9–13%)
$\geq$1 model dissents	632	63% (60–66%)
$\geq$2 models dissent	386	39% (36–42%)

Table 1: Frontier verdict-agreement pattern across the five-model panel ($n = 997$ claims). Wilson 95% CIs in parentheses.

4.2 Maximum Verdict Distance — Calibration vs Substantive Disagreement

Disagreement can be of varying degrees. The gap between "True" and "Mostly True" indicates a difference in calibration, while that between "True" and "False" indicates an actual dispute over the claim. In order to differentiate between the two, we determine for each claim its *maximum verdict distance*: the largest gap between any two of the five verdicts, ranking the categories as follows: True (0) $\rightarrow$ Mostly True (1) $\rightarrow$ Mixed (2) $\rightarrow$ Mostly False (3) $\rightarrow$ False (4). Table 2 reports the resulting distribution.

Distance	Interpretation	Claims	Share
0	Full unanimity (all 5 same verdict)	365	37% (34–40%)
1	Nuance only (e.g. True ↔ Mostly True)	400	40% (37–43%)
2	Substantive (e.g. True ↔ Mixed)	123	12% (10–15%)
3	Very substantive (e.g. True ↔ Mostly False)	75	8% (6–9%)
4	Polar (True ↔ False)	34	3% (2–5%)
≥2 verdicts apart (substantive or polar)		232	23% (21–26%)

Table 2: Maximum pairwise verdict distance across the five frontier verdicts per claim, on the ordinal scale: True (0) → Mostly True (1) → Mixed (2) → Mostly False (3) → False (4).

Note: The distance between verdicts assumes an equal-interval ordinal scale for the five categories, which is an approximation. A distance of size 2 may come about due to the vagueness of the rubric, differing time frames, or different understandings of what "Mixed" means. This distance is provided just as a rough marker of substance versus nuance, not as a measure of error size.

4.3 Pairwise Agreement

Table 3 illustrates the frequency of identical verdict labels chosen by each pair of frontier models within the corpus. Pairwise agreement reaches up to 76% and drops to 49%.

Model	Claude Fable 5	GPT-5.6-Sol	Gemini 3.1 Pro + Search	Sonar Deep Research	Grok 4.5
Claude Fable 5	—	67% (64–70%)	69% (66–72%)	57% (54–60%)	74% (71–76%)
GPT-5.6-Sol	67% (64–70%)	—	61% (58–64%)	64% (61–67%)	65% (62–68%)
Gemini 3.1 Pro + Search	69% (66–72%)	61% (58–64%)	—	49% (46–52%)	76% (73–78%)
Sonar Deep Research	57% (54–60%)	64% (61–67%)	49% (46–52%)	—	53% (50–56%)
Grok 4.5	74% (71–76%)	65% (62–68%)	76% (73–78%)	53% (50–56%)	—

Table 3: Pairwise label-agreement between frontier models (% of claims with an identical verdict). Wilson 95% CIs in parentheses.

4.4 Per-Model Verdict Distribution

Table 4 shows the share of claims each model assigned to each verdict category.

Model	True	Mostly True	Mixed	Mostly False	False
Claude Fable 5	51% (48–54%)	20% (18–23%)	7% (5–8%)	10% (8–12%)	13% (11–15%)
GPT-5.6-Sol	41% (38–44%)	26% (23–29%)	5% (4–6%)	14% (12–16%)	14% (12–17%)
Gemini 3.1 Pro + Search	60% (57–63%)	7% (5–8%)	4% (3–6%)	5% (4–7%)	23% (21–26%)
Sonar Deep Research	32% (30–35%)	28% (25–31%)	12% (11–15%)	16% (14–19%)	11% (9–13%)
Grok 4.5	59% (56–62%)	14% (12–16%)	3% (2–4%)	7% (5–8%)	18% (16–20%)

Table 4: Per-model verdict distribution. Wilson 95% CIs in parentheses.

4.5 Per-Model Majority Alignment

For each model, we examine how often its verdict matches the strict majority ($\geq 3/4$) of the remaining four, on claims on which there is such a majority. The per-model majority match rates range from 63% to 86% (Table 5). This measure is an indication of match to the panel for this corpus, not correctness: no model is treated as ground truth, and the eligible n varies by row.

Model	Agreement w/ peer majority	Eligible n	Ineligible
Claude Fable 5	86% (84–89%)	680	317
GPT-5.6-Sol	78% (75–81%)	730	267
Gemini 3.1 Pro + Search	80% (77–82%)	712	285
Sonar Deep Research	63% (59–66%)	772	225
Grok 4.5	84% (81–87%)	709	288

Table 5: How often each model's verdict matches the strict majority (≥3/4) of the other four. *Eligible* n is claims where such a majority exists.

4.6 Breakdowns by Domain and Majority Verdict

The three tables below present additional statistics. Table 6 presents the disagreement rates based on the eight domains the corpus' claims fall into. The claims were divided into domains directly within the Lenz platform. Table 7 looks only at claims on which a strict majority formed — at least three of the five models on the same verdict — and reports, for each verdict label, how often that majority was unanimous (5/5) rather than majority-only (3–4/5). Table 8 shows the verdict distribution within the fully unanimous claims.

Domain	Claims	Any disagreement	Substantive (≥2)	No majority
Finance	53	62% (49–74%)	26% (16–40%)	21% (12–33%)
General	150	59% (51–66%)	24% (18–31%)	10% (6–16%)
Health	181	75% (68–81%)	30% (24–37%)	14% (10–20%)
History	164	54% (46–61%)	13% (9–19%)	4% (2–9%)
Legal	62	65% (52–75%)	29% (19–41%)	15% (8–25%)
Politics	173	66% (59–73%)	25% (20–32%)	13% (9–19%)
Science	156	61% (53–68%)	22% (16–29%)	13% (8–19%)
Tech	58	66% (53–76%)	19% (11–31%)	3% (1–12%)

Table 6: Per-domain frontier disagreement. Denominator per row is the claims in that domain.

Majority verdict	Eligible n	Unanimous (5/5)	Majority only (3–4/5)
True	502	54% (50–59%)	46% (41–50%)
Mostly True	129	5% (3–11%)	95% (89–97%)
Mixed	30	7% (2–21%)	93% (79–98%)
Mostly False	76	11% (5–19%)	89% (81–95%)
False	148	51% (43–59%)	49% (41–57%)

Table 7: Per-verdict panel agreement. Denominator is claims with a strict ≥3/5 frontier majority on that verdict.

Unanimous verdict	Claims	Share of unanimous
True	273	75% (70–79%)
Mostly True	7	2% (1–4%)
Mixed	2	1% (0–2%)
Mostly False	8	2% (1–4%)
False	75	21% (17–25%)

Table 8: Out of the 365 claims on which all five models agreed, the distribution of verdicts.

5 Results by Confidence Level

In providing every verdict, each model also reports its level of confidence in that verdict, in the form of a number between 1 and 10 (1 being the lowest confidence, 10 the highest). This is the models' confidence in their own answer, not the accuracy of that answer. The following section presents the collected data on these numbers, their distribution, and their relation to disagreement.

5.1 Confidence Level Distribution

Below are the tables which show the distribution of the confidence levels for each individual model. The figures are heavily skewed towards higher confidence levels: 76% of all answers have a confidence level of 9 or 10, with mean confidence levels per model varying between 8.2 and 9.5 (the *Mean* row of Table 9; counts in Table 10). Low confidence is rare.

Level	Fable 5	GPT-5.6-Sol	Gemini	Sonar	Grok 4.5	All
1	0%	0%	2%	0%	0%	0%
2	0%	0%	0%	0%	0%	0%
3	1%	0%	0%	0%	0%	0%
4	1%	0%	0%	0%	0%	0%
5	2%	0%	0%	0%	0%	0%
6	8%	0%	0%	1%	1%	2%
7	15%	1%	0%	5%	5%	5%
8	23%	15%	1%	15%	22%	15%
9	33%	44%	27%	41%	43%	38%
10	18%	39%	69%	37%	29%	38%
Mean	8.23	9.20	9.48	9.08	8.95	—

Table 9: Per-model self-reported confidence distribution: share of each model's claims at each 1–10 level (columns are models, rows are levels). The *Mean* row is each model's mean confidence; *All* pools every answer (4,985 = 997 claims × 5 models).

Level	Fable 5	GPT-5.6-Sol	Gemini	Sonar	Grok 4.5	All
1	0	0	24	0	0	24
2	0	0	0	0	0	0
3	6	0	0	0	0	6
4	12	0	0	0	1	13
5	20	1	0	2	0	23
6	83	3	0	9	5	100
7	149	13	0	52	50	264
8	225	153	13	152	220	763
9	325	439	272	412	432	1,880
10	177	388	688	370	289	1,912
Total	997	997	997	997	997	4,985

Table 10: Per-model confidence distribution as raw answer counts (companion to Table 9).

5.2 Confidence in Relation to Disagreement

The following tables present the confidence scores against the levels of disagreement present between models. Grouping claims by their *lowest* model confidence given when accessing them (Tables 11 and 12), substantive verdict disagreement falls as confidence rises: claims in the low-to-mid range (1-6) show 54% substantive disagreement, versus 3% for the very-high range (9-10), and claims where all five models report maximum confidence show essentially none.

Confidence floor	Claims	Any disagreement	Substantive	Krippendorff's α
1+	997	63% (60–66%)	23% (21–26%)	0.77
2+	973	64% (61–67%)	24% (21–27%)	0.77
3+	973	64% (61–67%)	24% (21–27%)	0.77
4+	967	64% (61–67%)	24% (21–26%)	0.77
5+	954	63% (60–66%)	23% (20–26%)	0.77
6+	934	63% (60–66%)	22% (19–24%)	0.77
7+	848	59% (56–63%)	18% (15–21%)	0.77
8+	680	52% (49–56%)	12% (10–14%)	0.75
9+	451	36% (32–40%)	3% (2–5%)	0.73
10	126	9% (5–15%)	0% (0–3%)	0.86

Table 11: Verdict agreement grouped by each claim's *lowest* model confidence (cumulative floor). Substantive = $\geq$2 verdict labels apart.

Confidence band	Claims	Any disagreement	Substantive	Krippendorff's α
Very low/low/mid (1-6)	149	86% (79–91%)	54% (46–62%)	0.57
High/very high (7-10)	848	59% (56–63%)	18% (15–21%)	0.77
Very high (9-10)	451	36% (32–40%)	3% (2–5%)	0.73

Table 12: Verdict agreement grouped into named confidence bands (on each claim's lowest model confidence). Bands overlap by design.

5.3 Where Confidence Concentrates

Table 13 reports the confidence distribution of every model answer grouped by the claim's domain. Table 14 groups model answers by verdict and similarly reports distribution. In both, the 1–10 scale is collapsed into five bands: Very Low (1–2), Low (3–4), Mid (5–6), High (7–8), and Very High (9–10); and each cell reports the answer count above its share of the row, so every row sums to 100% (small differences due to rounding). The domain breakdown is largely uniform: the Very High band alone accounts for between 70% (Health) and 85% (History) of answers in every domain, and no domain attracts materially more caution than the others. The verdict breakdown separates far more sharply. The definitive poles concentrate their confidence scores in the Very High range: 92% of *True* and 84% of *False* answers fall in it. In contrast, only 49% of *Mostly True*, 46% of *Mixed*, and 55% of *Mostly False* answers reach Very High, with the remainder distributed across the High (7–8) band rather than the lower ones. These results are in agreement with reasonable expectations, as they show that presumably models are more assured in their answer when giving more polarising, less-nuanced verdicts. Table 15 presents the same information in the opposite direction: the verdict composition of each confidence band, rather than the confidence distribution of each verdict (Table 14). Each row represents one band and totals 100%, with every cell giving the number of answers above its share of that band. The two lowest bands together contain fewer than fifty answers, so their shares rest on a small base.

Domain	Total	Very Low	Low	Mid	High	Very High
Finance	265	1	1	3	69	191
		(0%)	(0%)	(1%)	(26%)	(72%)
General	750	2	6	32	177	533
		(0%)	(1%)	(4%)	(24%)	(71%)
Health	905	4	5	27	237	632
		(0%)	(1%)	(3%)	(26%)	(70%)
History	820	6	1	8	108	697
		(1%)	(0%)	(1%)	(13%)	(85%)
Legal	310	0	1	9	59	241
		(0%)	(0%)	(3%)	(19%)	(78%)
Politics	865	0	0	24	200	641
		(0%)	(0%)	(3%)	(23%)	(74%)
Science	780	8	4	18	132	618
		(1%)	(1%)	(2%)	(17%)	(79%)
Tech	290	3	1	2	45	239
		(1%)	(0%)	(1%)	(16%)	(82%)

Table 13: Confidence level by claim domain, over all 4,985 model answers. *Total* is the domain's answer count; each remaining cell: count, with its share of that domain's answers below.

Verdict	Total	Very Low	Low	Mid	High	Very High
True	2,433	18	0	1	171	2,243
		(1%)	(0%)	(0%)	(7%)	(92%)
Mostly True	941	0	6	51	421	463
		(0%)	(1%)	(5%)	(45%)	(49%)
Mixed	306	0	8	30	128	140
		(0%)	(3%)	(10%)	(42%)	(46%)
Mostly False	517	0	4	29	198	286
		(0%)	(1%)	(6%)	(38%)	(55%)
False	788	6	1	12	109	660
		(1%)	(0%)	(2%)	(14%)	(84%)

Table 14: Confidence level by the answer's own verdict, over all 4,985 model answers (companion to Table 13). *Total* is the verdict's answer count.

Confidence band	Total	True	Mostly True	Mixed	Mostly False	False
Very Low	24	18 (75%)	0 (0%)	0 (0%)	0 (0%)	6 (25%)
Low	19	0 (0%)	6 (32%)	8 (42%)	4 (21%)	1 (5%)
Mid	123	1 (1%)	51 (41%)	30 (24%)	29 (24%)	12 (10%)
High	1,027	171 (17%)	421 (41%)	128 (12%)	198 (19%)	109 (11%)
Very High	3,792	2,243 (59%)	463 (12%)	140 (4%)	286 (8%)	660 (17%)

Table 15: Verdict composition of each confidence band, over all 4,985 model answers (transpose of Table 14). *Total* is the band's answer count; each remaining cell: count, with its share of that band's answers below.

5.4 Models' Agreement on Confidence

Table 16 examines whether the models agree with one another on confidence rather than on the verdict, comparing the confidence scores a claim received regardless of the verdict labels. They differ by at least one point on 87% of claims and by three or more points on 21%, and the ordinal Krippendorff's α for confidence is 0.44, far below the verdict-level α of 0.77.

Measured on	Claims	Any disagreement	Significant	Krippendorff's α
Confidence (1–10)	997	87% (85–89%)	21% (19–24%)	0.44

Table 16: Model agreement on confidence. Any disagreement = spread $\geq$1 point; significant (disagreement) = spread $\geq$3 points.

6 Analysis

6.1 Verdict

A lower bound on model error. For each claim, exactly one of the five verdict categories is correct, or may at least be regarded as most nearly correct. Under the most charitable assumption available, that the panel's modal verdict is the correct one, still at least one model assigns an "*incorrect*" verdict on 63% of claims, at least two do so on 39%, and at least three on the 11% of claims (for which no verdict category attains a majority). These figures constitute a lower bound as relaxing the modal-correctness assumption can only decrease overall accuracy.

Note: Regardless of the above reasoning, we do have to acknowledge that without specific correct classifications of the claims, it is impossible to give concrete statistics. There is also the possibility that there are more nuanced claims that may reasonably fit into more than one of the five categories as well as a degree of boundary-ambiguity to the scale. The reasoning presented above is meant to show that regardless of these factors, even only on the basis of disagreement, it is clear that there must be a significant amount of error.

The middle of the rubric is where the panel fractures. When the panel lands on a middle category, it almost never converges to unanimity, while polar (True / False) majorities do so far more often (Table 7). Table 8 shows the same tendency. Of the 365 unanimous claims the overwhelming majority (95%) are unanimous-True or unanimous-False. This pattern admits two explanations. The first is that claims with a determinate, well-established verdict are easier to adjudicate, so models converge on them. The second is that the models disagree on where one verdict label ends and the next begins, as the 5-point scale is ultimately a simplification of a spectrum. This should be mitigated, at least to some degree, by the verdict category definitions present in the prompt, but it cannot be eliminated fully as a

factor. Regardless, what can be said without qualification is that the panel agrees far more on definitive claims than on nuanced ones.

Per-model differences. Table 4 shows that all of the models tend to favour True and Mostly True verdicts. While one might draw conclusions about the models, this more likely reflects the corpus. It is likely that the claims tend to lean towards being true and that was reflected in the models' answers. This can ultimately be determined with human labelling of the corpus, a natural step for future work.

Domain relevance. Table 6 presents the disagreement rates per domain and most relevant for analysis is the "Health" domain, since it observed the highest level of substantive disagreement (30%). This can be concerning, since health-related claims would be considered some of the most important to be able to credibly verify. That said, health claims are also among the more nuanced and hence difficult to adjudicate and label. Since we have a relatively small corpus, once it is divided into eight categories, its statistical precision is reduced. This is reflected in the wide Wilson 95% CIs. Due to the small sample size, we cannot draw firm conclusions, but it is still relevant to note the persistence of substantial inter-model disagreement within each of the domains.

Panel reliability. Krippendorff's α (ordinal) $= 0.77$ places the panel below the conventional reliability threshold. Krippendorff's own guidance treats $\alpha \geq 0.80$ as the floor for relying on coded data and $0.667 \leq \alpha < 0.80$ as warranting only tentative conclusions [7]; the panel falls in this lower band. The verdicts are therefore structured rather than random, the models are responding to a shared signal in each claim, but not concordant enough to be treated as exchangeable measurements of a single underlying verdict. This is the quantitative form of the claim made in Section 1: comparable aggregate benchmark accuracy does not make frontier models interchangeable as assessors, since two models of similar standing still return ordinally divergent verdicts often enough to keep panel-level reliability under the accepted bar.

6.2 Confidence

High overall confidence coexists with substantial disagreement. The five models reported high confidence throughout the corpus. 76% of all answers were assigned a confidence of 9 or 10, and per-model mean confidence ranged from 8.2 to 9.5 (Table 9). The same concentration is evident in the raw counts, where the great majority of the 4,985 answers fall at the top of the scale (Table 10). This high overall confidence is difficult to reconcile with the disagreement established in Section 4, where the panel diverges on 63% of claims and disagrees substantively on 23%. The models are thus, in aggregate, both highly confident and frequently in disagreement, which indicates that, contrary to what one might intuitively assume, a high reported confidence does not by itself imply that a verdict is uncontested.

Disagreement is concentrated in the low-confidence claims. This disagreement is not distributed evenly across the confidence scale but is concentrated among the claims the models were least certain about. Grouping each claim by its confidence floor, the lowest confidence any of the five models assigned to it, isolates the pattern (Table 11). On the claims where every model was very confident, corresponding to a floor of 9 or 10, substantive disagreement falls to 3%, and on those where all five reported the maximum it practically disappears. These high-confidence claims constitute close to half of the corpus yet contribute almost none of its substantive disagreement, which entails that the disagreement measured across the full corpus arises overwhelmingly from the lower-confidence claims. This is shown directly in Table 12, where claims whose floor falls in the lower half of the scale (1–6) disagree substantively at 54%, more than double the corpus-wide rate of 23%, and any disagreement among them is elevated as well. Even in broader band "High/very high (7–10)", substantive disagreement remains at 18%, which, although below the overall 23%, is still considerably higher than the 3% observed once only the 9–10 scores are isolated. Confidence measured at the level of the claim can therefore be used to separate the contested claims from the settled ones, even though a model's confidence in isolation, as established above, does not.

Confidence in relation to domain and verdict. Table 13 shows no major departure from the pooled distribution, as reported confidence remains concentrated in the Very High band across every domain. Between 70% (Health) and 85% (History) of each domain's answers fall in the Very High band, with a further 13% to 26% in High. The concentration of high confidence therefore persists when the answers are disaggregated by domain. The differences between domains are minor, and none is large enough to support a conclusion, particularly given the small number of claims in each domain (see the *Domain relevance* discussion in §6.1).

Table 14 instead groups the answers by their own verdict, and here the distribution separates sharply. The two polar verdicts are overwhelmingly concentrated in the Very High band, with 92% of True and 84% of False answers rated 9 or 10. The three intermediate verdicts are considerably less certain, as only 49% of Mostly True, 46% of Mixed, and

55% of Mostly False answers reach the Very High band, and most of the remainder sits in the High (7–8) band rather than lower. Table 15 expresses the same coupling in reverse. The Very High band, which holds roughly three quarters of all answers, is composed predominantly of the two poles (59% True and 17% False), whereas the mid band is made up almost entirely of the intermediate verdicts (89% combined). The two lowest bands rest on fewer than fifty answers and are not interpreted here as the sampling is far too small.

Taken together, these tables indicate that the models express their strongest confidence on the polar verdicts and reserve their uncertainty for the claims that call for an intermediate judgement. This parallels the structure found for verdict agreement in Section 4, where the panel converges at the poles and fractures across the middle. Confidence and agreement therefore move together, both tracking where on the verdict scale the panel lands rather than the claim's subject domain.

The models do not use the confidence scale in the same way. The preceding tables concern how confidence relates to the verdict a model assigns; Table 16 instead asks whether the models agree with one another on confidence itself, and they do so only weakly. The five models differ by at least one point on 87% of claims and by three or more points on 21%, and the ordinal Krippendorff's α for confidence is 0.44, far below the 0.77 obtained for the verdicts. The panel therefore agrees less on how confident to be than on what the verdict is.

Much of this disagreement sits at the top of the scale, between adjacent values rather than across wide gaps; it is nonetheless persistent, and no coarsening of the scale removes it. A reported confidence is not comparable across models, since the same numerical value need not correspond to the same degree of certainty from one model to the next, and the panel offers no shared scale of confidence on which a downstream user could rely. This does not contradict the earlier finding that a claim's confidence floor tracks verdict disagreement, as that relationship holds in aggregate across claims, whereas the present result concerns whether the models assign similar confidence to the same claim, which they largely do not.

This pattern is consistent with prior work on calibration. Kadavath et al. [5] find that self-reported confidence is reasonably calibrated on familiar formats but degrades on novel task formats and generalises only partially across tasks. The claims examined here are recent and previously unseen, a regime in which such degradation is expected, and the confidence α of 0.44, well below the verdict α, is the corresponding signature at the level of the panel.

7 Conclusion

This study set out to measure how consistently frontier LLMs judge real-world claims, and it found that consistency to be limited. Across 997 recent user-submitted claims, the five models returned substantively divergent verdicts on 23% of cases and dissented in some form on 63%, and the panel's ordinal Krippendorff's α of 0.77 places its agreement below the threshold at which a set of raters can be treated as interchangeable. The disagreement is therefore neither negligible nor random. It is a structured and persistent feature of how these systems adjudicate claims, and it survived the methodological revisions made over the earlier version of this work.

What the two dimensions of the study share is more informative than either alone. Verdict agreement and reported confidence move together. The panel converges and reports near-maximal confidence on the polar claims, and it fractures and grows less certain across the intermediate ones, whereas the subject domain of a claim has far less bearing on either. Confidence, however, offers no reliable basis for resolving this disagreement. Although a claim's lowest reported confidence anticipates whether the panel will divide, the models do not use the confidence scale in the same way, so the same numeric value carries different weight from one system to the next.

The practical consequence is that reliance on any single frontier model inherits this disagreement without exposing it. A different system would return a materially different verdict on a substantial share of claims, and that share is largest precisely on the nuanced, partially-true claims for which an accurate judgement matters most and on which no external signal, confidence included, flags the disagreement to the user. As these systems are increasingly consulted as arbiters of factual accuracy, the identity of the model becomes a determinant of the verdict in its own right.

These conclusions concern consistency rather than correctness, which the absence of human-labelled verdicts leaves out of reach. Yet the two are linked. Because only one verdict can be *most correct* on any claim, the disagreement measured here is a lower bound on the error that at least one model must commit, and any future assessment of accuracy would inherit the same divergence as its floor. Establishing those labelled verdicts is the natural next step, and it would convert the present account of where the models disagree into an account of where, and how often, they are wrong.

8 Reproducibility

All the claim-level data for the 997 complete claims (claim id and URL, atomic claim text, the five frontier verdicts, maximum verdict distance, domain, creation date) is available in a CSV file format: https://lenz.io/research/llm-disagreement/v1.1/data.csv. This is snapshot v1.1, data as of July 18, 2026, captured on July 22, 2026. The archival page will serve the snapshot indefinitely for citation purposes. The harvesting tool used to create the underlying data is available separately as a standalone repository, described in Appendix C. Research snapshots are versioned. Each snapshot will remain indefinitely accessible at its own archive URL; each new version will be accompanied by a change log when it is released (with additional claims, updated models, etc.).

Changes from v1.0. Snapshot v1.0 (May 2026) employed a 4-point rating scale (True / Mostly True / Misleading / False) with a pre-existing set of models (GPT-5.4, Claude Opus 4.7, Gemini 3 Pro, Gemini 3 Pro + Search, Sonar Pro) based on a one-label prompt (`usr_v2`) with no structured output or extended thinking. v1.1 scales up to a 5-point rating scale (Mixed and Mostly False are added to replace the Misleading category), updates the set of models to the five listed above and employs all models at a mid-level reasoning tier (mid-level thinking, retrieval, native JSON). v1.0 is permanently stored at its versioned URL [1]; numbers cannot be compared between versions due to the difference in the scale.

9 Ethics and Data Use

Only public-facing claim fields are used: the atomic claim text and the claim's creation date. No personal data. Private and staff claims are excluded (§3.1). Frontier models received only the claim text and the as-of date — no submitter identity or analytics signal.

10 Limitations

The main limitation of this study is that it lacks human-labelled answers for the claims. The consequence of this is that we are only able to measure disagreement, not accuracy, which is ultimately a primary concern for AI models. Treating the five verdict categories as equally spaced along that ordinal scale is a simplification. Part of the disagreement measured is reflective of the task, more than the specific models, as even human annotators have been found to have a $\kappa = 0.619$ [4]. In terms of the corpus and testing, the claims are in some way selected based on Lenz's users, so it is not a direct random sampling. It might also be a concern that the claims aren't in their raw format (§3.1), limiting how representative the models' replies are of what regular users would receive. Finally, another larger issue is that we ran each claim once per model; a rerun would most likely shift the numbers due to model inconsistency [8].

11 Future Work

The most substantial direction for future work to take is establishing human-labelled "*correct*" labels for each of the claims. This would address much of the current study's limitations and broaden its scope. We are looking into getting the claims in the current corpus labelled and creating a study that also encompasses accuracy rates, as well as examining how and where inaccuracy and disagreement with human-labelling clusters. We would also ideally rerun claims with every model in future versions. This would allow us to better account for inconsistencies within each model and possibly analyse how and where they form.

Acknowledgements

We would like to thank Beloslava Malakova for her guidance in the process of writing this study. Special thanks to Simon Willison and Benjamin Han for the valuable feedback on the first revision of this research. We would also like to express our appreciation to the authors of the works included in our bibliography, which served as a valuable source of knowledge. We would also like to acknowledge that AI tools were used to develop the harvesting and aggregation infrastructure, as well as serving as an additional editor for this report.

References

[1] Kosta Jordanov. Lenz frontier-LLM disagreement benchmark (v1.0), 2026. Dataset.

[2] Premtim Sahitaj, Iffat Maab, Junichi Yamagishi, Jawan Kolanowski, Sebastian Möller, and Vera Schmitt. Towards automated fact-checking of real-world claims: Exploring task formulation and assessment with LLMs. *arXiv preprint arXiv:2502.08909*, 2025.

[3] Eddie Yang and Dashun Wang. Benchmark illusion: Disagreement among LLMs and its scientific consequences. *arXiv preprint arXiv:2602.11898*, 2026.

[4] Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.13117.

[5] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

[6] Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

[7] Klaus Krippendorff. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433, 2004. doi: 10.1111/j.1468-2958.2004.tb00738.x.

[8] Rajarshi Haldar and Julia Hockenmaier. Rating roulette: Self-inconsistency in LLM-as-a-judge frameworks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025. arXiv:2510.27106.

A Example Claims

The twenty claims with the widest spread between the verdict categories, ordered by maximum verdict distance.

[General, distance 4, no majority] *Improving a firm's image does not qualify as Further Production in economics when determining whether something is a producer good.*
Claude Fable 5: Mostly True; GPT-5.6-Sol: False; Gemini 3.1 Pro + Search: True; Sonar Deep Research: False; Grok 4.5: True

[Tech, distance 4, no majority] *There are published articles describing the use of Python-based models for dimensional optimization of river crossing bridges for flood control, which can be adapted for use on different rivers by inputting relevant parameters.*
Claude Fable 5: Mostly False; GPT-5.6-Sol: Mostly True; Gemini 3.1 Pro + Search: False; Sonar Deep Research: Mixed; Grok 4.5: True

[General, distance 4, no majority] *Equal Measures 2030's 2024 SDG Gender Index provides a downloadable dataset that includes a field labeled "required annual change".*
Claude Fable 5: Mixed; GPT-5.6-Sol: False; Gemini 3.1 Pro + Search: False; Sonar Deep Research: Mostly False; Grok 4.5: True

[General, distance 4, no majority] *The animated television series "The Boondocks" was produced with PAL video standards in mind for Season 1, and with NTSC video standards in mind for later seasons.*
Claude Fable 5: Mixed; GPT-5.6-Sol: Mostly True; Gemini 3.1 Pro + Search: False; Sonar Deep Research: False; Grok 4.5: True

[Science, distance 4, no majority] *Under ASTM D924 test conditions, the dielectric dissipation factor (power factor) of an in-service (aged) sample of Nynas Nytro 10XN transformer mineral oil at 70°C is greater than 0.01.*
Claude Fable 5: Mostly False; GPT-5.6-Sol: True; Gemini 3.1 Pro + Search: False; Sonar Deep Research: Mostly False; Grok 4.5: False

[General, distance 4, no majority] *No lossless (FLAC) archive of the album "malii" by Draft.__ exists, and the album is available only in MP3 quality.*
Claude Fable 5: Mixed; GPT-5.6-Sol: Mostly False; Gemini 3.1 Pro + Search: False; Sonar Deep Research: Mostly True; Grok 4.5: True

[General, distance 4, no majority] *Academic research on mega-event bidding and FIFA governance has given limited attention to whether the structure of FIFA's 2026 bid evaluation framework systematically favored bids with inherited commercial and infrastructural advantages.*
Claude Fable 5: Mostly True; GPT-5.6-Sol: True; Gemini 3.1 Pro + Search: False; Sonar Deep Research: True; Grok 4.5: Mostly True

[Health, distance 4, no majority] *A recommended initial course of 6–10 sessions at clinics in Auckland, New Zealand, typically totals NZ$480–NZ$1,400 before any maintenance sessions.*
Claude Fable 5: Mostly True; GPT-5.6-Sol: Mostly True; Gemini 3.1 Pro + Search: False; Sonar Deep Research: Mixed; Grok 4.5: True

[General, distance 4, no majority] *The Turkish YouTuber known as "Atkafası" delayed publishing a pre-recorded video because the sale of his home accelerated his moving timeline.*
Claude Fable 5: False; GPT-5.6-Sol: True; Gemini 3.1 Pro + Search: False; Sonar Deep Research: Mostly False; Grok 4.5: True

[General, distance 4, no majority] *Jay Chou's songs contain collectivist lyrical themes that are clear examples of collectivism in C-pop storytelling.*
Claude Fable 5: Mostly False; GPT-5.6-Sol: Mostly True; Gemini 3.1 Pro + Search: False; Sonar Deep Research: Mostly True; Grok 4.5: True

[Science, distance 4, no majority] *Physiological costs of expressing sexually selected traits (for example, elevated thermal loads) can oppose natural selection that would otherwise favor smaller or less ornamented phenotypes, especially under environmental stress.*
Claude Fable 5: True; GPT-5.6-Sol: Mostly False; Gemini 3.1 Pro + Search: False; Sonar Deep Research: Mixed; Grok 4.5: True

[Science, distance 4, no majority] *"Dihydrogen dioxide" is an incorrect or nonstandard name for hydrogen peroxide (H2O2) in chemical nomenclature.*
Claude Fable 5: Mostly False; GPT-5.6-Sol: False; Gemini 3.1 Pro + Search: True; Sonar Deep Research: Mostly False; Grok 4.5: False

[Health, distance 4, no majority] *Based on an oral LD50 greater than 2000 mg/kg body weight, EJUPAX is classified under the Globally Harmonized System as acute oral toxicity Category 5 with hazard statement H303 ("May be harmful if swallowed").*
Claude Fable 5: Mostly True; GPT-5.6-Sol: Mostly False; Gemini 3.1 Pro + Search: False; Sonar Deep Research: Mixed; Grok 4.5: True

[Legal, distance 4, no majority] *Slavery is illegal in every country in the world.*
Claude Fable 5: Mostly True; GPT-5.6-Sol: False; Gemini 3.1 Pro + Search: True; Sonar Deep Research: False; Grok 4.5: Mostly True

[Science, distance 4, no majority] *For a reversible Hamiltonian flow with reversor R, for any point x on a reversible orbit, the points x and R(x) are related by a conjugacy of the dynamics restricted to that orbit.*
Claude Fable 5: Mostly True; GPT-5.6-Sol: True; Gemini 3.1 Pro + Search: False; Sonar Deep Research: Mostly True; Grok 4.5: True

[Science, distance 4, no majority] *Chickens existed before chicken eggs existed.*
Claude Fable 5: Mixed; GPT-5.6-Sol: Mixed; Gemini 3.1 Pro + Search: False; Sonar Deep Research: Mostly False; Grok 4.5: True

[Politics, distance 4, no majority] *The United States Senate has approved a resolution halting United States Armed Forces from hostilities within or against the Islamic Republic of Iran.*
Claude Fable 5: True; GPT-5.6-Sol: Mostly False; Gemini 3.1 Pro + Search: True; Sonar Deep Research: Mostly False; Grok 4.5: False

[Finance, distance 4, no majority] *Financial Ombudsman Service guidance says that deception (being tricked into authorising a payment) is the key factor when assessing protection or reimbursement for authorised payments, rather than whether the payer pressed a 'confirm' button.*
Claude Fable 5: Mostly True; GPT-5.6-Sol: Mostly True; Gemini 3.1 Pro + Search: True; Sonar Deep Research: Mostly False; Grok 4.5: False

[Politics, distance 4, no majority] *Galab Donev said that previous Bulgarian governments decided that allocation of money under Bulgaria's EU Recovery and Resilience Plan would be preceded by reforms on which the funds depend.*
Claude Fable 5: True; GPT-5.6-Sol: Mostly True; Gemini 3.1 Pro + Search: False; Sonar Deep Research: Mostly False; Grok 4.5: True

[Legal, distance 4, no majority] *Courts in Sierra Leone recognize the doctrine of agency of necessity as a legal basis for imposing a spouse's financial obligation to pay for the other spouse's necessaries.*
Claude Fable 5: Mostly True; GPT-5.6-Sol: Mostly True; Gemini 3.1 Pro + Search: False; Sonar Deep Research: Mostly False; Grok 4.5: True

B Rejected Claims

These are the three claims on which neither Claude Fable 5 nor, where it was called, the Claude Opus 4.8 fallback returned a usable verdict; they are therefore excluded from the 997-claim cohort.

[Science] *Marsupial mice (antechinuses) are found in eastern and southeastern Australia, including parts of Queensland, New South Wales, and Victoria.*
Verification ID: 4af188ca; errored cell: claude-fable-5; lenz.io/c/antechinus-distribution-eastern-australia-4af188ca

[Science] *Linalyl acetate and alpha-bisabolol acetate can destabilize the outer membrane of Xanthomonas citri subsp. citri.*
Verification ID: 6b59c840; errored cell: claude-fable-5; lenz.io/c/linalyl-alpha-bisabolol-acetate-xanthomonas-membrane-6b59c840

[Tech] *A Sony PlayStation 4 can be jailbroken on system software version 13.50.*
Verification ID: 9de4c899; errored cell: claude-fable-5; lenz.io/c/ps4-system-software-13-50-jailbreak-9de4c899

C Complete Code

What is released. All the code required to reproduce the harvest is public. The harness that generated it, along with the 1,000-claim corpus and the raw results, is available at https://github.com/lenzhq/lenz-research. The harvest consists of 5,000 rows, one for each (claim, model) cell, holding the verdict of the model, the confidence score 1 to 10, the text of the reasoning, the sources found, the cost and latency of inference, in JSON and spreadsheet format. No verdict label, no submitter identity, and no analytics signal accompany it, as prescribed by §9. Since every row represents a cell, the release includes no cross-model comparisons; the claim-level statistics reported in this paper come from the snapshot of §8, which is also the source of all statistics and tables shown here — none of which is computed on a live run. The setup and file structure, along with costs per model, are described in the README of the repository.