Evaluating Language Models for Social-Media Fact-Checking

A Cross-Vendor Concordance Study

Perclaim Research

May 14, 2026
(Revised May 19, 2026)

Contents

Abstract

We replayed 157 production fact-checks from the Perclaim corpus against five language models — Claude Sonnet 4.6, Claude Haiku 4.5, GPT-5.4-mini, gpt-oss:120b (local), and gemma4:26b (local) — to evaluate the accuracy–cost frontier for LLM-based fact-checking. Using the most accurate available model (Sonnet 4.6) as proxy ground truth, we decomposed inter-model disagreements into calibration noise (1-notch verdict shifts), orthogonal category swaps (OPINION ↔ UNVERIFIABLE), and substantive factual disagreements (2+ notches on the factual spectrum). Manual adjudication of substantive disagreements revealed three distinct findings: (1) the retrieval-equipped frontier models cluster tightly with 89.5–96.5% effective concordance on factual content and produce nearly identical conclusions with different calibration choices; (2) Haiku 4.5 makes zero factual errors against Sonnet on substantive disagreements but suffers an 8.9% JSON parse failure rate that blocked production use at the time of the original study; (3) offline candidates without web search exhibit systematic false-negative bias — 92–95% of their "FALSE" verdicts land on rows where the retrieval-equipped reference identifies real, verifiable content. We conclude that the divide between retrieval-equipped and retrieval-free models is categorical, not gradational.

May 19 follow-up (§6). Subsequent engineering work on Haiku's structured-output reliability — a parser-side cascade-recovery system combining regex-based verdict extraction with prose-refusal synthesis — reduced the null-verdict rate from 8.9% to 1.6% across a 188-row post-rollout production corpus, an 82% reduction. Haiku 4.5 is now production-viable and is deployed as the cross-vendor backup in Perclaim's production stack. The updated deployment recommendation (§6.4) supersedes the May 14 recommendation in §5.4.

1. Introduction

Perclaim is a fact-checking application that analyzes social-media posts and produces verdicts on a six-point factual spectrum (TRUE, MOSTLY_TRUE, MIXED, MISLEADING/MOSTLY_FALSE, FALSE) plus two orthogonal categories (OPINION, UNVERIFIABLE). The system runs in production against user-submitted posts via a Chrome extension and a web interface, currently using GPT-5.4-mini as its inference model with OpenAI's hosted `web_search` tool for current-events grounding.

Three questions motivated this evaluation:

1. **Accuracy–cost tradeoff for the free tier.** Is GPT-5.4-mini the right production model, or could a cheaper or more accurate alternative serve the same role?
2. **Pro-tier candidacy.** With Stripe integration on the roadmap, which model best justifies premium pricing through accuracy lift on hard cases?
3. **Offline viability.** Can locally-hosted open-weight models substitute for cloud frontier models, potentially eliminating per-request costs?

This study reports findings from a head-to-head replay of 157 production fact-checks across five models spanning three categories: retrieval-equipped frontier (Sonnet 4.6, Haiku 4.5, GPT-5.4-mini), retrieval-free reasoning (gpt-oss:120b), and retrieval-free multimodal (gemma4:26b).

2. Background

2.1 Fact-checking with LLMs

Current-events fact-checking presents a structural challenge for language models: training data has a cutoff date, but the claims requiring verification are typically about events post-cutoff. Two architectural responses dominate:

- **Hosted web search tools** integrate retrieval into the model's inference loop. OpenAI offers `web_search`; Anthropic offers `web_search_20260209` (with code-executed dynamic filtering on supported models) and `web_search_20250305` (legacy, broader model support).
- **Retrieval-Augmented Generation (RAG)** retrieves context app-side before inference. Perclaim currently does not use app-side retrieval; this was identified as a v2 prerequisite for local-model viability.

Perclaim's production architecture currently delegates retrieval to the model via OpenAI's `web_search` tool. Source URLs and citations in the produced verdict are emitted by the model, with downstream classifiers (`lib/sources-classifier.ts`, `lib/post-url-classifier.ts`) catching hallucinated or malformed source references.

2.2 The retrieval tool version question

During provider integration, we discovered that Anthropic's newer `web_search_20260209` tool — which uses code-executed dynamic filtering for higher-quality retrieval — is only supported on a subset of frontier models (`claude-opus-4-7`, `claude-opus-4-6`, `claude-sonnet-4-6`, `claude-mythos-preview`). Haiku 4.5 returns a 400 error with this tool but works with the legacy `web_search_20250305`. The provider implementation routes per-model via a `PTC_CAPABLE_MODELS` set at module scope, ensuring each model gets a compatible tool version.

This is a non-trivial finding: any cross-vendor evaluation that uses one tool version for all Anthropic models will either fail outright on Haiku or undersell Sonnet's retrieval capability.

3. Methodology

3.1 Dataset

The evaluation corpus comprises 157 production fact-checks from the Perclaim corpus, selected via the constraint `created_at >= 2026-05-01`. These checks were originally produced by GPT-5.4-mini in production. The corpus includes a mix of:

- Text-only posts (60 rows, 39%)
- Posts with one or more images (97 rows, 61%)
- Image-only posts (posts where the textual content is empty and the visual content carries the entire claim)
- Posts with hyperlinks, attributed quotes, video embeds, and mixed media

Topic distribution skews political — Trump administration policy, Iran war coverage, Ukraine ceasefire negotiations, immigration policy, election claims — reflecting the natural distribution of factually-disputed social-media content during the evaluation window. This skew is a limitation worth surfacing: the corpus is not representative of all social-media content, only of content users submit for fact-checking, which is itself selected for being claim-like and likely contested.

3.2 Replay protocol

For each model, the eval harness (`scripts/eval/replay.ts` in the perclaim-web repository) replayed each row's input — the claim text, image fetch status, and source URL — against the candidate model using the same system prompt and tool configuration as production. The model's verdict, summary, suggested reply, claims breakdown, and source citations were captured in a structured JSON result file.

Anthropic provider implementation (`scripts/eval/providers.ts`) handles per-model tool version selection, retry logic on transient failures, and structured-output parsing. Ollama provider implementation handles local-model inference via the OpenAI-compatible API surface exposed by Ollama; web search tools are not available in this configuration.

3.3 The Sonnet-as-proxy-baseline framing

The traditional approach to model evaluation uses ground-truth labels from human adjudicators. For this study, we use Claude Sonnet 4.6 as proxy ground truth instead, on the following reasoning:

1. Sonnet 4.6 has the strongest retrieval (`web_search_20260209` with dynamic filtering) of any tested model.
2. Manual adjudication of substantive disagreements between Sonnet and GPT-5.4-mini ($n = 9$) found Sonnet correct on 6 cases, GPT correct on 0 cases, and 3 cases legitimate-either-way. Sonnet did not lose any adjudicated case.
3. Producing genuine human-adjudicated ground-truth labels for 157 politically-charged posts is expensive and inter-rater agreement on contested political claims is itself noisy.

This is a limitation, not ground truth. Using Sonnet as the reference biases results toward "models that calibrate similarly to Sonnet rank well." We surface this caveat throughout and present absolute concordance numbers rather than accuracy claims.

3.4 Disagreement decomposition

Verdict differences between any candidate and Sonnet were categorized by spectrum distance:

- **Strict agreement**: identical verdict
- **Adjacent (1-notch shift)**: TRUE ↔ MOSTLY_TRUE, MIXED ↔ MOSTLY_TRUE, etc.
- **Same-severity relabel**: MISLEADING ↔ MOSTLY_FALSE (both = severity 2)
- **Two-step**: e.g., MISLEADING → MOSTLY_TRUE
- **Big-swing (3+ notches)**: e.g., FALSE → MOSTLY_TRUE
- **Orthogonal**: any pair where one side is OPINION/UNVERIFIABLE/CRITIQUE — these are category swaps, not factual disagreements

The "substantive disagreement" category combines two-step, big-swing, and same-severity relabel — the disagreement types that represent genuinely different factual readings rather than calibration noise.

3.5 Adjudication

For each candidate, we manually reviewed all substantive disagreements with Sonnet, examining:

- The claim text (or image, where applicable)
- Sonnet's verdict, summary, and suggested reply
- The candidate's verdict, summary, and suggested reply
- Web-search verification of the underlying factual claim where checkable

Adjudication was performed by the same author working from the result data — the claude.ai chat assistant available during the analysis sessions (May 13–14 for the original study, May 19 for the follow-up), an instance from the Claude 4.6 model family. The adjudicator role is deliberately distinct from the proxy-baseline role: the corpus-replay baseline (§3.3) is the programmatic output of Sonnet 4.6 via the Anthropic API, while the adjudicator is a separate claude.ai chat assistant instance reviewing disagreements with web-search verification. This is acknowledged as a methodological weakness — the adjudicator and the baseline share vendor architecture and may share biases. Independent human adjudication by professional fact-checkers is listed as future work (§8.5).

4. Results

4.1 Verdict distribution per model

The raw verdict distributions reveal substantially different calibration regimes:

Verdict	Sonnet	Haiku	GPT	gpt-oss	gemma4
TRUE	3	11	10	4	12
MOSTLY_TRUE	31	52	16	4	4
MIXED	71	41	65	21	9
MISLEADING	11	5	12	1	8
MOSTLY_FALSE	1	6	6	4	1
FALSE	5	10	6	41	52
OPINION	12	7	18	16	23
UNVERIFIABLE	22	11	19	52	17
(null/failed)	1	14	0	9	26
Total	157	157	152	152	152

Sonnet and GPT-5.4-mini concentrate verdicts in MIXED (45% and 43% respectively), consistent with retrieval-equipped models that surface nuance in real-world claims. Haiku skews toward MOSTLY_TRUE (33%), suggesting slightly more permissive calibration. The offline candidates show a striking inversion: gpt-oss puts 27% of verdicts in FALSE and 34% in UNVERIFIABLE; gemma4 puts 34% in FALSE. The frontier models combined put only 3–7% of verdicts in FALSE.

4.2 Headline concordance matrix

Concordance with Sonnet, computed on the subset of rows where both models produced a verdict:

Candidate	Paired rows	Strict concord.	+ Adjacent (1-notch)	+ Orthogonal	Substantive disagree.
Haiku 4.5	143	56.6%	89.5%	96.5%	5 (3.5%)
GPT-5.4-mini	151	52.3%	79.5%	92.7%	11 (7.3%)
gpt-oss:120b	142	31.0%	44.4%	76.8%	33 (23.2%)
gemma4:26b	125	29.6%	47.2%	63.2%	46 (36.8%)

The cumulative-collapse columns reveal that strict concordance numbers dramatically understate practical agreement for the frontier pair (Haiku and GPT) while not changing the picture significantly for the offline pair. When 1-notch calibration shifts and orthogonal category swaps are folded in as "not-actually-factually-different," Haiku reaches 96.5% effective concordance with Sonnet, GPT reaches 92.7%, but the offline candidates remain below 80%.

4.3 Image vs no-image concordance

All four candidates do worse on image-bearing rows than text-only rows, but the gap is much wider for offline models:

Candidate	Image-row concordance	No-image concordance	Gap
GPT-5.4-mini	51.1%	54.1%	3.0pp
Haiku 4.5	52.6%	64.6%	12.0pp
gpt-oss:120b	21.4%	44.8%	23.4pp
gemma4:26b	22.1%	38.6%	16.5pp

Image rows are harder for all candidates — the claim is often image-embedded text (a quote-card, a screenshot of a Truth Social post, etc.) that requires both vision capability and retrieval to verify. The offline candidates' image-row collapse (∼22%) is severe and represents a compounding limitation: limited vision + no retrieval = systematic failure on image-only posts.

4.4 Disagreement type decomposition

The same concordance numbers, decomposed by what kind of disagreement they contain:

Candidate	Total disag.	Adjacent	Orthog.	Two-step	Big-swing	Same-relabel
Haiku 4.5	62	75.8%	16.1%	4.8%	1.6%	1.6%
GPT-5.4-mini	72	56.9%	27.8%	9.7%	2.8%	2.8%
gpt-oss:120b	98	19.4%	46.9%	24.5%	9.2%	0%
gemma4:26b	88	25.0%	22.7%	42.0%	10.2%	0%

The fingerprints are categorically different. Frontier-model disagreements are 73–84% calibration-or-orthogonal (adjacent + orthogonal). Offline-model disagreements are 30–52% substantive (two-step + big-swing + same-relabel). These are not different points on the same accuracy axis — they are different failure regimes.

4.5 Manual adjudication: Sonnet vs GPT-5.4-mini ($n = 11$ substantive disagreements)

Manual review with web-search verification yielded the following:

Pattern	Count	Description
Sonnet correct, GPT wrong	6	E.g., Sonnet correctly identifies a real Dan Rather quote that GPT calls fabricated; Sonnet identifies a real Lead Stories fact-check that GPT misses entirely
Both legitimate, different calibration	5	Same factual conclusions, different verdict labels reflecting different calibration thresholds
GPT correct, Sonnet wrong	0	—

The wins for Sonnet show a consistent pattern: more specific dates, more specific fact-checker citations (Lead Stories, PolitiFact, Snopes, EDMO, NewsGuard), and more accurate identification of the actual claim being made — particularly in image-only posts where pattern-matching on partial cues led GPT to misread the topic.

4.6 Manual adjudication: Sonnet vs Haiku ($n = 5$ substantive disagreements)

All five Haiku-vs-Sonnet substantive disagreements turned out to be calibration choices, not factual errors:

short_id	Sonnet / Haiku	Pattern
ck_g8buojxi	Misleading / Mostly_True	Same facts (Trump's "soccer might be better" quote real). Sonnet penalizes meme's overstatement; Haiku anchors on quote authenticity.
ck_rr0evv7a	Misleading / True	Same facts (Trump's "not motivated by Americans' finances" quote real). Sonnet penalizes meme stripping Iran context; Haiku anchors on quote authenticity.
ck_sfkjxj22	Misleading / Mostly_False	Same factual conclusions ($1.5T = full CR cost, federal law bars undocumented Medicaid). Same-severity relabel.
ck_x6nhrwf4	Misleading / Mostly_True	Same factual conclusions (Bessent quote real; "wage increase" framing misleading). Calibration call.
ck_z38wjbkp	Mixed / False	Same factual breakdown of multi-claim meme. Sonnet credits kernels of truth; Haiku rules on framing.

Haiku produces **zero factual errors** against Sonnet on these substantive disagreements. All five are calibration calls about whether to anchor on "real underlying fact" or "distorted framing." Both anchorings are defensible.

4.7 Offline candidate failure pattern: systematic false negatives

The dominant disagreement pair for both offline candidates is Mixed → False: gpt-oss has 22 cases where Sonnet says Mixed and gpt-oss says False; gemma4 has 31. Sampling these cases consistently revealed the same pattern: the offline candidate denies a real event because it can't verify the event via web search, and defaults to False.

Three illustrative cases:

Case 1: Ramaswamy and the Ohio gubernatorial primary (`ck_1g5xbpn5`, Sonnet MIXED / gpt-oss FALSE). Sonnet correctly identifies that Vivek Ramaswamy won the 2026 Ohio GOP gubernatorial primary, with Amy Acton as the Democratic nominee. gpt-oss says: "Ramaswamy is a presidential candidate with no primary win, he is not running for Ohio governor, and Amy Acton is not on the ballot." gpt-oss is using outdated training-data knowledge and is factually wrong.

Case 2: FireAid $100M concert (`ck_xwe3y9uu`, Sonnet MIXED / gemma4 FALSE). Sonnet confirms the FireAid concert raised approximately $100M and that Rep. Kevin Kiley's allegations were investigated, while flagging that "laundered" is legally inaccurate language. gemma4 says: "There is no evidence that a 'Pacific Palisades Fire Aid' concert raised $100 million. . . No such large-scale fund or corresponding official statements exist in the public record." The FireAid concert is a widely-covered real event. gemma4 is factually wrong.

Case 3: Virginia Democratic redistricting spending (`ck_cnq2kxfu`, Sonnet MIXED / gemma4 FALSE). Sonnet confirms the approximately $62M Democratic spending figure from VPAP data, slightly below the meme's $70M claim. gemma4 says: "No official financial disclosures or reports from credible news organizations corroborate such an enormous expenditure." VPAP data exists and is public. gemma4 is factually wrong.

Quantifying the pattern: 95.1% of gpt-oss's FALSE verdicts (39 of 41) and 92.3% of gemma4's FALSE verdicts (48 of 52) land on rows where Sonnet's verdict is something other than FALSE. The offline candidates are not exhibiting more conservative skepticism — they are systematically denying real events they have no way to learn about.

4.8 Reliability profile

Failure rates per candidate, including both hard errors (no replayed block produced) and JSON parse failures (block present but no verdict extractable):

Candidate	Rows	Hard errors	JSON parse failures	Total failure rate
Sonnet 4.6	157	0	1	0.6%
GPT-5.4-mini	152	0	0	0.0%
Haiku 4.5	157	0	14	8.9%
gpt-oss:120b	152	8	1	5.9%
gemma4:26b	152	4	22	17.1%

Haiku's failure profile is striking: zero hard errors, but 14 rows (8.9%) where the model produced output that failed structured-output parsing. This pattern suggests the failure mode is malformed JSON (truncation, escape errors, syntax mistakes) rather than capability gaps. Schema enforcement, prefill scaffolding, or validate-and-retry are plausible mitigations that could push this rate below 1%.[1]

4.9 Performance and cost

Wall-clock per row (mean / median / p95):

Candidate	Median	Mean	p95	Total wall (157 rows)
GPT-5.4-mini	12.7s	13.7s	25.1s	34.6 min
Haiku 4.5	21.5s	21.4s	38.0s	56.0 min
gpt-oss:120b	43.1s	44.8s	82.6s	113.5 min
gemma4:26b	68.6s	71.1s	110.6s	180.2 min
Sonnet 4.6	115.2s	117.8s	253.0s	308.2 min

[1] This hypothesis was confirmed by the follow-up work documented in §6.

Cost estimates (in USD per check; eval-derived with `web_search_20260209` uncapped on Anthropic models):

Candidate	Per-check cost	Notes
GPT-5.4-mini	$0.047	Production-calibrated over 123 checks, May 7–10
Haiku 4.5	~$0.06–0.10	Eval-extrapolated; observed $0.0905 in May 19 replay (§6)
Sonnet 4.6	~$0.30–0.55	Eval-extrapolated; subsequent analysis (§6) revised the capped-deployment estimate upward to ~$0.50
gpt-oss:120b	~$0.005	Power-only cost on owned hardware (Dell Pro Max GB10, ~50W)
gemma4:26b	~$0.005	Same hardware

5. Discussion

5.1 The retrieval cliff is categorical, not gradational

The most consequential finding is that the divide between retrieval-equipped and retrieval-free models is not a smooth degradation along an accuracy axis but a categorical shift in failure mode. Retrieval-equipped models (Sonnet, Haiku, GPT) disagree with each other primarily on calibration — whether to anchor on the underlying fact's authenticity or on the framing's distortion. Retrieval-free models (gpt-oss, gemma4) disagree on whether real events occurred at all.

This has direct implications for any product that fact-checks current events. Without retrieval, an offline model cannot distinguish "claim about real event that I don't have training-data knowledge of" from "claim about fabricated event." Its default — observed empirically at 92–95% rates for the offline candidates' FALSE verdicts — is to deny.

For Perclaim, this means local-model viability is gated on app-side retrieval (Tavily, Brave, or Google CSE) being shipped first, and then a re-evaluation with retrieval-equipped offline models. The current architecture, which delegates retrieval entirely to the model via OpenAI's `web_search` tool, has no plausible local-substitution path.

5.2 Concordance numbers understate frontier-model agreement

The headline 56.6% Haiku-vs-Sonnet concordance figure, taken at face value, suggests Haiku is substantially less accurate than Sonnet. Manual adjudication contradicts this: Haiku makes zero factual errors against Sonnet on substantive disagreements. The 43.4% disagreement is overwhelmingly 1-notch calibration noise.

A more useful summary statistic for frontier-model evaluation is "effective concordance after collapsing 1-notch shifts and orthogonal category swaps." On this measure, Haiku reaches 96.5%, GPT reaches 92.7%, and the offline candidates remain below 80%. This measure is closer to "do these models reach the same factual conclusion" than strict concordance is.

5.3 Haiku 4.5's reliability problem is engineering, not capability

Haiku's 8.9% JSON parse failure rate is the single largest barrier to its production use, and it sits inside a model with otherwise-excellent accuracy. The pattern is consistent with structured-output reliability issues observed in earlier Anthropic models — the model produces semantically correct content but mistypes JSON delimiters, truncates, or generates non-spec characters.

Several engineering mitigations are likely effective:

- **Prefill scaffolding**: prefill the response with the opening `{` to anchor the model's output mode.
- **Schema enforcement at the API level**: Anthropic's structured output features (where available for the model and tool combination) provide grammar-constrained decoding.
- **Validate-and-retry**: detect JSON parse failure at the response layer and re-invoke with explicit "respond in valid JSON only" instructions.
- **Output post-processing**: apply lenient JSON parsers (e.g., `partial-json`, `json5`) that can recover from common minor errors.

A focused engineering investigation could plausibly push Haiku's failure rate below 1%, at which point Haiku-with-retrieval becomes a strong candidate for either free-tier replacement or a cheaper Pro-tier option.[2]

5.4 Implications for tier strategy

The recommendations in this subsection reflect the analysis as of May 14, 2026. §6.4 presents the revised deployment recommendation following the May 19 Haiku follow-up work.

Combining accuracy, cost, and reliability findings, three deployment positions are clearly defensible:

Free tier: GPT-5.4-mini stays. Cost-calibrated at $0.047/check, 92.7% effective concordance with Sonnet, zero reliability issues, never factually wrong against Sonnet on adjudicated substantive cases. The current production model is the right free-tier choice on the current evidence.

Pro tier: Sonnet 4.6. When Stripe ships, Pro-tier accuracy lift is real — Sonnet wins 6–0 against GPT-5.4-mini on substantive disagreements requiring retrieval-derived knowledge. At a capped-`max_uses` production cost of approximately $0.30/check and a 500/month fair-use cap, Pro-tier economics work within the $15/mo target for typical usage. The premium positioning ("Pro uses the same model that won 6–0 against our standard model in head-to-head evaluation") is empirically defensible.

Haiku 4.5: deferred candidate. Pending JSON reliability investigation, Haiku is a strong contender for either a free-tier upgrade or a cheaper Pro-tier option. The accuracy data already supports it; the engineering work to stabilize structured output is the remaining gate.

Local candidates (gpt-oss, gemma4): not viable without app-side retrieval. Until v2 retrieval architecture ships, these models cannot serve current-events fact-checking without producing systematic false negatives.

6. Follow-up: Closing the Haiku reliability gap

This section, added May 19, 2026, documents engineering work performed after the original study to address Haiku 4.5's JSON parse failure rate. The findings here supersede the deferred-candidate framing in §5.4 and inform the revised deployment recommendation in §6.4.

6.1 Motivation

The original study (§5.3) hypothesized that Haiku's 8.9% structured-output failure rate was an engineering problem rather than a capability gap and listed plausible mitigations. Two facts motivated treating this as a high-priority follow-up:

[2] §6 documents the follow-up work that empirically resolved this question, achieving a 1.6% null-verdict rate via parser-side cascade recovery.

1. Haiku's accuracy profile was otherwise indistinguishable from Sonnet's on adjudicated substantive disagreements (zero factual errors out of five cases), at roughly one-fifth the per-check cost.
2. No other tested model occupied this position: high accuracy, low cost, blocked only on output reliability.

If the structured-output failure rate could be pushed below 2%, Haiku would become a viable candidate for the production stack — either as primary or as a same-vendor-class backup for the current production model.

6.2 Failure mode characterization

We re-ran the corpus three times to characterize the failure mode before committing to a fix. Two observations dominated:

The failures are not deterministic. The same row replayed on different runs produced different failures. Of the 11 nulls from the post-fix run (§6.3), 7 recovered directly via the LLM in the next run with no recovery intervention; same model, same inputs, different outputs. Per-row recovery is noise; aggregate failure rate is signal.

The dominant failure shape is prose refusal, not malformed JSON. Of the rows that didn't parse cleanly, the majority were the model declining to fact-check in conversational English ("I can't verify this without more context") rather than producing structurally malformed JSON. A smaller subset involved trailing commas, unescaped quotes inside summary strings, or other JSON-spec violations where the verdict was clearly present but unparseable.

The two failure shapes call for different recoveries. Malformed JSON needs structural extraction (find the verdict field via regex even when the surrounding JSON is broken). Prose refusal needs synthesis (recognize that the model declined and produce a corresponding UNVERIFIABLE verdict with the refusal preserved as evidence).

6.3 Cascade recovery design

We implemented a parser-side recovery cascade as an evaluation-only helper (`scripts/eval/recovery.ts`), wired into the eval providers but deliberately not into the production inference path. Production already has a cross-vendor fallback mechanism that activates on `parsed.error`, and synthesizing a verdict at the parser layer would suppress that fallback. The recovery helper is therefore eval-scoped, with an active anti-tidiness comment documenting why moving it to `lib/` would degrade production behavior.

The cascade fires when the primary JSON parser returns an error and applies two recovery paths in fixed order:

1. **Regex recovery layer.** A structural anchor matching the verdict field — the pattern

   ```
   /"overall_verdict"\s*:\s*"(TRUE|MOSTLY_TRUE|MIXED|MISLEADING|
           MOSTLY_FALSE|FALSE|UNVERIFIABLE|OPINION)"/
   ```

   extracts the verdict from the raw response text along with a best-effort summary string. Returns a structured result tagged with `_synthesized_from: 'regex_recovery'`.

2. **Prose-refusal synthesis fallback.** If the regex finds nothing, the cascade assumes the model produced a prose refusal and synthesizes a clean UNVERIFIABLE verdict. The summary is set to "Model declined to fact-check this submission." The raw response text (truncated to 500 characters) is preserved as `_raw_excerpt` for audit. The originating parse error is preserved as `_recovery_trigger`. Tagged with `_synthesized_from:`

`'prose_refusal'`.

The fixed-order constraint is load-bearing. Regex extraction must run first because a structurally-malformed JSON response may contain a real verdict; reordering would synthesize UNVERIFIABLE over a recoverable result. The cascade's top-of-file comment documents this constraint and includes an active anti-tidiness directive.

6.4 Results: trajectory across three replay runs

We measured the null-verdict rate across three replays of the post-2026-05-14 production corpus (188 canonical rows after filtering cache-reference tracking rows that lack verdicts by construction):

Run	Date / Scope	Wall	Null rate (canonical)
Original (pre-fix)	2026-05-14, 157 rows	56.0 min	14/157 = **8.9%**
Post-fix replay	2026-05-19T14:04, 200 rows	∼80 min	11/188 = **5.85%**
Post-cascade replay	2026-05-19T20:54, 200 rows	∼77 min	3/188 = **1.6%**

The 8.9% → 5.85% transition reflects two interventions applied between the original study and the May 19 replay: a prompt-side rule explicitly directing the model to emit UNVERIFIABLE as JSON rather than refusing in prose, and a replay-harness fix that was occluding some valid responses. The 5.85% → 1.6% transition is attributable to the cascade-recovery work documented in §6.3. Total reduction from baseline: **82%**.

Cascade fire breakdown. On the cascade-validation run, six of the canonical rows fired the prose-refusal synthesis path and landed as UNVERIFIABLE with the model's prose refusal preserved as evidence. The regex-recovery path fired zero times — the two rows from the post-fix run that had malformed-but-extractable JSON came back with cleanly-formed JSON the next time around, an instance of the stochastic-variance observation in §6.2. The regex path's correctness is verified by unit tests in `test/eval-recovery.test.ts` using the actual malformed-fixture data from the post-fix run.

Remaining failure mode. Three canonical rows in the cascade-validation run produced null verdicts and did not trigger the cascade. These are cases where the lenient JSON parser succeeded structurally but the resulting object had no `overall_verdict` field — a partial-parse mode that returns no `.error` signal and therefore bypasses the cascade. The shape is distinct from both regex-recoverable malformation and prose refusal. Closing this last 1.6% would require extending the cascade to fire on `(!parsed.error && !parsed.overall_verdict)` as well, which is queued as a small follow-up but not a deployment blocker.

6.5 Cost data update

Per-row cost was directly observed during the cascade-validation run: $18.10 over 200 rows = $0.0905/check. This is within the eval-extrapolated range from §4.9 ($0.06–0.10) and confirms Haiku as the lowest-cost cloud frontier option in the tested set, at roughly one-fifth the cost of Sonnet.

Sonnet 4.6's cost estimate was not directly re-measured during the follow-up, but cost modeling for the deployment matrix below uses ∼$0.50/check as the planning figure, slightly above the §4.9 capped-deployment estimate. This is the more conservative number and the one that drives the Pro-tier economics analysis in §6.6.

6.6 Revised deployment recommendation

The deployment recommendation in §5.4 is superseded by the recommendation below.

The updated production stack uses GPT-5.4-mini as primary inference across both free and Pro tiers, with Haiku 4.5 as the cross-vendor backup activated on transient errors from the primary. Sonnet 4.6 is removed from the production stack and retained only as the cross-vendor evaluation baseline.

Tier	Primary	Backup
Free (25/mo cap)	GPT-5.4-mini	Haiku 4.5
Pro (250/mo cap)	GPT-5.4-mini	Haiku 4.5
BYOK	User's API key	—

Why Sonnet was removed from the production stack. The May 14 recommendation placed Sonnet 4.6 behind the Pro tier as the accuracy upgrade. The economics did not survive contact with the pricing environment surveyed in the May 16 competitive scan: the general-purpose AI subscriptions (Perplexity Pro, ChatGPT Plus, Claude Pro, Google AI Pro) all converge on $20/month, and a niche fact-checking tool cannot price above that ceiling. With a Pro-tier price target of $9.99/month and a 250-check cap, Sonnet at ∼$0.50/check would produce an inference-cost ceiling of $125/month against $9.99 in revenue. Even with cache hits and sub-cap usage pulling average cost down, the cap-hitter risk is unsustainable.

Why Haiku is the backup and not the primary. The 1.6% null-verdict rate is low enough to be acceptable as a fallback (rare activation, transparent recovery via the production cross-vendor mechanism) but not as primary. GPT-5.4-mini's 0% failure rate over 152 rows in the original study and its cost-calibrated $0.047/check make it the safer primary. Same-vendor diversification (using two Anthropic models) would not have provided the resilience benefit of cross-vendor backup.

Pro-tier accuracy positioning. The Pro tier no longer promises a model-quality upgrade. Pro and free differ on usage cap (250 vs. 25 checks/month) and surrounding features (history, polish). This is a strategic call about market positioning more than a technical one — the per-call cost economics ruled out model-differentiated tiers at the $9.99 price point.

7. Limitations

7.1 Sample size

The corpus is 157 rows for the original study and 188 rows for the Haiku follow-up replay (after filtering cache-reference rows). The substantive-disagreement subset adjudicated per candidate ranged from 5 (Haiku vs Sonnet) to 46 (gemma4 vs Sonnet). For the smaller subsets, conclusions about "zero factual errors" should be read as "zero factual errors observed in the sampled cases" rather than as a population-level claim.

7.2 Sonnet as proxy ground truth

Sonnet 4.6 is the most accurate available model among those tested, by virtue of having the best retrieval. But it is not infallible. Using it as the reference biases the matrix toward "models that calibrate like Sonnet rank high." A model with different but legitimate calibration would score worse on this matrix without being less accurate in absolute terms.

Mitigation: we report decomposition (calibration vs orthogonal vs substantive) alongside strict concordance, so calibration-style differences are visible separately from substantive disagreements.

7.3 Single-replay variance

Each row was replayed once per model in the original study and three times for the Haiku follow-up. LLM outputs vary across replays, especially for models with non-zero temperature or sampling-based tool use (web search query selection varies). Some fraction of the observed disagreements are within-model run-to-run variance rather than across-model differences.

The Haiku follow-up made this variance directly observable: seven of the eleven nulls from the post-fix run recovered to clean structured output in the next run with no intervention. The cascade-recovery system is validated on the failures that recur in aggregate; per-row stability is not claimed.

7.4 Adjudicator bias

Manual adjudication was performed by an LLM-based assistant (the claude.ai assistant model) reviewing the result data and performing web-search verification. This assistant shares architectural similarity with Sonnet and Haiku and may share biases. Cases where all Anthropic models converge on a verdict that is wrong but plausible-sounding would be hard for an Anthropic-family adjudicator to detect.

Independent human adjudication by experienced fact-checkers would address this. A subset of 20–30 adjudicated cases reviewed by professional fact-checkers (e.g., at PolitiFact or Snopes) would provide a useful calibration check.

7.5 Cost estimates are eval-derived

The cost numbers reported in §4.9 are extrapolated from a single eval run, not from production traffic. The Sonnet eval ran with uncapped `max_uses` on `web_search_20260209`, which inflated input-token usage well beyond what a capped-retrieval production deployment would incur. The Haiku follow-up directly observed $0.0905/check across 200 rows, which is the more reliable number for that model. Production-traffic cost measurement remains future work for all models in the deployment stack.

7.6 Corpus topical skew

The 157-row corpus is heavily political — Trump administration policy, Iran war coverage, election claims. This reflects user-submitted fact-check requests during the eval window. Models may calibrate differently on, for example, scientific claims, health misinformation, financial scams, or local news. Generalization beyond political fact-checking is not supported by this evaluation.

8. Future work

8.1 Haiku JSON reliability investigation (resolved)

This item is closed by the work documented in §6. A focused audit of the original 14 Haiku JSON-parse failures yielded a parser-side cascade-recovery system that reduced the null-verdict rate from 8.9% to 1.6% across a 188-row post-rollout corpus. Haiku 4.5 is now production-viable.

8.2 Partial-parse cascade extension

Three rows in the cascade-validation run remain null after the cascade because they produced structurally-valid JSON with no `overall_verdict` field — a different failure mode than the cascade was designed for. Extending the cascade to fire on `(!parsed.error && !parsed.overall_verdict)`

would close these remaining cases. Estimated work: small follow-up dispatch + 200-row re-replay.

8.3 Multi-replay variance characterization

Run each model 3–5 times against a fixed corpus. Compute within-model verdict variance per row and report alongside cross-model disagreement. This isolates "calibration differences" from "sampling noise" cleanly. The Haiku follow-up demonstrates this is non-trivial: seven of eleven nulls in one run recovered with no intervention in the next.

8.4 App-side retrieval integration

Ship app-side retrieval (Tavily, Brave, or Google CSE) as a v2 architecture, then re-evaluate offline candidates with retrieved context. The hypothesis is that retrieval-equipped offline candidates will close most of the gap to retrieval-equipped frontier models, making local-model deployment viable for cost-sensitive deployments.

8.5 Professional fact-checker adjudication

Recruit 2–3 professional fact-checkers to independently adjudicate a 30-row sample of substantive disagreements. Use this to calibrate the LLM-based adjudication used in this study and quantify any systematic bias.

8.6 Larger corpus

Extend the corpus to 500–1000 rows with explicit topical stratification (politics, science, health, finance, local news). Re-run the matrix with this expanded corpus.

8.7 Newer Anthropic models

Claude Opus 4.7 was not evaluated in this study but may offer additional accuracy lift over Sonnet 4.6 at higher cost. A targeted Opus-vs-Sonnet study on 50 hard cases would establish whether Opus is a meaningful upgrade for the cross-vendor evaluation baseline.

8.8 Kimi K2.x pilot

A Kimi K2.x evaluation was scoped but deferred during the original study. A 25-row pilot (∼$0.88 estimated cost) is the recommended next step before committing to a full 200-row evaluation.

9. Conclusion

We evaluated five language models for social-media fact-checking on a corpus of 157 production fact-checks, using Claude Sonnet 4.6 as proxy ground truth. A follow-up study (§6) re-evaluated Haiku 4.5 after engineering work on its structured-output reliability. The main findings are:

1. Concordance numbers significantly understate practical agreement between retrieval-equipped frontier models. Haiku 4.5 reaches 96.5% effective concordance with Sonnet 4.6 once 1-notch calibration shifts and orthogonal category swaps are folded in. GPT-5.4-mini reaches 92.7%.
2. Haiku 4.5 makes zero factual errors against Sonnet on substantive disagreements. Its 8.9% JSON parse failure rate, identified in the original study as the only barrier to production

use, was reduced to 1.6% via parser-side cascade recovery (§6). Haiku is now production-viable.

3. Offline candidates without web search exhibit systematic false-negative bias. 92–95% of their FALSE verdicts land on rows containing real, verifiable events that they cannot learn about due to retrieval absence. This is a categorical failure mode, not a calibration difference.
4. The deployment strategy revised in §6.4 is two-model: GPT-5.4-mini as primary inference across both free and Pro tiers, with Haiku 4.5 as cross-vendor backup. Sonnet 4.6 is removed from production due to per-check cost economics incompatible with the Pro-tier price point; it remains the reference model for evaluation work. Local-model viability is gated on app-side retrieval, which is a v2 architectural prerequisite.

Appendix A: Eval harness artifacts

The eval harness lives in `scripts/eval/` in the perclaim-web repository. Key files:

- `scripts/eval/replay.ts` — main replay loop
- `scripts/eval/providers.ts` — provider implementations for OpenAI, Anthropic, Ollama
- `scripts/eval/diff.ts` — pairwise diff and concordance computation
- `scripts/eval/recovery.ts` — cascade recovery helper (added May 19, 2026; eval-only by design — see top-of-file documentation)
- `scripts/eval/README.md` — usage documentation

Relevant commits during this evaluation:

- `82c81e3` `feat(eval): Anthropic provider with hosted web_search`
- `cc517fc` `docs(backlog): Anthropic provider integration`
- `af7fe97` `fix(eval): AnthropicProvider per-model web_search tool version`
- `029bf4b` `docs(backlog): per-model tool version fix`
- `1764faf` `fix(eval): replace image-context-equal metric tautology`
- `900019f` `docs(backlog): metric fix`
- `8d7c456` `fix(eval): cascade recovery for parse failures in replay providers` (May 19)
- `899002a` `docs(backlog): track eval parse-failure recovery cascade` (May 19)

Appendix B: Result file inventory

Result JSON files used in this analysis:

- `baseline-openai-2026-05-13.json` (152 rows, GPT-5.4-mini)
- `replay-anthropic-claude-sonnet-4-6-2026-05-14T01-19-36-901Z.json` (157 rows)
- `replay-anthropic-claude-haiku-4-5-20251001-2026-05-14T00-22-48-132Z.json` (157 rows)
- `candidate-ollama-gpt-oss-120b-2026-05-13.json` (152 rows)

- `candidate-ollama-gemma4-26b-2026-05-13.json` (152 rows)
- `replay-anthropic-claude-haiku-4-5-20251001-2026-05-19T13-25-07-811Z.json` (25-row pilot, May 19 follow-up)
- `replay-anthropic-claude-haiku-4-5-20251001-2026-05-19T14-04-08-314Z.json` (200 rows, post-fix)
- `replay-anthropic-claude-haiku-4-5-20251001-2026-05-19T20-54-29-896Z.json` (200 rows, post-cascade)

Appendix C: Selected adjudicated cases

The full adjudication record for the GPT-vs-Sonnet $n = 11$ substantive disagreements is available in session transcripts. Three representative cases are reproduced here.

Case 1: McConaughey fabricated quote (`ck_pybbzmm8`)

Claim: image-only post on Facebook attributing a quote to Matthew McConaughey: "If Kamala Harris had become president, Trump would be in prison, the Straits of Hormuz would be open, Ukraine would be victorious, gas would be under $3 a gallon. . ."

GPT-5.4-mini verdict: MIXED. Summary: "The post mixes one real fact with several speculative or unsupported predictions. Trump was convicted on 34 felony counts in 2024, but claims about prison, the Strait of Hormuz, Ukraine, gas under $3, and restored global respect are not established as factual outcomes."

Sonnet 4.6 verdict: FALSE. Summary: "The quote attributed to Matthew McConaughey is fabricated. His representatives confirmed to the fact-checking outlet Lead Stories that he never made this statement, and the image itself has been flagged as likely AI-generated."

Adjudication: Sonnet correct. Lead Stories published a fact-check on May 11, 2026 confirming the quote is fabricated and the image is AI-generated; GPT misread the claim entirely as a Trump-conviction discussion and missed the McConaughey attribution.

Case 2: Dan Rather quote (`ck_iwhcc5me`)

Claim: image-only post attributing a quote to Dan Rather rejecting Trump as "a cheating, lying, racist, treasonous, fascist, and vile man."

GPT-5.4-mini verdict: FALSE. Summary: "The image presents a quote attributed to Dan Rather, but I couldn't find a public source for this wording on his site or in major coverage."

Sonnet 4.6 verdict: MOSTLY_TRUE. Summary: "Dan Rather did post this statement on his verified X (Twitter) account in April 2024, and the quote in the meme matches his actual words very closely."

Adjudication: Sonnet correct. The Dan Rather X post dates to April 3–4, 2024 and was covered by Newsweek, The Wrap, The Daily Beast, and Blaze Media. The quote in the meme matches Rather's verified X-account post.

Case 3: Ramaswamy / Ohio (`ck_1g5xbpn5`)

Claim: text post discussing Vivek Ramaswamy's Ohio GOP primary win and a forthcoming general election against Democratic nominee Amy Acton.

Sonnet 4.6 verdict: MIXED. Summary: "The core factual claims — that Ramaswamy won the Ohio GOP gubernatorial primary and that Amy Acton is the Democratic nominee — are

accurate and confirmed by multiple news outlets."

gpt-oss:120b verdict: FALSE. Summary: "The post claims Vivek Ramaswamy won an Ohio Republican primary, is seeking the governor's mansion, and that Amy Acton is the Democratic candidate. None of these statements are true."

Adjudication: Sonnet correct. gpt-oss is applying outdated training-data knowledge (Ramaswamy as a 2024 presidential candidate) and incorrectly denying his 2026 Ohio gubernatorial primary win, which is documented in multiple sources Sonnet's retrieval surfaced.

This evaluation was conducted as part of Perclaim's pre-launch model selection process. All result files and analysis scripts are available in the perclaim-web repository under `scripts/eval/`. The May 19 follow-up (§6) extended the original study with parser-side cascade recovery work that closed the Haiku reliability gap identified in §5.3 of the original.