LENZ RESEARCH · SNAPSHOT V1.0 · DATA AS OF MAY 21, 2026

Beyond Benchmarks: Disagreement Among Frontier LLMs on Real-World Fact-Checks

Kosta Jordanov · Lenz · kosta@lenz.io

We presented 1,000 recent real user claims to the five top frontier LLMs and asked each one for a verdict. These aren't benchmark items with public answer keys — they're claims real users submitted for verification to a fact-checking platform. Only one verdict bucket can be correct per claim, so any disagreement among the panel means at least one model's verdict is **label-inconsistent under this 4-bucket rubric** (True / Mostly True / Misleading / False). On 67% of claims, the panel splits.

KEY FINDINGS

- **67% of claims** (672 / 1,000) have at least one frontier model dissenting from the panel majority — or no majority forms at all.
- **34% of claims** (343 / 1,000) involve a 2+ bucket gap between the most-disagreeing pair of frontier verdicts — a substantive disagreement on the answer, not just a calibration shift.
- **The panel converges on definitive verdicts; the middle of the rubric is where it fractures.** Within the 328 unanimous claims, only 4 are unanimous-Misleading and 0 are unanimous-Mostly-True.
- **Some models concentrate verdicts at the True/False poles;** others spread across the middle two buckets.

1. How often the frontier disagrees

On **67% of claims** (672 / 1,000; 95% CI: 64–70%), the frontier panel doesn't agree — at least one model dissents from the majority verdict, or no strict majority forms at all. The breakdown:

For each claim we looked at the five frontier verdicts and asked: did at least three pick the same answer (a strict majority)? If yes, how many of the remaining models dissented? If no clear majority emerged at all — verdicts split across three or four different buckets — the claim falls in the *Models split, no majority* row. Most of these claims are unlikely to appear in any training corpus with a gold label attached — there's no canonical answer key to pattern-match against, no benchmark leaderboard to anchor to.

We refer below to the "majority" and to "dissent from the majority." A majority of frontier models is not ground truth. The majority verdict is sometimes wrong; an individual dissenting model is sometimes

right. We use the majority as a structural reference point for measuring disagreement, not as a stand-in for correctness.

Frontier verdict pattern	Claims	Share of corpus
All 5 agreed (unanimity)	**328**	**33%** 30–36%
1 of 5 dissented	**224**	**22%** 20–25%
2 of 5 dissented	**316**	**32%** 29–35%
Models split, no majority (e.g. 2-2-1 or 2-1-1-1)	**132**	**13%** 11–15%
≥1 model dissents (incl. splits)	**672**	**67%** 64–70%
≥2 models dissent (incl. splits)	**448**	**45%** 42–48%

Panel agreement: Krippendorff's α (ordinal) = 0.639 (n=1000 claims, 5 raters). This indicates nontrivial but limited agreement: the models' verdicts are structured rather than random, but not consistent enough to treat the panel as a single interchangeable judge. Ordinal α is the standard Krippendorff variant for an ordered categorical scale (True / Mostly True / Misleading / False). See §7.5 Statistical analysis for choice of metric.

Lower bound on model error. For each claim, exactly one of the four verdict buckets is the correct answer. If we assume the panel's most popular bucket is the correct one — the most charitable assumption — the minimum number of models that picked a wrong verdict is:

> **≥1 model wrong** on 67% of claims (any non-unanimous panel)
>
> **≥2 wrong** on 45% of claims (3-2, 3-1-1, or no-majority splits)
>
> **≥3 wrong** on 13% of claims (no bucket reaches a majority, so at most 2 can be right)

Relaxing the "most popular is correct" assumption can only *raise* these counts, never lower them. The actual error rates are likely higher still: even the 33% of cases where all five agree can and likely does include shared blind spots.

2. Substantive vs nuance disagreement

On **34% of claims** (343 / 1,000; 95% CI: 31–37%), at least two frontier models pick verdicts that are 2 or more buckets apart in our 4-bucket rubric — a disagreement that goes beyond calibration.

Not every disagreement is equal. A "True" vs "Mostly True" split is a confidence-calibration shift. A "True" vs "False" split is a substantive disagreement about the answer. We measure this as the **max pairwise bucket distance** across the 5 verdicts on each claim, where the verdicts are ordered `True (0) → Mostly True (1) → Misleading (2) → False (3)`.

Distance	Interpretation	Claims	Share
0	Full unanimity (all 5 picked the same bucket)	328	33% 30–36%
1	Nuance only (e.g. True ↔ Mostly True)	329	33% 30–36%
2	Substantive (True ↔ Misleading, or Mostly True ↔ False)	132	13% 11–15%
3	Polar (True ↔ False)	211	21% 19–24%
≥2 buckets apart (substantive or polar)		343	34% 31–37%

Caveat. Bucket distance treats True / Mostly True / Misleading / False as an ordinal scale; an equal-spaced interpretation is a simplification. A 2-bucket gap can still reflect rubric ambiguity, temporal-framing differences, or differing interpretations of "Misleading." We report it as a coarse "substantive vs nuance" indicator, not a metric of error magnitude.

3. Model-vs-model agreement

Highest peer agreement: **Gemini 3 Pro × Gemini 3 Pro + Search (75%)** — unsurprising, since they share a base model. Lowest: **Claude Opus 4.7 × Gemini 3 Pro, Claude Opus 4.7 × Gemini 3 Pro + Search and Gemini 3 Pro × Sonar Pro (53%)** — three pairs tie at the floor.

How often each pair of frontier models picked the same verdict label, across all claims in the corpus.

	GPT-5.4	Claude Opus 4.7	Gemini 3 Pro	Gemini 3 Pro + Search	Sonar Pro
GPT-5.4	—	65% 62–68%	65% 62–68%	60% 57–63%	60% 57–63%
Claude Opus 4.7	65% 62–68%	—	53% 50–56%	53% 50–56%	58% 55–61%
Gemini 3 Pro	65% 62–68%	53% 50–56%	—	75% 72–77%	53% 50–56%
Gemini 3 Pro + Search	60% 57–63%	53% 50–56%	75% 72–77%	—	58% 55–61%

	GPT-5.4	Claude Opus 4.7	Gemini 3 Pro	Gemini 3 Pro + Search	Sonar Pro
Sonar Pro	60% 57–63%	58% 55–61%	53% 50–56%	58% 55–61%	—

4. Per-model behavior

Two angles on the same five models: how each one distributes its verdicts (4.1), and how often each one's verdict matches the strict majority of the other four (4.2).

4.1 Verdict distribution

Some models concentrate verdicts at the True/False poles; others distribute more broadly across the middle two buckets. This reflects model-level decision priors interacting with the specific claims — without ground truth, we can't separate the two. The table below shows the share of claims each model assigned to each bucket, with 95% Wilson CIs underneath each cell.

Model	True	Mostly True	Misleading	False
GPT-5.4	42% 39–45%	16% 14–19%	12% 10–14%	30% 28–33%
Claude Opus 4.7	38% 35–41%	26% 23–29%	19% 17–22%	17% 15–20%
Gemini 3 Pro	54% 51–57%	3% 2–4%	3% 2–4%	40% 37–43%
Gemini 3 Pro + Search	52% 49–55%	4% 3–5%	9% 7–11%	35% 32–38%
Sonar Pro	35% 32–38%	23% 21–26%	16% 14–18%	26% 23–28%

4.2 Agreement with the rest of the panel

Across the five models, peer-majority agreement ranges from **69% to 81%**. This is peer-alignment in this corpus, not correctness — no model is treated as ground truth here, and *eligible n* differs per row.

For each model, how often does its verdict match the strict majority (≥3/4) of the other four? A claim is eligible only when a ≥3/4 majority exists among the other four.

Model	Agreement w/ peer majority	Eligible *n*	Ineligible	Tier
GPT-5.4	81% 78–84%	650	350	parametric

Model	Agreement w/ peer majority	Eligible *n*	Ineligible	Tier
Claude Opus 4.7	70% 67–74%	691	309	parametric
Gemini 3 Pro	77% 74–80%	683	317	parametric
Gemini 3 Pro + Search	76% 73–79%	693	307	retrieval
Sonar Pro	69% 66–73%	675	325	retrieval

5. Detailed results

5.1 Per-domain frontier disagreement

Denominator per row: claims in that domain (the *Claims* column).

Domain	Claims	Any disagreement	Substantive (≥2 buckets)	No majority
Finance	75	67% 55–76%	39% 28–50%	20% 13–30%
General	179	68% 60–74%	40% 33–48%	12% 8–17%
Health	171	71% 64–78%	29% 23–36%	12% 8–17%
History	131	53% 44–61%	24% 17–32%	13% 8–20%
Legal	48	77% 63–87%	40% 27–54%	19% 10–32%
Politics	168	70% 62–76%	38% 31–46%	8% 5–13%
Science	151	68% 60–75%	36% 29–44%	21% 15–28%
Tech	77	69% 58–78%	31% 22–42%	8% 4–16%

5.2 Per-verdict panel agreement

When the panel does land on a middle bucket, it almost never converges. Mostly True and Misleading majorities reach unanimity at most 5% of the time, vs 43–47% for True and False

majorities.

Consistent with this, work on a different real-world corpus (17,856 PolitiFact claims with a single-family Llama-3 ablation, Schwab et al. 2025) finds nuanced labels are where fact-check verdict models concentrate their errors — a related observation from a different methodological setup (single-family ablation, not a frontier panel).

Denominator: claims with a strict ≥3/5 frontier majority on this verdict. Reveals which verdict zones the panel is most/least confident about.

Majority verdict	Eligible n	Unanimous (5/5)	Majority only (3-4 of 5)
True	438	47% 42–51%	53% 49–58%
Mostly True	76	0% 0–5%	100% 95–100%
Misleading	74	5% 2–13%	95% 87–98%
False	280	43% 37–49%	57% 51–63%

Viewed from the other direction — of the 328 claims where all 5 frontier models converged on the same verdict, the distribution across verdicts:

Unanimous verdict	Claims	Share of unanimous
True	204	62% 57–67%
Mostly True	0	0% 0–1%
Misleading	4	1% 0–3%
False	120	37% 32–42%

6. Data

1,000 claims — the most recent real-world user submissions to a fact-checking platform that pass every eligibility filter listed under Exclusions below. None of these claims is older than February 15, 2026. Unless otherwise stated, every metric on this page uses this set as its denominator; tables that use a different denominator (e.g. claims with a strict ≥3/5 frontier majority on a verdict) state it inline.

Provenance

These claims were submitted to **Lenz**, a fact-checking platform. We chose this corpus because it represents organic real-world fact-check requests rather than curated benchmark items. **Lenz's own verdict on each claim is not used in this analysis** — this paper measures frontier-model disagreement only, not Lenz vs the frontier.

Claim normalization

The `atomic_claim` field in the CSV is not the user's raw submission. It's the output of Lenz's framing step, which strips emotional language and bias and distills the input into a single neutral, testable proposition anchored to the submission date. Frontier models were rated against the framed claim, not the raw text. A user who types *"Canadian authorities are throwing Christians in jail for quoting the Bible!!!"* is rated on the proposition *"As of April 4, 2026, Canadian authorities have jailed individuals for publicly quoting the Bible because of their Christian beliefs."*

Exclusions

The corpus excludes:

- Claims marked private by the submitting user
- Claims from platform staff, internal accounts, or agent/API submissions (only real user web submissions appear in the corpus)
- Claims with editorial status `pending` (not yet reviewed) or `hidden` — either depublished after editorial review or auto-flagged at submission time by Lenz's PII screening step (containing personal information about non-public individuals)
- Near-duplicate claims — pairs within a cosine distance of `0.2` on OpenAI `text-embedding-3-small` embeddings (1536-dim) of the `atomic_claim` are collapsed to a single canonical row. The newer claim becomes canonical when the proposition is time-dependent; otherwise the existing claim with the most views on Lenz wins. Only canonicals appear in this corpus.
- Claims where at least one of the five frontier models failed to produce a parseable verdict, even after one retry. Most of these residual errors come from Gemini's grounded-search API, which occasionally returns malformed responses; the rest are rare Anthropic refusals. Only claims with all five models successful are included in the cohort.
- Claims older than 180 days (recency window applied at harvest time)

7. Methodology

7.1 Model selection

Five frontier models, chosen to cover two capability surfaces:

- **Parametric** (training-only): GPT-5.4 (OpenAI), Claude Opus 4.7 (Anthropic), Gemini 3 Pro (Google)
- **Retrieval-augmented**: Gemini 3 Pro + Search (Google), Sonar Pro (Perplexity)

7.2 Prompt

Each claim is presented with an "as of YYYY-MM-DD" anchor matching the submission date, asking the model to pick one of four verdicts:

```
Classify this claim as of <date>: "<atomic claim>"

Output exactly one label: True, Mostly True, Misleading, or False.
No explanations, no qualifiers.
```

Verbatim prompt template version: `usr_v2`. No Abstain option is offered (a forced choice keeps the comparison symmetric across models). Unparseable outputs are not reclassified into a verdict bucket; claims with any parse error are excluded from the complete-claim cohort.

7.3 LLM call configuration

All five models received the same system placeholder (`.`) and the same user prompt template (`usr_v2`). No structured-output schema, tool-call schema, seed, top-p, or logit-bias controls were used. The harvester requested deterministic decoding where supported (`temperature=0.0`); GPT-5.4 and Claude Opus 4.7 were called without an explicit temperature because their provider adapters reject custom temperature settings. Output length was capped at 16 tokens for GPT-5.4, Claude Opus 4.7, and Sonar Pro; Gemini 3 Pro and Gemini 3 Pro + Search used a 1024-token cap (lower caps produced provider-side errors during harvester development). Gemini 3 Pro + Search enabled Google Search grounding; Sonar Pro was treated as retrieval-augmented through Perplexity's search-backed API. Parseable outputs had to equal exactly one of the four labels after normalization.

7.4 Grading

No LLM grader. All measurements derive from direct parsed-label equality between the 5 frontier verdicts on the same claim. No reference label or ground truth is used.

7.5 Statistical analysis

Sampling frame & inferential target. The corpus is the 1,000 most recent eligible claims submitted to this single fact-checking platform (per the filters in §6) — not a probability sample from any wider population, and not a complete enumeration (older eligible claims exist but are excluded by the cap). Reported Wilson 95% CIs are nominal binomial intervals under a model where each claim is an independent draw from a hypothetical stream of similar eligible submissions to this same platform under the same screening rules. They are not coverage statements about "all real-world fact-checks."

Non-iid caveat. Lenz claims are not independently and identically distributed: users cluster submissions around news events, screening selects for certain topics, and individual users often submit multiple related claims in a single session. True sampling variability under a more honest cluster model (e.g. cluster bootstrap) would likely be larger than what Wilson reports. We surface CIs as a *minimum precision floor*, not a guaranteed coverage interval.

Wilson 95% confidence intervals on every reported rate. We use the Wilson score interval [1] rather than the Wald (normal-approximation) interval because it has better small-N behavior and handles boundary cases (`p=0/n`, `p=n/n`) without producing degenerate zero-width intervals. It is the de-facto standard in modern ML evaluation literature. Wilson CIs appear inline next to every rate in §1, §2, §3, §4.2, §5, and the appendix; the printed bounds are exact, not centered on the raw point estimate.

Inter-rater reliability — Krippendorff's α (ordinal). The verdict scale (True / Mostly True / Misleading / False) is ordinal, so we score with Krippendorff's α at the ordinal level of measurement [2] rather than Fleiss' κ (which treats categories as nominal and would underestimate agreement — a True ↔ Mostly True 1-bucket disagreement is much smaller than a True ↔ False polar split, and the ordinal metric reflects that). α is reported as a single panel-level number alongside the §1 results table.

No model-vs-model significance testing. We report pairwise agreement rates with 95% Wilson CIs as descriptive statistics rather than treating the page as a model leaderboard. Pairwise significance tests are sensitive to the comparison target and eligibility set: for example, peer-majority agreement is a paired claim-level outcome, but each model has a different set of claims where the other four models form a strict majority.

References.

[1] Wilson, E.B. (1927). "Probable inference, the law of succession, and statistical inference." *Journal of the American Statistical Association* 22, 209–212.
[2] Krippendorff, K. (2004). "Reliability in Content Analysis: Some Common Misconceptions and Recommendations." *Human Communication Research* 30(3), 411–433.

8. Reproducibility

Full per-claim data: **download CSV**. One row per claim — claim ID and URL, atomic claim text, the 5 frontier verdicts, max pairwise bucket distance, domain, and creation date. Strictly rectangular, no preamble comments. The `claim_url` column links each row back to the original claim page on Lenz; some pages may be unavailable if the user who submitted the claim later deleted or privatized it.

PDF artifact: **download PDF**. Browser-independent rendering of this page for offline reading, citation, or arxiv-style preprint hosting. Hash-pinned in the snapshot manifest (`pdf_sha256`) so the PDF served at `/v1.0/pdf` is byte-identical across re-deploys.

This snapshot is **v1.0**, data as of May 21, 2026. The archival URL `/research/llm-disagreement/v1.0` permanently serves the v1.0 snapshot — citation-stable even when the bare URL bumps to a future version.

Harvester prompt version: `usr_v2`. Grader: direct parsed-label equality across the 5 frontier verdicts. No LLM grader, no reference verdict.

9. Limitations

- The pigeonhole rate is a *floor* on rubric inconsistency, not an absolute "model X is factually wrong" judgement on any specific claim. Only one of {True, Mostly True, Misleading, False} can be the correct bucket, so any disagreement implies at least one inconsistent verdict — but we don't claim to know which model is wrong on which claim.
- **Bucket-distance ordinality.** §2 treats True / Mostly True / Misleading / False as an equal-spaced ordinal scale. That's a simplification. A 2-bucket gap can reflect rubric ambiguity, temporal-framing differences, or different interpretations of "Misleading" — not necessarily a factually larger error.
- **Verdict ambiguity is partly a task property, not just an LLM property.** On a major published real-world fact-check corpus (AVeriTeC, 4,568 claims annotated through multi-round review against 50 fact-checking organizations), inter-annotator agreement on verdicts reaches κ=0.619 — substantial but well short of perfect. Some fraction of frontier disagreement on this page reflects underlying difficulty of these labels for any rater, not just for LLMs.
- The eligibility filter in §4.2 (strict ≥3/4 majority among the other four) excludes claims where the other four split. *eligible_n* varies per row; the disclosed counts make this explicit.
- Domain stratification reflects the source platform's traffic patterns, not a uniform sampling of fact-checkable claims generally.
- This is a snapshot, frozen on a specific date with specific model versions. Frontier LLMs are non-deterministic, so even a re-run with the same models and prompts would produce somewhat different numbers; re-running with newer models or different prompts moves the numbers more. That's why we version the snapshot and ship an archival URL.
- Retrieval-enabled models may have looked up sources at inference time. We do not control or audit what they retrieved.

10. FAQ

Why no Lenz-vs-frontier comparison? You're a fact-checking platform.

A meaningful accuracy comparison requires human-labeled ground truth. We're working on a follow-up study (see below) that human-labels every claim in this corpus and compares both the frontier panel and Lenz's own verdicts against those labels. Until that ships, we'd rather publish nothing about Lenz's relative accuracy than publish a comparison that can't actually answer "who's right." This paper measures only what is measurable without ground truth: how the frontier panel behaves on real-world claims.

Has anyone measured frontier-LLM disagreement before?

Yang & Wang (2026) show top frontier models disagree on 16-38% of MMLU-Pro and GPQA items even at matched aggregate accuracy, and demonstrate that switching the annotation model in downstream scientific re-analyses can flip estimated treatment-effect signs. On real-world claim verification with rigorous human annotation, the canonical reference is AVeriTeC (4,568 fact-checked claims, multi-round annotation against 50 organizations, inter-annotator κ=0.619). Larger fact-check corpora exist — for example, 17,856 PolitiFact claims under a single-family Llama-3 ablation.

Why not use a standard fact-checking benchmark like AVeriTeC instead of building a corpus?

Two reasons. First, AVeriTeC, PolitiFact, and similar fact-check corpora have been publicly available for years and almost certainly appear in current frontier-model training data — measured disagreement on them confounds true inference disagreement with memorization. Lenz's corpus is structurally fresh: real-user submissions from the past 180 days, indexed only on lenz.io, never paired with canonical verdicts in any public training set. Second, those corpora draw from a narrower distribution (political claims from US-centric fact-checkers, often pre-screened for newsworthiness) than what real users actually ask about — Lenz claims span health, science, finance, history, tech, and legal questions in the same 4-bucket rubric.

What about benchmark contamination — did the models see these claims during training?

These are recent real-user submissions, not curated benchmark items from SimpleQA, TruthfulQA, FActScore, or other public datasets. Some claims may overlap topically with material seen in training, but they aren't paired with canonical answer keys the way benchmark items are. Retrieval-enabled models *can* still find sources on the live web — including Lenz's own public claim pages — though this corpus isn't a controlled contamination audit.

Why these five models?

Three frontier parametric models (GPT-5.4, Claude Opus 4.7, Gemini 3 Pro) and two retrieval-augmented (Gemini 3 Pro + Search, Sonar Pro). The split is intentional — it covers both inference modes that are common in production AI systems.

Why four buckets instead of five (with Abstain)?

In preliminary harvesting we saw some frontier models routinely decline to answer harder claims while others always committed to a verdict. An Abstain bucket would have made cross-model comparison asymmetric — a model's "I won't answer" lands in a different category from another model's confident-but-wrong answer, even though both behaviors carry epistemic weight. We force the same 4-choice set on every model so the disagreement we measure is over verdicts, not over willingness-to-verdict.

Will you re-run this?

This is a frozen snapshot (v1.0, data as of May 21, 2026). The archival URL `/research/llm-disagreement/v1.0` will always serve this exact version. When v2 ships — with more claims, refreshed model versions, or methodology changes — it'll appear with a clear changelog entry; v1.0 stays at its archival URL.

What's the planned follow-up?

We're working on a companion study that human-labels every claim in this corpus and uses those labels as ground truth to evaluate *both* the five frontier models *and* Lenz's own verdict. The point isn't a leaderboard. The point is to map the *structure* of disagreement: where do frontier panels systematically diverge from a human consensus, where does Lenz diverge from both, how each individual model and Lenz align with the same human reference, and what categories of claims drive each kind of divergence (rubric ambiguity, temporal framing, domain specialization, calibration drift). The current paper says *that* the frontier disagrees on real-world claims; the follow-up will say *how*, on the same corpus, with humans as the reference.

11. Ethics & 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 per §6. Frontier models received only the claim text and the as-of date — no submitter identity, no analytics signal.

If a claim is later privatized or deleted by its submitter, we can drop it from this snapshot and from any future downloads. The CSV is generated from the snapshot at download time, so removing a claim from the snapshot removes it from the public page in a single update.

12. Changelog

- **v1.0** (May 21, 2026, code `a6b78be`): initial frozen snapshot. Frontier-disagreement only; no Lenz-vs-frontier comparison.

Appendix: Example claims where the frontier fractures

The twenty claims in this corpus with the widest spread between the highest- and lowest-bucket frontier verdicts. These are claims where the panel doesn't just disagree — it disagrees substantively, with at least one model picking a verdict ≥2 buckets away from another.

Ordered by max pairwise bucket distance (descending), no-majority cases tie-broken first, then by stable hash of the claim ID. Deterministic — the page renders the same examples on every load until the next snapshot.

Muthiah Muralidaran said that the Indian Premier League is purely a business and that flat pitches are prepared because low-scoring matches are boring for sponsors. →

Politics · max bucket distance: **3** · no majority

GPT-5.4	CLAUDE OPUS 4.7	GEMINI 3 PRO	GEMINI 3 PRO + SEARCH	SONAR PRO
True	Mostly True	False	Misleading	Misleading

The World Bank's active portfolio in Nigeria stands at over $16.4 billion as of 2025. →

Finance · max bucket distance: **3** · no majority

GPT-5.4	CLAUDE OPUS 4.7	GEMINI 3 PRO	GEMINI 3 PRO + SEARCH	SONAR PRO
Mostly True	True	False	Misleading	Misleading

Individuals who prefer music with less positive emotional content tend to have higher intelligence. →

Science · max bucket distance: **3** · no majority

GPT-5.4	CLAUDE OPUS 4.7	GEMINI 3 PRO	GEMINI 3 PRO + SEARCH	SONAR PRO
Misleading	Mostly True	False	True	Misleading

Humans systematically overestimate short time intervals. →

Science · max bucket distance: **3** · no majority

GPT-5.4	CLAUDE OPUS 4.7	GEMINI 3 PRO	GEMINI 3 PRO + SEARCH	SONAR PRO
Mostly True	Mostly True	True	True	False

There exist published research papers on unsupervised regime identification in multivariate oceanic current time series, particularly focusing on coastal regions and methods that infer the number of regimes from data, which are relevant for forecasting applications in areas such as Bahia de Santos, Brazil. →

Science · max bucket distance: **3** · no majority

GPT-5.4	CLAUDE OPUS 4.7	GEMINI 3 PRO	GEMINI 3 PRO + SEARCH	SONAR PRO
True	Mostly True	True	False	False

Equal Measures 2030's 2024 SDG Gender Index provides a downloadable dataset that includes a field labeled "required annual change". →

General · max bucket distance: **3** · no majority

GPT-5.4	CLAUDE OPUS 4.7	GEMINI 3 PRO	GEMINI 3 PRO + SEARCH	SONAR PRO
True	Mostly True	True	False	False

The FDI World Dental Federation confirms that daily oral hygiene routines, including mouthwash use, significantly reduce the incidence of gingivitis, periodontal disease, and dental caries. →

Health · max bucket distance: **3** · no majority

GPT-5.4	CLAUDE OPUS 4.7	GEMINI 3 PRO	GEMINI 3 PRO + SEARCH	SONAR PRO
Mostly True	Mostly True	True	False	Misleading

A group known as "Khanna Coolies" operated as bicycle-riding food porters delivering meals in Calcutta. →

History · max bucket distance: **3** · no majority

GPT-5.4	CLAUDE OPUS 4.7	GEMINI 3 PRO	GEMINI 3 PRO + SEARCH	SONAR PRO
Misleading	Misleading	False	False	True

Hostels in Kota, Rajasthan commonly use caged ceiling fans as a preventive measure against student suicides. →

General · max bucket distance: **3** · no majority

GPT-5.4	CLAUDE OPUS 4.7	GEMINI 3 PRO	GEMINI 3 PRO + SEARCH	SONAR PRO
Mostly True	True	False	Misleading	False

A study led by Yadan Li at Southwest University in Chongqing found that exposure to frightening images and sounds at night (20:00) produced greater increases in skin conductance, heart rate, and blood pressure than the same exposure during the day (08:00), regardless of room lighting conditions. →

Science · max bucket distance: **3** · no majority

GPT-5.4	CLAUDE OPUS 4.7	GEMINI 3 PRO	GEMINI 3 PRO + SEARCH	SONAR PRO
True	Misleading	True	False	Misleading

Generator performance standards parameters are the responsibility of the Network Planning and Design department, not the Asset Management department. →

Tech · max bucket distance: **3** · no majority

GPT-5.4	CLAUDE OPUS 4.7	GEMINI 3 PRO	GEMINI 3 PRO + SEARCH	SONAR PRO
False	Misleading	True	False	Misleading

Teams in the esports game Valorant that select agent compositions with balanced roles such as duelist, controller, initiator, and sentinel have a higher probability of winning compared to teams with unbalanced compositions, according to statistical analysis of professional match data as of April 2026. →

Science · max bucket distance: **3** · no majority

GPT-5.4	CLAUDE OPUS 4.7	GEMINI 3 PRO	GEMINI 3 PRO + SEARCH	SONAR PRO
Misleading	Mostly True	True	False	Mostly True

Kenyan President William Ruto has stated that Kenya has a total of 20,000 kilometers of tarmacked (paved) roads. →

Politics · max bucket distance: **3** · no majority

GPT-5.4	CLAUDE OPUS 4.7	GEMINI 3 PRO	GEMINI 3 PRO + SEARCH	SONAR PRO
False	Misleading	Mostly True	True	True

SIGMAS raised a $1 million seed funding round in 2026, co-led by Mucker Capital and HongShan Capital (formerly Sequoia China). →

Finance · max bucket distance: **3** · no majority

GPT-5.4	CLAUDE OPUS 4.7	GEMINI 3 PRO	GEMINI 3 PRO + SEARCH	SONAR PRO
False	Misleading	False	True	True

The diagnostic literature on autism describes autistic people who are frequently devastated by accidentally breaking social rules they were trying hard to follow. →

Health · max bucket distance: **3** · no majority

GPT-5.4	CLAUDE OPUS 4.7	GEMINI 3 PRO	GEMINI 3 PRO + SEARCH	SONAR PRO
True	True	False	False	Mostly True

Donald Trump said that an attack on Iran was postponed at the request of Gulf allies. →

Politics · max bucket distance: **3** · no majority

GPT-5.4	CLAUDE OPUS 4.7	GEMINI 3 PRO	GEMINI 3 PRO + SEARCH	SONAR PRO
False	Mostly True	False	True	True

Wildlife species in Vietnam, including elephants, rhinoceroses, and tigers, face significant threats from habitat loss and are classified as endangered. →

Science · max bucket distance: **3** · no majority

GPT-5.4	CLAUDE OPUS 4.7	GEMINI 3 PRO	GEMINI 3 PRO + SEARCH	SONAR PRO
Mostly True	True	False	False	Mostly True

Volodymyr Zelensky was nominated for the Nobel Peace Prize for 2026. →

Politics · max bucket distance: **3** · no majority

GPT-5.4	CLAUDE OPUS 4.7	GEMINI 3 PRO	GEMINI 3 PRO + SEARCH	SONAR PRO
False	Mostly True	False	True	True

Amadeo historically served as a logistical transition point between the urbanized lowlands and the mountainous hinterlands of Cavite, Philippines. →

History · max bucket distance: **3** · no majority

GPT-5.4	CLAUDE OPUS 4.7	GEMINI 3 PRO	GEMINI 3 PRO + SEARCH	SONAR PRO
Mostly True	Mostly True	True	False	True

As of May 6, 2026, Muslims from multiple countries have gathered in Hooghly district, West Bengal, India. →

Politics · max bucket distance: **3** · no majority

GPT-5.4	CLAUDE OPUS 4.7	GEMINI 3 PRO	GEMINI 3 PRO + SEARCH	SONAR PRO
True	Mostly True	False	Misleading	True

On 67% of real-world user fact-checks in this corpus, the five strongest frontier LLMs disagree. Rely on any single one and you inherit that disagreement.

Snapshot v1.0 · data as of May 21, 2026 · code `a6b78be`. Citation-stable archive: **/research/llm-disagreement/v1.0**. Full per-claim CSV: **data.csv**. PDF: **pdf**.