

Large Language Models Require Curated Context for Reliable Political Fact-Checking—Even with Reasoning and Web Search

Matthew R. DeVerna
Stanford University

Kai-Cheng Yang
Binghamton University

Harry Yaojun Yan
Texas A&M University

Filippo Menczer
Indiana University

Abstract

Large language models (LLMs) have raised hopes for automated end-to-end fact-checking, but prior studies report mixed results. As mainstream chatbots increasingly ship with reasoning capabilities and web search tools—and millions of users already rely on them for verification—rigorous evaluation is urgent. We evaluate 15 recent LLMs from OpenAI, Google, Meta, and DeepSeek on more than 6,000 claims fact-checked by PolitiFact, comparing standard models with reasoning- and web-search variants. Standard models perform poorly, reasoning offers minimal benefits, and web search provides only moderate gains, despite fact-checks being available on the web. In contrast, a curated RAG system using PolitiFact summaries improved macro F1 by 233% on average across model variants. These findings suggest that giving models access to curated high-quality context is a promising path for automated fact-checking.

1 Introduction

Misinformation is a persistent and consequential challenge in the digital age, distorting public opinion and eroding trust in democratic institutions (Ecker et al., 2024; World Economic Forum, 2024). The rapid rise of large language models (LLMs) has intensified these concerns by making it easy to generate convincing but false content at scale (Menczer et al., 2023). At the same time, LLMs have been lauded for their potential to counter misinformation, with growing interest in their use for information seeking (Chatterji et al., 2025) and fact-checking (Augenstein et al., 2024; Chen and Shu, 2024).

Early computational approaches to fact verification typically decomposed the task into sub-tasks such as claim detection, evidence retrieval, and veracity classification. Datasets and tools like FEVER (Thorne et al., 2018), LIAR (Wang,

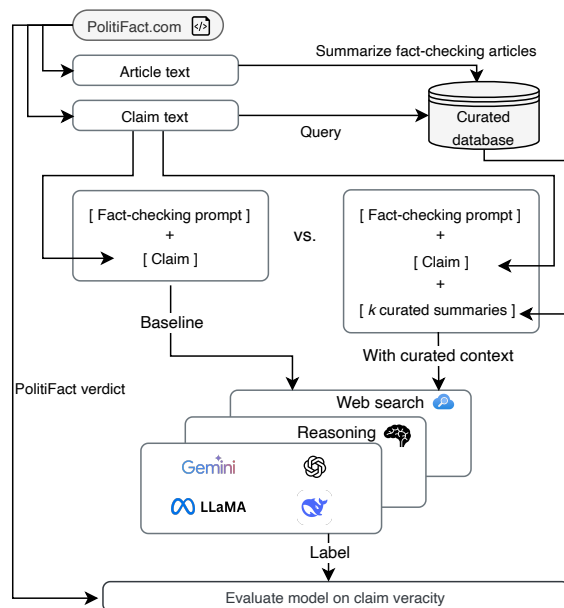


Figure 1: Data pipeline and study design. We collect PolitiFact claims, verdicts, article text, and metadata; generate evidence-focused summaries of fact-checking articles; and build a curated evidence database. We then evaluate 15 LLMs from four major providers with varying capabilities in two conditions, baseline (no retrieval) and with k curated fact-checking article summaries ($k \in 3, 6, 9$). Models predict claim veracity, and we compare their predictions to PolitiFact’s verdicts.

2017), RumorEval (Gorrell et al., 2019), CLEF CheckThat! (Nakov et al., 2022), and ClaimBuster (Hassan et al., 2017) spurred this line of work. Transformer-based models such as BERT produced major accuracy gains on these benchmarks (Soleimani et al., 2020; Nie et al., 2019; Zeng et al., 2021). The emergence of LLMs has recently opened the door to effective end-to-end fact-checking (Pan et al., 2023; Lee et al., 2020).

Here, we focus on political fact-checking, specifically on testing how well AI models can predict the veracity labels ascribed to political claims by professional fact-checkers. For example, Hoes et al. (2023) evaluated GPT-3 on claims verified by Poli-

tiFact¹ and reported nearly 70% accuracy when the task was simplified to a binary true/false classification. For the same task, [Quelle and Bovet \(2024\)](#) found that pairing models with a custom Google-search pipeline improved performance. However, performance drops on PolitiFact’s original six-point scale: [Hoes et al. \(2023\)](#) observed class accuracies from 10% (True claims) to 47% (Mostly True claims). [Quelle and Bovet \(2024\)](#) likewise found low and inconsistent results (0–86%), even with Google search access. More recently, using a small PolitiFact dataset ($n = 228$), [Bombassei De Bona et al. \(2025\)](#) reported macro F1 scores of 0.09–0.48 across models such as Llama2-70B and Mixtral-8x7B. The LLM landscape has since evolved rapidly, with newer, more capable models and additional platforms now widely available.

Many mainstream LLMs now ship with two key features: built-in reasoning modes ([Kavukcuoglu, 2025](#); [OpenAI, 2025](#)) and an integrated retrieval-augmented generation (RAG) system leveraging web-search functionality ([OpenAI, 2024](#); [Mallick and Kilpatrick, 2024](#)). RAG aims to improve factuality by incorporating relevant information from retrieved documents ([Jing et al., 2025](#); [Jiang et al., 2023](#)). It can be implemented via live web search, which enables flexible coverage of emerging topics ([Augenstein et al., 2019](#)), or via a vector database, which offers speed and control over source quality. However, RAG systems have yielded inconsistent improvements for political fact-checking ([Quelle and Bovet, 2024](#); [Fontana et al., 2025](#)).

At the same time, professional organizations are testing LLMs for political fact-checking ([Snopes, 2025](#)) despite little evidence of their reliability for such a demanding task. Others are integrating AI into fact-checking workflows ([Nenno et al., 2025](#); [De et al., 2025](#); [Choi and Ferrara, 2024](#); [Zhou et al., 2026](#)). Yet experimental evidence shows that inaccurate or poorly justified fact-checks can impair people’s ability to judge news headlines ([DeVerna et al., 2024](#)).

Given such concerns, this paper explores two concrete questions: *Can mainstream LLMs with reasoning and web-search capabilities reliably conduct political fact-checking?* If not, *can their accuracy be improved by providing carefully curated context?* We answer these questions by evaluating 15 LLMs from DeepSeek, Meta, Google, and Ope-

nAI, specifically three model categories: (1) *standard* LLMs without advanced reasoning or search; (2) *reasoning* models designed for enhanced inference; and (3) *web-search-enabled* models that augment generation with live internet results. We test these models on 18 years of claims fact-checked by PolitiFact, using a six-label veracity scale aligned with PolitiFact’s labeling system. To provide high-quality context, we also introduce a curated RAG pipeline built on GPT-3.5-generated summaries of PolitiFact fact-checking articles. This design is explicitly intended to isolate and compare the contributions of internal knowledge, reasoning, and web search, each with and without curated high-quality context.

In addition to fact-checking accuracy, LLMs have raised concerns about ideological bias ([Rozado, 2024](#); [Bang et al., 2024](#); [Fulay et al., 2024](#)) and inaccurate citations ([Jaźwińska and Chandrasekar, 2025](#); [Byun et al., 2024](#); [Liu et al., 2023](#)). For the models with web-search capabilities, a related concern is the choice of cited sources ([Yang, 2025b](#); [Jaźwińska and Chandrasekar, 2025](#)). We therefore wish to analyze the sources these models rely on, how reliable those sources are, and to what extent their selection might introduce unintended systematic biases.

Our study (Fig. 1) makes three contributions:

1. We compare 15 recent LLMs from four major providers on a fact-checking task with six labels aligned with PolitiFact’s veracity scale. Our analysis includes both closed commercial and open-weight LLMs of different sizes.
2. We systematically evaluate reasoning and web-search capabilities, both with and without a curated RAG system.
3. We investigate citation practices in search-enabled models, analyzing their impact on performance, alignment with original fact-checking articles, source reliability, and ideological orientation.

Our findings reveal that standard mainstream models, even those with reasoning capabilities, perform poorly on the fact-checking task. While web search capabilities can moderately improve model performance, the cited sources exhibit a strong liberal bias. The models also struggle when there is insufficient information to make a decision. In contrast, providing high-quality curated context significantly improves performance, yielding F1 score increases of 21–351% (mean: 233%) across different settings.

¹politifact.com

These results call for caution in the use of current commercial AI chatbots for nuanced political fact-checking by everyday users; fact-checking may not be a task that can be automatically resolved as the general intelligence and capabilities of LLMs improve. However, when provided with a curated fact-checking database, today’s models already perform well. Therefore, we argue that future efforts in automated fact-checking should focus on improving web search by incorporating and prioritizing high-quality fact-checking context.

2 Data and Methods

2.1 Politifact Data

We evaluate the political fact-checking capabilities of large language models using the complete archive of PolitiFact claims and fact-checks, covering the period from its launch in 2007 through October 2024. We collected these data by systematically crawling the PolitiFact fact-check archive.²

For each claim, we extracted seven fields: the statement, the PolitiFact Truth-O-Meter verdict (six-point veracity labels “True,” “Mostly true,” ..., “Pants on Fire,”³ the claim source, the statement date, the publication date of the fact-check, the topic tags, and the article link (URL). We also retrieved the full text of each fact-checking article using the Python package newspaper3k, as these articles provide the evidence and reasoning used by professional fact-checkers to justify their verdicts. We refer to the fact-checking article associated with a specific claim as the “matching” article.

The crawl yielded 25,224 claims. To ensure quality, we excluded 298 “Flip-o-meter” items (which assess position changes rather than veracity⁴), 55 Spanish-language fact-checks that could not be parsed reliably, and 260 claims for which we encountered article summarization issues (see Fact-Checking Database Construction for details). The final dataset contains 24,611 veracity-focused claims, with 4,941 distinct topic tags and 4,759 unique claim sources, ranging from individual speakers to social-media items (e.g., “Facebook post,” “Viral image”).

²politifact.com/factchecks/list

³politifact.com/article/2018/feb/12/principle-s-truth-o-meter-politifact-methodology-i

⁴politifact.com/article/2008/aug/05/introducing-flip-o-meter

2.2 Curated RAG System for Fact-Checking

A key part of our analysis relies on the construction of a *Curated RAG* system leveraging a database of high-quality information, namely, summarized fact-checking articles. This system differs from RAG systems that rely on web-search functionality provided by off-the-shelf LLMs. Because the web-search and Curated RAG pipelines operate independently, models can draw information from either of them or both simultaneously, as demonstrated in our experimental design (Fig. 1).

Article Summarization Procedure. To create a searchable knowledge base for the Curated RAG system, we summarized each fact-checking article in our collection using OpenAI’s GPT-3.5 (gpt-3.5-turbo-0125; system prompt in Appendix). This procedure was designed to (1) extract key evidence supporting each verdict while minimizing model bias, and (2) handle web scraping artifacts such as malformed or missing text. We formatted the summaries as “<SUMMARY> (PolitiFact verdict: <VERDICT>).” This way, each prompt supplied both article text and PolitiFact verdict, guiding the model to identify the reasoning and evidence underlying the official conclusion.

Faithfulness of Article Summaries. We assessed the reliability of our summaries with established methods (Chen et al., 2024). We focused on summary faithfulness—the extent to which summaries accurately reflect source articles without introducing errors or hallucinations. One author and a trained student coder independently compared 100 randomly sampled summaries to their source articles, assigning each to one of four categories: faithful, minor inaccuracies, major inaccuracies, or hallucinated content (definitions are provided in the Appendix). Two annotators concurred on 97 of 100 cases (97%). One labeled all as “Faithful,” while the other labeled 97 as “Faithful” and three as “Minor Inaccuracies.” These results indicate the summarization process yields reliable representations of the original content.

Fact-Checking Database Construction. Before finalizing the database, we removed 260 claims with unusable fact-checking articles. Of these, 148 returned empty text, and 12 contained fewer than 200 characters, as identified during manual review. Additionally, 100 claims were not captured due to data-collection errors. This left 24,451 high-quality summaries, converted into sentence embeddings

and stored in a Chroma vector database.⁵ Embeddings were generated using Sentence Transformers and the all-MiniLM-L6-v2 model (Reimers and Gurevych, 2019), enabling similarity search via cosine distance.⁶

Although only part of the dataset is used in evaluation (see § 2.4 for details), the database includes all summaries. This requires the system to retrieve relevant items from a much larger pool of information, reflecting likely deployment conditions.



Database Retrieval Accuracy. For each claim, we queried the database with the text <STATEMENT ORIGINATOR> claimed <STATEMENT> and recorded the rank of the “matching summary,” i.e., the fact-checking article summary associated with the tested claim. We queried the database for the top $k = 3, 6,$ or 9 most similar summaries, matching the settings used in our fact-checking analyses (§ 2.4). Performance was measured with top- k accuracy—the fraction of queries in which the correct summary appeared within the top- k results (also known as hit rate@ k).



2.3 Tested LLMs

We evaluated 15 LLMs from four prominent language model providers: OpenAI, Google, Meta, and DeepSeek (Table 1). This selection included models of varying sizes, from compact to large versions. In addition to the standard models, we included widely available models with reasoning and web search functionality. For Google models, the search capability corresponds to their grounding feature,⁷ which allowed us to test these models with and without search enabled. We accessed Google and OpenAI models through their official APIs, while DeepSeek and Meta models were accessed via the TogetherAI platform.⁸

2.4 Fact-Checking Analyses

Experimental Setup. We evaluated each model’s ability to predict PolitiFact’s veracity labels across multiple conditions (Fig. 1). In the zero-shot setting, models received only the claim and its originator, mirroring typical usage scenarios; search-enabled systems could access their built-in web search capabilities and rely on whatever evidence

Table 1: List of tested LLMs, their providers, and capabilities. A ✓ in the  and  columns indicates reasoning and search capabilities, respectively. Model release dates and knowledge cutoffs are provided in the Appendix.

| Provider | Model |  |  |
|----------|---------------------------------------|---|---|
| DeepSeek | DeepSeek-V3 | | |
| DeepSeek | DeepSeek-R1 | ✓ | |
| Meta | Llama-3.2-3B-Instruct-Turbo | | |
| Meta | Llama-3.2-11B-Vision-Instruct-Turbo | | |
| Meta | Llama-3.2-90B-Vision-Instruct-Turbo | | |
| Google | gemini-2.0-flash | | ✓ |
| Google | gemini-2.0-flash-lite | | |
| Google | gemini-2.0-flash-thinking-exp-01-21 | ✓ | ✓ |
| Google | gemini-2.0-pro-exp-02-05 | | |
| OpenAI | gpt-4o-2024-11-20 | | |
| OpenAI | gpt-4o-mini-2024-07-18 | | |
| OpenAI | gpt-4o-mini-search-preview-2025-03-11 | | ✓ |
| OpenAI | gpt-4o-search-preview-2025-03-11 | | ✓ |
| OpenAI | o1-2024-12-17 | ✓ | |
| OpenAI | o3-mini-2025-01-31 | ✓ | |

they retrieved. In the Curated RAG setting, all models were additionally provided with article summaries from our retrieval system and instructed to evaluate claims using the most relevant summaries. By comparing these configurations, we can tease out the effects of internal knowledge, web search, and curated retrieval on fact-checking performance.

We augmented PolitiFact’s Truth-O-Meter scale (hereafter referred to as the Truth Scale) with a “Not Enough Information” label, enabling models to abstain from providing a veracity label rather than forcing a potentially unreliable classification. This design tests whether models can communicate useful uncertainty signals that may allow developers to build safer, more effective fact-checking systems (Kotonya and Toni, 2020; Zhao, 2025). See the Appendix for an analysis of performance across coarser veracity schemes (binary and ternary).

Our protocol used system prompts that define the task and available labels, with claims presented as separate user prompts. Full prompt details are provided in the Appendix, along with a robustness analysis confirming that results hold under a simpler prompt design. Each claim was tested independently with no prior conversation history. We set the temperature to 0 for all tests and used the default web search options.

All analyses were executed on Extreme Science and Engineering Discovery Environment (XSEDE) Jetstream virtual machines (VMs) (Towns et al., 2014; Stewart et al., 2015).

⁵trychroma.com

⁶docs.trychroma.com/docs/embeddings/embedding-functions#default-all-minilm-l6-v2

⁷ai.google.dev/gemini-api/docs/google-search

⁸together.ai

Data Preparation. To reduce computational costs, we created two stratified samples. From the full dataset, we randomly selected half of the claims from each year, yielding a set of 12,306 claims (‘12k’) for evaluating standard models. More computationally intensive reasoning and web-search models were tested on 6,153 claims (‘6k’), drawn from the 12k set by again randomly selecting half the claims from each year. We encountered repeated API errors for a small set of claims, mostly with reasoning models. To ensure fair comparisons, we calculated performance metrics based only on claims that returned valid outputs across all models and tested scenarios, leaving 12,275 and 6,137 claims in the 12k and 6k sets.

Structured Responses and Response Cleaning. We used each API’s structured response features to standardize outputs, but some models occasionally returned malformed JSON. Across six models, these formatting errors produced 1,035 invalid responses: 675 from Llama-3.2-3B-Instruct-Turbo, 250 from Gemini-2.0-flash-thinking-exp-01-21, 61 from Llama-3.2-11B-Vision-Instruct-Turbo, 40 from Gemini-2.0-flash, 6 from Llama-3.2-90B-Vision-Instruct-Turbo, and 3 from DeepSeek-V3. We employed GPT-4o-mini (gpt-4o-mini-2024-07-18)—which did not exhibit these issues—to extract the original classifications and return them in structured form (see Appendix). We validated this process by manually reviewing a random sample of up to 25 problematic responses per model (109 total). The extracted labels matched the originals in all cases (100% agreement).

Performance Metrics. We evaluated performance using standard multi-class metrics: macro precision, macro recall, and macro F1 across Truth Scale classes, weighting each class equally. We also calculated weighted F1 scores, but found no meaningful differences between them and the macro version (see Appendix). The “Not Enough Information” (NEI) label was excluded from macro calculations because no ground truth exists for this category; instead, we analyze NEI usage patterns directly in the Appendix.

2.5 Source Citation Patterns of Web Search-Enabled LLMs

We analyzed the sources cited by web search-enabled LLMs along multiple dimensions: how often they cited original PolitiFact articles, the type

of website used, the source’s reliability, and its political leanings.

Source Type Classification. We classified cited domains into seven categories (fact-checking, news, government, Wikipedia, educational, research, and other) using a rule-based system developed through manual review. The system combines curated third-party datasets with pattern matching.

News domains were identified by matching against a list of over 23,000 news organizations compiled from multiple sources (Yang, 2025a), including NewsGuard,⁹ MBFC,¹⁰ ABYZ,¹¹ and Media Cloud,¹² as well as academic studies (Robertson et al., 2018; Le Quéré et al., 2022; Fischer et al., 2020; Horne et al., 2022).

Other domain types were classified using string patterns applied in priority order: fact-checking sites (e.g., “factcheck.org,” “snopes,” “politifact”); government domains (.gov, .mil, agency-specific patterns); educational institutions (.edu, university-related terms); and research organizations (think tanks, policy institutes, international organizations). See Appendix for additional details.

Source Reliability and Political Orientation. Source reliability is defined at the domain level, based on the March 2025 snapshot of NewsGuard’s Reliability Ratings database.¹³ NewsGuard assigns news domains reliability scores ranging from 0 to 100, with higher values indicating greater credibility, based on independent evaluations by professional journalists.

For ideological orientation, we used the DomainDemo dataset, which profiles over 129,000 domains by the demographics of their Twitter audiences (2011–2022) in line with established classifications (Yang et al., 2025). Political leaning scores range from –1 to +1 (exclusively shared by Democratic or Republican users, respectively).

3 Results

Our analysis proceeds in two parts. First, we compare model performance in zero-shot and Curated RAG settings (§ 3.1). Second, we analyze citation practices of web search-enabled models (§ 3.2).

⁹newsguardtech.com

¹⁰mediabiasfactcheck.com

¹¹abyznewslinks.com

¹²mediacloud.org

¹³newsguardtech.com/solutions/news-reliability-ratings

3.1 Fact-Checking Performance

To balance coverage and cost, we used a two-stage evaluation. First, we tested the less expensive standard models across multiple Curated RAG configurations ($k = 3, 6, 9$) on the full 12k set. Second, guided by those results, we evaluated the more costly reasoning and web search-enabled models on the 6k set with simplified configurations (only $k = 6$). For comparability, we report macro F1 results on the 6k set. The Appendix reports full results on the 12k set, showing highly similar performance across both sets, as well as performance broken down by veracity verdict.

Standard Models and Curated RAG Accuracy.

Figure 2(a) shows macro F1 fact-checking performance for standard models across configurations. In the zero-shot setting, where models rely solely on internal knowledge, performance is consistently low ($F1 \approx 0.1$ – 0.3).

Curated RAG context leads to substantial improvements across all models. F1 gains range from $+0.26$ (Llama-3.2-3B) to $+0.66$ (GPT-4o mini), with little variation across k values ($SD = 0.002$ – 0.037). The best configurations are GPT-4o at $k = 9$ ($F1 = 0.90$) and DeepSeek-V3 at $k = 9$ ($F1 = 0.89$). These gains from Curated RAG far exceed the differences in performance between models. Within the Llama-3.2 family, larger variants not only start from higher baselines but also achieve greater improvements from Curated RAG, suggesting that model capacity amplifies the benefits of high-quality retrieval. Viewed as relative improvements, the largest gains within each model family are substantial: approximately $+295\%$ for GPT, $+250\%$ for Gemini, $+351\%$ for Llama, and $+259\%$ for DeepSeek (see Appendix for details).

Figure 2 also highlights retrieval effectiveness. Performance is stable across k values, reflecting a retrieval system that reliably ranks the correct summary near the top: the median rank is 1 and the mean rank is 3 (Fig. 2(b)). Top- k accuracy is correspondingly high: 0.96 at $k = 3$ and 0.98 at $k = 6$ or 9 (Fig. 2(c)).

Reasoning and Web Search-enhanced Models.

We next examine how reasoning and web search affect macro F1 scores. Figure 3(a) shows that reasoning models do not consistently outperform the baselines. When gains occur, they are modest ($+0.06$ on average), and performance can even decline (-0.03 for Gemini 2.0 Flash Thinking).

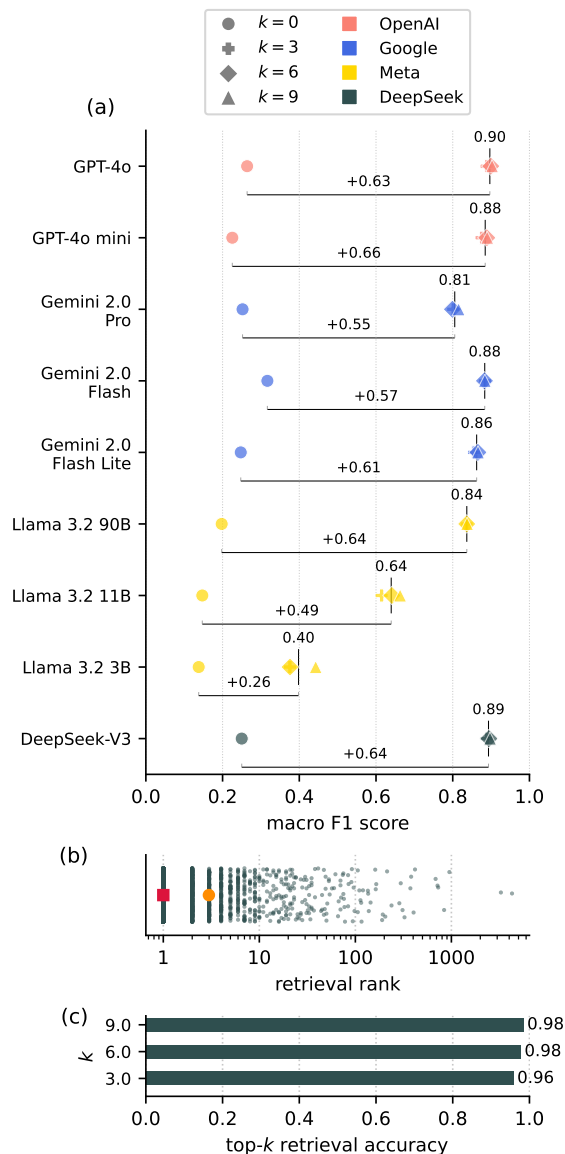


Figure 2: Fact-checking performance of standard LLMs and retrieval performance of the Curated RAG system. (a) Macro F1 scores in zero-shot ($k = 0$) and Curated RAG ($k > 0$) conditions. Shapes and colors denote Curated RAG settings and model providers. Vertical bars show average F1 scores across Curated RAG settings ($k = 3, 6, 9$); horizontal annotations show improvement over zero-shot. (b) Distribution of retrieval ranks for matching summaries across tested claims; the red square marks the median and the orange circle marks the mean. The y-axis displays random jitter for visualization clarity. (c) Top- k retrieval accuracy for each setting.

These results suggest that advanced reasoning capabilities provide limited benefit in the zero-shot setting. On the other hand, Curated RAG summaries deliver substantial improvements for reasoning models, mirroring patterns observed in the evaluation of the standard models.

Search-enhanced models exhibit provider-

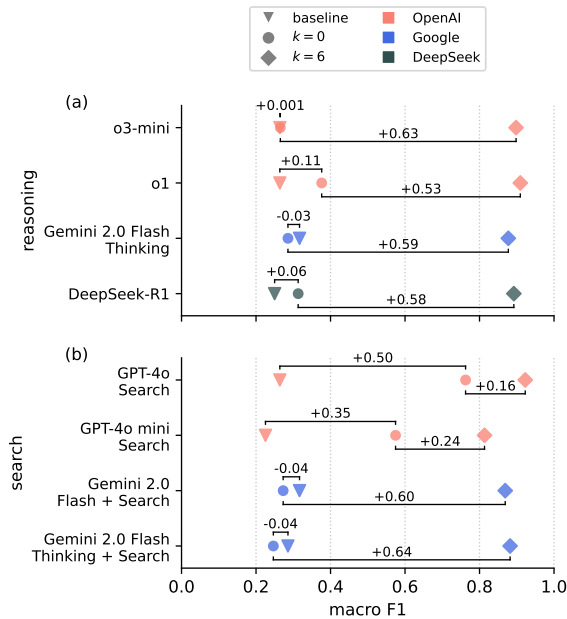


Figure 3: Fact-checking performance of LLMs with (a) reasoning and (b) web search capabilities. Triangles show zero-shot performance of the corresponding standard models used as baselines: GPT-4o for o3-mini and o1; Gemini 2.0 Flash for Flash Thinking; DeepSeek-V3 for R1; and the non-search equivalent for search-enabled models (e.g., GPT-4o for GPT-4o Search). Circles denote zero-shot performance ($k = 0$) of reasoning and search-enhanced models, respectively, in (a) and (b), while diamonds show their Curated RAG-enhanced performance at $k = 6$. Horizontal annotations indicate performance differences: zero-shot reasoning/search variants compared with baselines (above symbols) and with the Curated RAG setting ($k = 6$; below symbols).

specific differences (Fig. 3(b)). GPT search models substantially outperform their non-search versions, gaining +0.50 (GPT-4o) and +0.35 (GPT-4o mini). However, Gemini search models underperform relative to their standard variants. This is likely due to difficulties in retrieving effective citations, as discussed in § 3.2.

Curated RAG summaries continue to provide meaningful improvements: +0.24 for GPT-4o mini Search and +0.16 for GPT-4o Search. Gemini models also show large gains in the Curated RAG setting (+0.60 for Flash, +0.64 for Flash Thinking), though these partially reflect their weak zero-shot baselines.

3.2 Citation Usage

To interpret web-search-enabled model performance, we examine citation patterns. Table 2 reports URL citation rates in generated fact-checks. Gemini models rarely include URLs, helping ex-

Table 2: Citation statistics for search-enabled LLM-generated fact-checks. Columns show model name, Curated RAG setting (k), and counts and percentages of fact-checks containing any URL or an exact match to the original PolitiFact article.

| Model | k | Any URL (%) | Exact Match (%) |
|---------------------------|-----|-------------|-----------------|
| Gemini 2.0 Flash | 0 | 14 (0) | 0 (0) |
| Gemini 2.0 Flash | 6 | 0 (0) | 0 (0) |
| Gemini 2.0 Flash Thinking | 0 | 0 (0) | 0 (0) |
| Gemini 2.0 Flash Thinking | 6 | 0 (0) | 0 (0) |
| GPT-4o mini Search | 0 | 3,861 (63) | 2,691 (44) |
| GPT-4o mini Search | 6 | 3,638 (59) | 2,780 (45) |
| GPT-4o Search | 0 | 3,932 (64) | 2,814 (46) |
| GPT-4o Search | 6 | 4,448 (72) | 3,646 (59) |

plain why search fails to improve their performance (Fig. 3(b)). However, search-enabled GPT models cite URLs frequently: 59%–72% of responses include at least one citation, and 44%–59% link directly to the original PolitiFact article.

To understand how citation behavior shapes fact-checking performance, we focus on GPT models, as Gemini models seldom include any citations. We calculate macro F1 scores for search-enabled GPT models across three conditions: responses in which any citation matches the correct PolitiFact article, responses with one or more citations that do not match the original fact-checking article, and responses with no URLs cited. We examine these patterns across both zero-shot ($k = 0$) and curated RAG settings ($k = 6$). Figure 4 shows that performance is consistently highest when the correct article is cited, while citing other URLs is generally associated with the largest drop in macro F1 across models and settings (see Appendix for details). Notably, responses with no citations generally outperform those with other citations, suggesting that retrieving and relying on irrelevant evidence can hinder performance. This effect is substantially moderated when a curated RAG context is provided ($k = 6$).

To assess the sources that search-enhanced LLMs rely on beyond PolitiFact itself, we focus on GPT models, as Gemini models seldom include citations. We extract the top-level domains from all cited URLs (e.g., cnn.com for CNN) and present the results in Figure 5.

Using the classification approach described in § 2, we categorize cited sources by domain type. Figure 5(a) shows that both GPT models predominantly cite fact-checking sites, followed by news outlets, government websites, and other sources. Citations to Wikipedia, educational, and research websites are relatively infrequent.

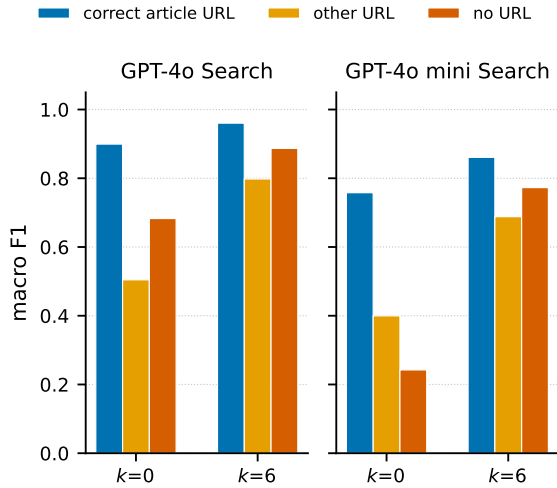


Figure 4: Macro F1 scores for web search-augmented models across three retrieval outcome categories: responses in which any citation matched the correct PolitiFact article, responses with one or more citations that did not match the original fact-checking article, and responses with no URLs cited.

Figures 5(b,c) highlight the sources cited most often. The top ten sources are consistent across both GPT models. PolitiFact is the most cited source, with [factcheck.org](https://www.factcheck.org), [snopes.com](https://www.snopes.com), and Wikipedia also appearing prominently. Mainstream news outlets such as Associated Press and CNN rank among the most cited sources, alongside government websites like [senate.gov](https://www.senate.gov) and [cdc.gov](https://www.cdc.gov).

Finally, to evaluate the reliability and political orientation of cited sources, we assign NewsGuard reliability scores and DomainDemo political leaning scores (Yang et al., 2025), respectively, to the extracted top-level domains of all cited URLs. Figure 6 shows the joint distribution of these metrics across all tested GPT scenarios, limited to domains for which both scores are available. The distribution peaks at a NewsGuard score of 100 and a political leaning score of -0.3 , corresponding to [politifact.com](https://www.politifact.com). In general, we observe a tendency to cite left-leaning sources with high credibility ratings. These patterns persist when we remove [politifact.com](https://www.politifact.com) from the dataset and when we re-run the analysis using an alternative domain-quality list (Lin et al., 2023) (see Appendix). They are also consistent with prior research related to AI news citation patterns (Yang, 2025b).

4 Discussion

Our study shows that despite rapid progress, major LLMs still perform poorly at fact-checking when

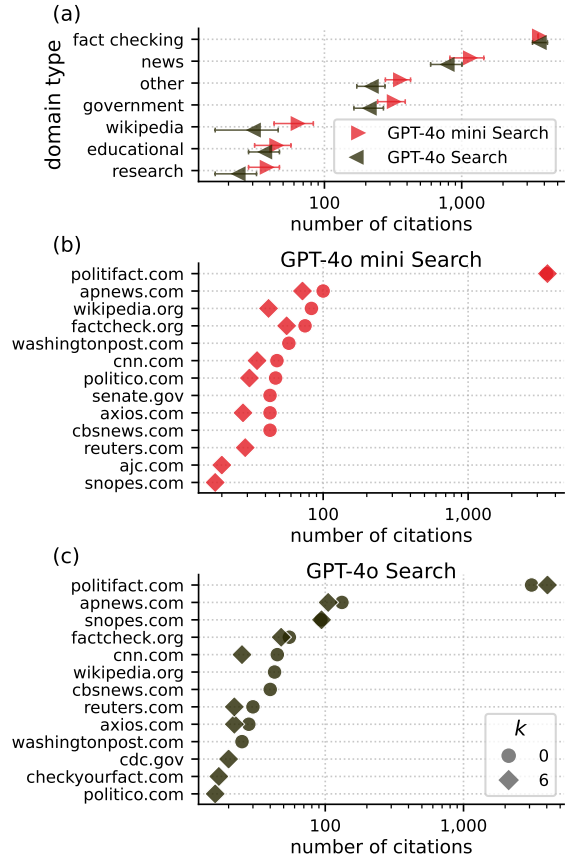


Figure 5: Sources cited by search-enhanced GPT models. (a) Average number of citations by domain type. Error bars indicate standard deviation across k values. (b) Top 10 sources cited by GPT-4o mini Search for $k = 0$ and $k = 6$. (c) Same as (b) for GPT-4o Search.

limited to internal knowledge, with weak zero-shot results across providers. Reasoning-enhanced models yielded only small and inconsistent gains.

Web search helps in some cases. In particular, GPT models improved substantially, often citing PolitiFact or other credible sources, whereas Gemini models struggled to surface relevant citations. Performance also varied substantially depending on whether models retrieved the correct fact-checking article, a different source, or no source at all. This gap shows that “search” is not a uniform capability: it depends on query formulation, source prioritization, and the integration of retrieved material into the model’s reasoning process.

GPT models’ overall citation mix skewed leftward, raising questions about ideological balance. Whether this reflects model or search-pipeline bias, or structural features of the information ecosystem (e.g., a correlation between accuracy and political leaning (Fulay et al., 2024)), it may undermine trust in systems that present themselves as arbiters of

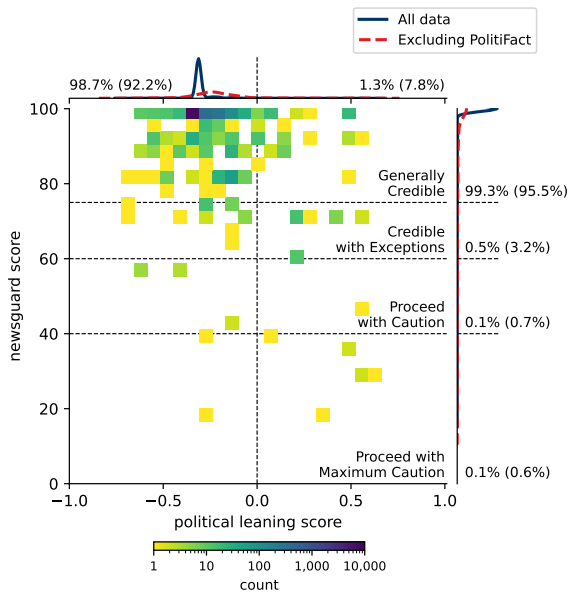


Figure 6: Joint distribution of NewsGuard reliability and political leaning scores for sources cited by search-enhanced GPT models. Marginal distributions are shown in the top and right panels for all citations (blue) and for citations excluding `politifact.com` (red). Black dashed lines separate NewsGuard group labels, and annotated percentages indicate the share of sources falling in each group; values in parentheses report the same percentages with PolitiFact excluded.

truth, especially in politically sensitive settings.

By supplying high-quality, claim-specific evidence, the Curated RAG system improves macro F1 by 233% on average across models, demonstrating that leading LLMs can map evidence to correct labels when quality information is available. In other words, the key limitation of LLMs is not how models reason over information, but whether they have access to the right information in the first place.

Overall, these findings suggest a nuanced picture. For everyday users, caution is warranted; standard LLMs remain unreliable fact-checkers, and web search—while helpful—does not ensure the right information is found or interpreted correctly. For researchers and system designers, the lesson is that carefully curated retrieval pipelines are currently the most dependable foundation for reliable fact-checking. Building such pipelines at scale, however, is non-trivial: they must cover diverse domains, update continuously, and manage conflicting or incomplete evidence. Moreover, the successful deployment of such a system creates a structural tension within the information ecosystem. Wide use of LLM-based fact-checking tools could

reduce direct traffic to fact-checking websites, undermining the revenue models of the organizations whose work these pipelines depend on. Addressing these challenges will require not only advances in how evidence is surfaced, but also careful attention to the sustainability of the broader fact-checking ecosystem on which such systems depend.

5 Limitations

This study has several limitations. First, our evaluation is limited to PolitiFact and may not generalize to other fact-checkers, platforms, or types of claims (Yu et al., 2023). Second, we do not address the “breaking news problem” (DeVerna et al., 2024): most tested claims predate evaluation, so performance on newly emerging claims may be worse; while web-search-enabled models could fetch live evidence in principle, we did not test this. Future work should evaluate performance in breaking news settings with incomplete or rapidly evolving information. Third, results may not generalize to other LLM models; reliance on vendor APIs means results may change as providers update or modify their models, limiting reproducibility over time. Fourth, results may not generalize across all prompting strategies; while our prompt robustness analysis (see Appendix) suggests that our results are not driven solely by our prompting strategy, more work is needed to understand how people actually use LLMs to fact-check information. Finally, we report system-level performance only and do not measure downstream effects on beliefs, trust, or sharing behavior, which are critical for assessing societal impact (Guay et al., 2023).

Code and data. Replication code and data are available on GitHub¹⁴ and Zenodo (DeVerna et al., 2025), respectively.

Acknowledgments. Funding was provided by the Institute for Humane Studies and the Knight Foundation. We thank Eduardo Blanco for feedback and Abhi Boda for assistance with the summary faithfulness analysis.

Author contributions. M.R.D. conceived the study; M.R.D., K.-C.Y., and H.Y.Y. designed the research; M.R.D. collected/analyzed data and created visualizations with feedback from all authors; M.R.D. wrote the original draft, reviewed and edited by all authors; M.R.D. and F.M. acquired funding; F.M. supervised.

¹⁴github.com/osome-iu/fact_check_rag_osome

References

- Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2024. [Factuality challenges in the era of large language models and opportunities for fact-checking](#). *Nature Machine Intelligence*, 6:852–863.
- Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. [Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 4684–4696. Association for Computational Linguistics.
- Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. [Measuring Political Bias in Large Language Models: What Is Said and How It Is Said](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11142–11159. Association for Computational Linguistics.
- Francesco Bombassei De Bona, David La Barbera, Stefano Mizzaro, and Kevin Roitero. 2025. [A Comparative Analysis of Retrieval-Augmented Generation and Crowdsourcing for Fact-Checking](#). In *Advances in Information Retrieval*, pages 446–454. Springer Nature Switzerland.
- Courtnei Byun, Piper Vasicek, and Kevin Seppi. 2024. [This Reference Does Not Exist: An Exploration of LLM Citation Accuracy and Relevance](#). In *Proceedings of the Third Workshop on Bridging Human-Computer Interaction and Natural Language Processing*, pages 28–39. Association for Computational Linguistics.
- Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. [How People Use ChatGPT](#). Working Paper 34255, National Bureau of Economic Research.
- Canyu Chen and Kai Shu. 2024. [Combating misinformation in the age of llms: Opportunities and challenges](#). *AI Magazine*, 45(3):354–368.
- Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024. [Complex claim verification with evidence retrieved in the wild](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, page 3569–3587. Association for Computational Linguistics.
- Eun Cheol Choi and Emilio Ferrara. 2024. [Automated Claim Matching with Large Language Models: Empowering Fact-Checkers in the Fight Against Misinformation](#). In *Companion Proceedings of the ACM Web Conference 2024, WWW '24*, page 1441–1449. Association for Computing Machinery.
- Soham De, Michiel A. Bakker, Jay Baxter, and Martin Saveski. 2025. [Supernotes: Driving consensus in crowd-sourced fact-checking](#). In *Proceedings of the ACM on Web Conference 2025, WWW '25*, page 3751–3761, New York, NY, USA. Association for Computing Machinery.
- Matthew DeVerna, Kai-Cheng Yang, Harry Yaojun Yan, and Filippo Menczer. 2025. [Data: Large Language Models Require Curated Context for Reliable Political Fact-Checking—Even with Reasoning and Web Search](#).
- Matthew R. DeVerna, Harry Yaojun Yan, Kai-Cheng Yang, and Filippo Menczer. 2024. [Fact-checking information from large language models can decrease headline discernment](#). *Proceedings of the National Academy of Sciences*, 121(50):e2322823121.
- Ullrich Ecker, Jon Roozenbeek, Sander van der Linden, Li Qian Tay, John Cook, Naomi Oreskes, and Stephan Lewandowsky. 2024. [Misinformation poses a bigger threat to democracy than you might think](#). *Nature*, 630(8015):29–32.
- Sean Fischer, Kokil Jaidka, and Yphtach Lelkes. 2020. [Auditing local news presence on Google News](#). *Nature Human Behaviour*, 4(12):1236–1244.
- Nicolo' Fontana, Francesco Corso, Enrico Zuccolotto, and Francesco Pierri. 2025. [Evaluating open-source large language models for automated fact-checking](#). *Preprint*, arXiv:2503.05565.
- Suyash Fulay, William Brannon, Shrestha Mohanty, Cassandra Overney, Elinor Poole-Dayana, Deb Roy, and Jad Kabbara. 2024. [On the Relationship between Truth and Political Bias in Language Models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 9004–9018. Association for Computational Linguistics.
- Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. [SemEval-2019 Task 7: RumourEval, Determining Rumour Veracity and Support for Rumours](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854. Association for Computational Linguistics.
- Brandon Guay, Adam J. Berinsky, Gordon Pennycook, and 1 others. 2023. [How to think about whether misinformation interventions work](#). *Nature Human Behaviour*, 7:1231–1233.
- Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane,

- Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. 2017. [ClaimBuster: the first-ever end-to-end fact-checking system](#). *Proc. VLDB Endow.*, 10(12):1945–1948.
- Emma Hoes, Sacha Altay, and Juan Bermeo. 2023. [Leveraging ChatGPT for Efficient Fact-Checking](#). Preprint [PsyArXiv].
- Benjamin D Horne, Maurício Gruppi, Kenneth Joseph, Jon Green, John P Wihbey, and Sibel Adali. 2022. [Nela-local: A dataset of US local news articles for the study of county-level news ecosystems](#). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1275–1284.
- Klaudia Jazwińska and Aisvarya Chandrasekar. 2025. [AI Search Has A Citation Problem](#). *Columbia Journalism Review*.
- Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. [Active Retrieval Augmented Generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992. Association for Computational Linguistics.
- Zhi Jing, Yongye Su, and Yikun Han. 2025. [When Large Language Models Meet Vector Databases: A Survey](#). In *2025 Conference on Artificial Intelligence x Multimedia (AIxMM)*, pages 7–13.
- Koray Kavukcuoglu. 2025. [Gemini 2.5: Our most intelligent AI model](#). Google The Keyword. Accessed: 2025-08-15.
- Neema Kotonya and Francesca Toni. 2020. [Explainable automated fact-checking: A survey](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, page 5430–5443. International Committee on Computational Linguistics.
- Marianne Aubin Le Quéré, Ting-Wei Chiang, and Mor Naaman. 2022. [Understanding Local News Social Coverage and Engagement at Scale during the COVID-19 Pandemic](#). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 560–572.
- Nayeon Lee, Belinda Z. Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabza. 2020. [Language Models as Fact Checkers?](#) In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 36–41. Association for Computational Linguistics.
- Hause Lin, Jana Lasser, Stephan Lewandowsky, Rocky Cole, Andrew Gully, David G. Rand, and Gordon Pennycook. 2023. [High level of correspondence across different news domain quality rating sets](#). *PNAS Nexus*, 2(9):pgad286.
- Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. [Evaluating Verifiability in Generative Search Engines](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025. Association for Computational Linguistics.
- Shrestha Basu Mallick and Logan Kilpatrick. 2024. [Gemini API and Google AI Studio now offer Grounding with Google Search](#). Google for Developers. Accessed: 2025-08-15.
- Filippo Menczer, David Crandall, Yong-Yeol Ahn, and Apu Kapadia. 2023. [Addressing the Harms of AI-Generated Inauthentic Content](#). *Nature Machine Intelligence*, 5:679–680.
- Preslav Nakov, Alberto Barrón-Cedeño, Giovanni da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Gautam Kishore Shahi, Hamdy Mubarak, Alex Nikolov, Nikolay Babulov, Yavuz Selim Kartal, Michael Wiegand, Melanie Siegel, and Juliane Köhler. 2022. [Overview of the CLEF–2022 CheckThat! Lab on Fighting the COVID-19 Infodemic and Fake News Detection](#). In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 495–520. Springer International Publishing.
- Sami Nenko, Cornelius Puschmann, Kamil Fuławka, and Philipp Lorenz-Spreen. 2025. [Content-based detection of misinformation expands its scope across politicians and platforms](#). SocArXiv[PrePrint].
- Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. [Combining Fact Extraction and Verification with Neural Semantic Matching Networks](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6859–6866.
- OpenAI. 2024. [Introducing ChatGPT Search](#). Product release. Accessed: 2025-08-15.
- OpenAI. 2025. [Introducing OpenAI o3 and o4-mini](#). Product release. Accessed: 2025-08-15.
- Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. [Fact-Checking Complex Claims with Program-Guided Reasoning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004. Association for Computational Linguistics.
- Dorian Quelle and Alexandre Bovet. 2024. [The perils and promises of fact-checking with large language models](#). *Frontiers in Artificial Intelligence*, 7.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 3980–3990. Association for Computational Linguistics.

- Ronald E. Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. [Auditing Partisan Audience Bias within Google Search](#). *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW).
- David Rozado. 2024. [The Political Preferences of LLMs](#). *PLOS ONE*, 19(7):e0306621.
- Snopes. 2025. [Factbot by snopes](#). Web Dashboard. Accessed: 2025-08-15.
- Amir Soleimani, Christof Monz, and Marcel Worring. 2020. [BERT for Evidence Retrieval and Claim Verification](#). In *Advances in Information Retrieval*, pages 359–366. Springer International Publishing.
- Craig A. Stewart, Timothy M. Cockerill, Ian Foster, David Hancock, Nirav Merchant, Edwin Skidmore, Daniel Stanzione, James Taylor, Steven Tuecke, George Turner, Matthew Vaughn, and Niall I. Gaffney. 2015. Jetstream: A self-provisioned, scalable science and engineering cloud environment. In *Proceedings of the XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*, XSEDE '15, pages 1–8. Association for Computing Machinery.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a Large-scale Dataset for Fact Extraction and VERification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819. Association for Computational Linguistics.
- J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr. 2014. Xsede: Accelerating scientific discovery. *Computing in Science & Engineering*, 16(5):62–74.
- William Yang Wang. 2017. [“Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426. Association for Computational Linguistics.
- World Economic Forum. 2024. [Global risks report 2024](#). Accessed: 2025-08-15.
- Kai-Cheng Yang. 2025a. [A list of news domains](#). Zenodo.
- Kai-Cheng Yang. 2025b. [News source citing patterns in ai search systems](#). *Preprint*, arXiv:2507.05301.
- Kai-Cheng Yang, Pranav Goel, Alexi Quintana-Mathé, Luke Horgan, Stefan D. McCabe, Nir Grinberg, Kenneth Joseph, and David Lazer. 2025. [DomainDemo: a dataset of domain-sharing activities among different demographic groups on Twitter](#). *Scientific Data*, 12(1):1251.
- Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hananeh Hajishirzi. 2023. [CREPE: Open-Domain Question Answering with False Presuppositions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10457–10480. Association for Computational Linguistics.
- Xia Zeng, Amani S. Abumansour, and Arkaitz Zubiaga. 2021. [Automated Fact-Checking: A Survey](#). *Language and Linguistics Compass*, 15(10):e12438.
- Celina Zhao. 2025. [Ai hallucinates because it’s trained to fake answers it doesn’t know](#). *Science*, 390(6773):558–559.
- Xinyi Zhou, Ashish Sharma, Amy X. Zhang, and Tim Althoff. 2026. [Correcting misinformation on social media with a large language model](#). *Preprint*, arXiv:2403.11169.

Appendix

1 Article Summary Prompts

We include here the prompts used to summarize the scraped PolitiFact fact-checking articles. In each case, we leverage both “system” and “user” prompts. System prompts provide high-level instructions, while user prompts contain claim-specific content.

System prompt

As an AI assistant, your task is to summarize text from a PolitiFact fact-checking article. The input text may contain incomplete sentences, HTML tags, and non-textual elements. First, clean the text by removing any irrelevant content or formatting issues. Then, write a concise, neutral summary focusing on the article’s main conclusion and supporting facts, covering who, what, when, where, and why.

The summary should be one paragraph, free of editorializing or subjective interpretation. Provide only the summary with no additional text or comments. If no article text is provided, respond with ‘No article text provided.’ Follow these instructions strictly to ensure an accurate, unbiased summary.

User prompt

*Verdict: <POLITIFACT CLAIM VERDICT>
<POLITIFACT ARTICLE TEXT>*

2 Summary Faithfulness

To assess the reliability of our summaries, we followed the framework described in [Chen et al. \(2024\)](#). Specifically, the method defines four categories of faithfulness: faithful, minor inaccuracies,

Table A1: Labels and definitions used to evaluate article summary faithfulness, taken from [Chen et al. \(2024\)](#).

| Label | Definition |
|----------------------|---|
| Faithful | Accurately reflects the meaning and details of the original article without any factual errors. |
| Minor Inaccuracies | Contains slight factual inaccuracies that do not significantly alter the main conclusion. |
| Major Inaccuracies | Contains significant factual errors that distort the core meaning of the article. |
| Hallucinated Content | Introduces content not present in the original article. |

major inaccuracies, and hallucinated content. Table A1 presents their definitions, which were used by the annotators to code the LLM-generated article summaries.

3 Fact-Checking Prompts

This section presents the prompts used in the zero-shot and Curated RAG fact-checking tests.

System prompt: Zero-shot

As an AI fact checker, your task is to evaluate the accuracy of a CLAIM by assigning a label from the ‘Truth Scale’ and providing a justification for that label. Each claim will include a ‘STATEMENT ORIGINATOR’ indicating the source of the claim to assist you.

Truth Scale:

- *True* - The statement is accurate and there’s nothing significant missing.
- *Mostly true* - The statement is accurate but needs clarification or additional information.
- *Half true* - The statement is partially accurate but leaves out important details or takes things out of context.
- *Mostly false* - The statement contains an element of truth but ignores critical facts that would give a different impression.
- *False* - The statement is not accurate.
- *Pants on fire* - The statement is not accurate and makes a ridiculous claim.
- *Not enough information* - There is not enough information to reliably apply one of the other labels.

Instructions:

1. Evaluate the claim using the most relevant information you have.
2. If you do not have enough information, use the ‘Not enough information’ label.
3. Consider nuances and subtle details that could influence the claim’s accuracy.
4. Choose the most appropriate label from the ‘Truth Scale’ and explain your reasoning concisely.

User prompt: Zero shot

*STATEMENT ORIGINATOR: <STATEMENT ORIGINATOR>
CLAIM: <CLAIM>*

System prompt: Curated RAG fact-checking

As an AI fact checker, your task is to evaluate the

accuracy of a CLAIM by assigning a label from the ‘Truth Scale’ and providing a justification for that label. Each claim will include a ‘STATEMENT ORIGINATOR’ indicating the source of the claim, along with ‘FACT-CHECKING INFORMATION’ summarizing relevant PolitiFact fact-checks to assist you.

Truth Scale:

- *True* - The statement is accurate and there’s nothing significant missing.
- *Mostly true* - The statement is accurate but needs clarification or additional information.
- *Half true* - The statement is partially accurate but leaves out important details or takes things out of context.
- *Mostly false* - The statement contains an element of truth but ignores critical facts that would give a different impression.
- *False* - The statement is not accurate.
- *Pants on fire* - The statement is not accurate and makes a ridiculous claim.
- *Not enough information* - There is not enough information to reliably apply one of the other labels.

Instructions:

1. Evaluate the claim using the most relevant ‘FACT-CHECKING INFORMATION’ provided.
2. If the provided ‘FACT-CHECKING INFORMATION’ is not relevant to the statement, use the ‘Not enough information’ label.
3. Consider nuances and subtle details that could influence the claim’s accuracy.
4. Choose the most appropriate label from the ‘Truth Scale’ and explain your reasoning concisely.

User prompt: Curated RAG fact-checking

STATEMENT ORIGINATOR: <STATEMENT ORIGINATOR>

CLAIM: <CLAIM>

FACT-CHECKING INFORMATION:

Summary 1: <RETRIEVED SUMMARY #1>

Summary 2: <RETRIEVED SUMMARY #2>

...

Summary k: <RETRIEVED SUMMARY #k>

Structured Responses. At the time of testing, structured response features were unavailable for

Gemini reasoning or Google’s “grounding” (web search) functionality, and the TogetherAI API likewise did not support them for Llama or DeepSeek models. To address this limitation, we augmented the fact-checking prompts with an additional instruction in the “Instructions” section of each system prompt. Specifically, we used the same prompts described above, adding the following item in both the zero-shot and Curated RAG settings:

5. Please provide your response in valid JSON format in plain text, without enclosing it in backticks or any other formatting markers. The response should include exactly two keys: ‘label’ and ‘justification’. Do not add any extra characters before or after the JSON object.

Given these API constraints, the adjustment was applied to Gemini models with reasoning or search enabled, as well as all Llama and DeepSeek models. To assess its effectiveness, we calculated the malformed-JSON rate for these scenarios only.

The Llama variants (3B, 11B, and 90B) produced 147,300 responses (12,275 claims × 4 k settings × 3 models). Gemini Thinking and Gemini 2.0 Flash with grounding produced 24,548 responses (6,137 claims × 2 k settings × 2 models). DeepSeek V3 contributed 49,100 responses (12,275 claims × 4 k settings), and DeepSeek R1 contributed 12,274 responses (6,137 claims × 2 k settings). Altogether, these scenarios yielded 233,222 responses, of which 1,035 were unparsable, corresponding to a rate of 0.44%—fewer than one in every 200 generations—indicating that our approach was highly effective.

4 Cleaning Malformed JSON Prompt

As discussed in the Data and Methods section, we employ the OpenAI API to extract fact-checking labels from unparsable responses. We call the gpt-4o-mini-2024-07-18 model with the following prompt.

System prompt: Label extraction

Your task is to extract information from unstructured AI fact checks and provide it, unaltered, in valid JSON format. This output must have exactly two keys:

1. ‘label’: the fact-checking label
2. ‘justification’: the rationale or explanation for that label

Instructions:

1. Read and understand the provided text, which contains malformed JSON.

2. *Extract the relevant information, preserving every piece of content exactly as it appeared (no additions, removals, or modifications).*
3. *Produce valid JSON that contains only the keys ‘label’ and ‘justification’ as described above.*
 - *Do not alter the text of the label or the justification in any way.*
 - *Do not include any additional keys, text, comments, or explanations.*
 - *The final output must be strictly valid JSON and nothing else.*

User prompt: Label extraction

<MALFORMED JSON>

5 Performance Statistics

Table A2 reports full performance statistics for all tested models and settings, and Table A3 reports relative improvements in macro F1 obtained with the Curated RAG system compared to baseline LLMs. Table A4 shows performance by PolitiFact label for the best- and worst-performing models. We omit the complete by-label breakdown, which would require 312 rows (52 tests × 6 labels), but provide aggregate performance in the [Performance by Veracity](#) section, and detailed statistics in our public repository.

6 Macro Versus Weighted F1

Aggregate performance in multi-label tasks can be reported in different ways. In the main text, we use macro F1, which weights each PolitiFact label equally. An alternative is weighted F1, which weights each class’s contribution by the number of items it contains, giving more influence to larger classes. Figure A1 shows that macro and weighted F1 are highly correlated and yield no meaningful differences in reported performance.

7 Performance on Samples with Different Sizes

The main text reports results on the 6k subset to align with reasoning and web search–enhanced models. Standard LLMs were also evaluated on the 12k dataset. Figure A2 shows high consistency between macro F1 scores on both datasets across zero-shot and Curated RAG settings, confirming that the 6k subset is representative of performance on the full dataset.

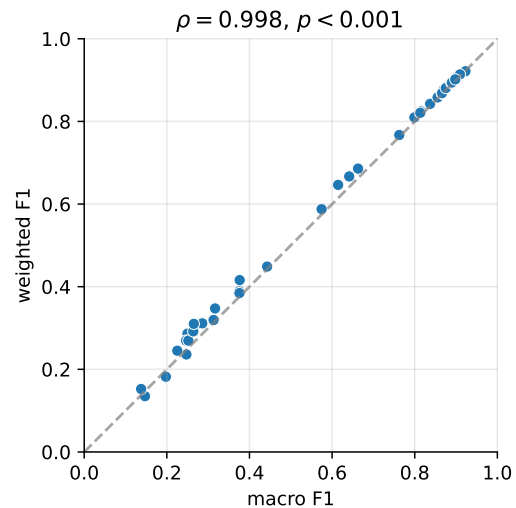


Figure A1: Comparison of macro and weighted fact-checking F1 scores on the 6k set of claims on all tests. Each dot represents a specific test setting. The Spearman’s correlation (ρ) and level of significance are included above.

8 Performance by Veracity

Model performance was also analyzed at the level of individual veracity labels using the 6k set of claims. We calculate the F1 score for claims in each original PolitiFact verdict across all test conditions reported in the main text. Figure A3 includes the average of these scores across all models and for each individual model family. On average, models perform worse on the “Pants on Fire” label, a pattern that appears consistently across models and settings. This finding likely reflects two factors. First, “Pants on Fire” represents an extreme category requiring models to distinguish ordinary falsehoods from claims deemed absurd or egregiously misleading, where human judgment may also vary. Second, as shown in Figure A4, retrieval for these claims appears slightly less reliable on average, increasing uncertainty in cases that demand the strongest classification.

9 Assessing Model Awareness of Missing Information

The introduction of “Not Enough Information” (NEI) responses complicates the assessment of response “correctness.” However, abstention is a critical component of responsible AI systems, and developers building fact-checking systems may prefer predictable behavior, such as consistently abstaining when retrieval fails to return relevant evidence. This perspective informs our evaluation framework.

Table A2: Fact-checking model performance. Columns show the retrieval setting (k), macro precision, macro recall, macro F1, weighted F1, accuracy, and support (number of claims). A \checkmark in the \mathcal{Q} and \mathcal{R} columns denote web search and reasoning capabilities.

| Model | \mathcal{Q} | \mathcal{R} | k | P (macro) | R (macro) | F1 (macro) | F1 (wt) | Acc. | Sup. |
|---------------------------|---------------|---------------|-----|-----------|-----------|------------|---------|------|------|
| DeepSeek-V3 | | | 0 | 0.36 | 0.22 | 0.25 | 0.29 | 0.26 | 6137 |
| DeepSeek-V3 | | | 3 | 0.97 | 0.83 | 0.89 | 0.89 | 0.84 | 6137 |
| DeepSeek-V3 | | | 6 | 0.97 | 0.84 | 0.89 | 0.90 | 0.85 | 6137 |
| DeepSeek-V3 | | | 9 | 0.97 | 0.84 | 0.90 | 0.90 | 0.85 | 6137 |
| DeepSeek-R1 | | \checkmark | 0 | 0.37 | 0.31 | 0.31 | 0.32 | 0.31 | 6137 |
| DeepSeek-R1 | | \checkmark | 6 | 0.96 | 0.84 | 0.89 | 0.90 | 0.85 | 6137 |
| Llama 3.2 3B | | | 0 | 0.43 | 0.21 | 0.14 | 0.15 | 0.20 | 6137 |
| Llama 3.2 3B | | | 3 | 0.62 | 0.44 | 0.38 | 0.39 | 0.43 | 6137 |
| Llama 3.2 3B | | | 6 | 0.62 | 0.42 | 0.38 | 0.38 | 0.41 | 6137 |
| Llama 3.2 3B | | | 9 | 0.62 | 0.48 | 0.44 | 0.45 | 0.47 | 6137 |
| Llama 3.2 11B | | | 0 | 0.40 | 0.23 | 0.15 | 0.14 | 0.20 | 6137 |
| Llama 3.2 11B | | | 3 | 0.80 | 0.62 | 0.61 | 0.65 | 0.64 | 6137 |
| Llama 3.2 11B | | | 6 | 0.80 | 0.65 | 0.64 | 0.67 | 0.66 | 6137 |
| Llama 3.2 11B | | | 9 | 0.80 | 0.67 | 0.66 | 0.69 | 0.68 | 6137 |
| Llama 3.2 90B | | | 0 | 0.30 | 0.27 | 0.20 | 0.18 | 0.24 | 6137 |
| Llama 3.2 90B | | | 3 | 0.91 | 0.80 | 0.84 | 0.84 | 0.80 | 6137 |
| Llama 3.2 90B | | | 6 | 0.91 | 0.80 | 0.84 | 0.84 | 0.81 | 6137 |
| Llama 3.2 90B | | | 9 | 0.90 | 0.81 | 0.84 | 0.84 | 0.81 | 6137 |
| Gemini 2.0 Flash Lite | | | 0 | 0.34 | 0.28 | 0.25 | 0.24 | 0.25 | 6137 |
| Gemini 2.0 Flash Lite | | | 3 | 0.90 | 0.82 | 0.86 | 0.86 | 0.82 | 6137 |
| Gemini 2.0 Flash Lite | | | 6 | 0.91 | 0.83 | 0.87 | 0.87 | 0.83 | 6137 |
| Gemini 2.0 Flash Lite | | | 9 | 0.91 | 0.83 | 0.87 | 0.87 | 0.83 | 6137 |
| Gemini 2.0 Flash | | | 0 | 0.38 | 0.30 | 0.32 | 0.35 | 0.32 | 6137 |
| Gemini 2.0 Flash | | | 3 | 0.94 | 0.84 | 0.88 | 0.89 | 0.84 | 6137 |
| Gemini 2.0 Flash | | | 6 | 0.94 | 0.84 | 0.88 | 0.89 | 0.84 | 6137 |
| Gemini 2.0 Flash | | | 9 | 0.94 | 0.84 | 0.88 | 0.89 | 0.84 | 6137 |
| Gemini 2.0 Flash | \checkmark | | 0 | 0.39 | 0.25 | 0.27 | 0.31 | 0.29 | 6137 |
| Gemini 2.0 Flash | \checkmark | | 6 | 0.93 | 0.82 | 0.87 | 0.88 | 0.83 | 6137 |
| Gemini 2.0 Flash Thinking | | \checkmark | 0 | 0.37 | 0.30 | 0.29 | 0.31 | 0.32 | 6137 |
| Gemini 2.0 Flash Thinking | | \checkmark | 6 | 0.94 | 0.83 | 0.88 | 0.88 | 0.84 | 6137 |
| Gemini 2.0 Flash Thinking | \checkmark | \checkmark | 0 | 0.38 | 0.28 | 0.25 | 0.27 | 0.29 | 6137 |
| Gemini 2.0 Flash Thinking | \checkmark | \checkmark | 6 | 0.95 | 0.83 | 0.88 | 0.89 | 0.84 | 6137 |
| Gemini 2.0 Pro | | | 0 | 0.38 | 0.29 | 0.25 | 0.27 | 0.29 | 6137 |
| Gemini 2.0 Pro | | | 3 | 0.86 | 0.77 | 0.80 | 0.81 | 0.77 | 6137 |
| Gemini 2.0 Pro | | | 6 | 0.85 | 0.78 | 0.80 | 0.81 | 0.78 | 6137 |
| Gemini 2.0 Pro | | | 9 | 0.86 | 0.79 | 0.82 | 0.82 | 0.79 | 6137 |
| GPT-4o mini | | | 0 | 0.30 | 0.27 | 0.23 | 0.25 | 0.27 | 6137 |
| GPT-4o mini | | | 3 | 0.93 | 0.83 | 0.88 | 0.88 | 0.85 | 6137 |
| GPT-4o mini | | | 6 | 0.94 | 0.85 | 0.89 | 0.89 | 0.86 | 6137 |
| GPT-4o mini | | | 9 | 0.94 | 0.85 | 0.89 | 0.89 | 0.86 | 6137 |
| GPT-4o | | | 0 | 0.41 | 0.25 | 0.26 | 0.29 | 0.27 | 6137 |
| GPT-4o | | | 3 | 0.96 | 0.83 | 0.89 | 0.89 | 0.84 | 6137 |
| GPT-4o | | | 6 | 0.97 | 0.85 | 0.90 | 0.90 | 0.86 | 6137 |
| GPT-4o | | | 9 | 0.97 | 0.85 | 0.90 | 0.91 | 0.86 | 6137 |
| GPT-4o Search | \checkmark | | 0 | 0.87 | 0.74 | 0.76 | 0.77 | 0.77 | 6137 |
| GPT-4o Search | \checkmark | | 6 | 0.96 | 0.90 | 0.92 | 0.92 | 0.92 | 6137 |
| GPT-4o mini Search | \checkmark | | 0 | 0.77 | 0.59 | 0.57 | 0.59 | 0.61 | 6137 |
| GPT-4o mini Search | \checkmark | | 6 | 0.91 | 0.81 | 0.81 | 0.82 | 0.84 | 6137 |
| o1 | | \checkmark | 0 | 0.44 | 0.37 | 0.38 | 0.42 | 0.42 | 6137 |
| o1 | | \checkmark | 6 | 0.98 | 0.86 | 0.91 | 0.91 | 0.87 | 6137 |
| o3-mini | | \checkmark | 0 | 0.33 | 0.25 | 0.27 | 0.31 | 0.32 | 6137 |
| o3-mini | | \checkmark | 6 | 0.96 | 0.85 | 0.90 | 0.90 | 0.86 | 6137 |

Table A3: Relative performance increases from RAG implementation across fact-checking models. Columns show the retrieval setting (k), baseline F1 score ($k = 0$), Curated RAG F1 score ($k > 0$), and percentage increase from baseline to Curated RAG. A ✓ in the **Q** and **🧠** columns denote web search and reasoning capabilities, respectively. The bottom row presents the mean and standard deviation of performance increase across all tests.

| Model | Q | 🧠 | k | F1 (macro) Baseline | F1 (macro) RAG | Increase (%) |
|---------------------------|---|---|-----|---------------------|----------------|--------------|
| DeepSeek-V3 | | | 3 | 0.25 | 0.89 | 255 |
| DeepSeek-V3 | | | 6 | 0.25 | 0.89 | 258 |
| DeepSeek-V3 | | | 9 | 0.25 | 0.90 | 259 |
| DeepSeek-R1 | | ✓ | 6 | 0.31 | 0.89 | 185 |
| Llama 3.2 3B | | | 3 | 0.14 | 0.38 | 173 |
| Llama 3.2 3B | | | 6 | 0.14 | 0.38 | 173 |
| Llama 3.2 3B | | | 9 | 0.14 | 0.44 | 221 |
| Llama 3.2 11B | | | 3 | 0.15 | 0.61 | 318 |
| Llama 3.2 11B | | | 6 | 0.15 | 0.64 | 336 |
| Llama 3.2 11B | | | 9 | 0.15 | 0.66 | 351 |
| Llama 3.2 90B | | | 3 | 0.20 | 0.84 | 323 |
| Llama 3.2 90B | | | 6 | 0.20 | 0.84 | 323 |
| Llama 3.2 90B | | | 9 | 0.20 | 0.84 | 324 |
| Gemini 2.0 Flash Lite | | | 3 | 0.25 | 0.86 | 246 |
| Gemini 2.0 Flash Lite | | | 6 | 0.25 | 0.87 | 250 |
| Gemini 2.0 Flash Lite | | | 9 | 0.25 | 0.87 | 250 |
| Gemini 2.0 Flash | | | 3 | 0.32 | 0.88 | 179 |
| Gemini 2.0 Flash | | | 6 | 0.32 | 0.88 | 179 |
| Gemini 2.0 Flash | | | 9 | 0.32 | 0.88 | 179 |
| Gemini 2.0 Flash | ✓ | | 6 | 0.27 | 0.87 | 218 |
| Gemini 2.0 Flash Thinking | | ✓ | 6 | 0.29 | 0.88 | 207 |
| Gemini 2.0 Flash Thinking | ✓ | ✓ | 6 | 0.25 | 0.88 | 258 |
| Gemini 2.0 Pro | | | 3 | 0.25 | 0.80 | 218 |
| Gemini 2.0 Pro | | | 6 | 0.25 | 0.80 | 217 |
| Gemini 2.0 Pro | | | 9 | 0.25 | 0.82 | 223 |
| GPT-4o mini | | | 3 | 0.23 | 0.88 | 288 |
| GPT-4o mini | | | 6 | 0.23 | 0.89 | 294 |
| GPT-4o mini | | | 9 | 0.23 | 0.89 | 295 |
| GPT-4o | | | 3 | 0.26 | 0.89 | 237 |
| GPT-4o | | | 6 | 0.26 | 0.90 | 241 |
| GPT-4o | | | 9 | 0.26 | 0.90 | 242 |
| GPT-4o Search | ✓ | | 6 | 0.76 | 0.92 | 21 |
| GPT-4o mini Search | ✓ | | 6 | 0.57 | 0.81 | 42 |
| o1 | | ✓ | 6 | 0.38 | 0.91 | 142 |
| o3-mini | | ✓ | 6 | 0.27 | 0.90 | 239 |
| Mean (SD) | | | | | | 233.2 (72.8) |

Table A4: By-label fact-checking performance for the best- and worst-performing models (based on macro F1). Columns show web search capability (\mathcal{Q}), retrieval setting (k), PolitiFact label, precision, recall, F1, and support (number of claims).

| Model | \mathcal{Q} | k | Label | P | R | F1 | Sup. |
|---------------|---------------|-----|---------------|------|------|------|------|
| GPT-4o Search | ✓ | 6 | True | 0.96 | 0.95 | 0.95 | 639 |
| GPT-4o Search | ✓ | 6 | Mostly true | 0.98 | 0.96 | 0.97 | 910 |
| GPT-4o Search | ✓ | 6 | Half true | 0.97 | 0.97 | 0.97 | 907 |
| GPT-4o Search | ✓ | 6 | Mostly false | 0.99 | 0.93 | 0.96 | 929 |
| GPT-4o Search | ✓ | 6 | False | 0.85 | 0.98 | 0.91 | 1931 |
| GPT-4o Search | ✓ | 6 | Pants on fire | 1.00 | 0.64 | 0.78 | 821 |
| Llama 3.2 3B | | 0 | True | 1.00 | 0.00 | 0.00 | 639 |
| Llama 3.2 3B | | 0 | Mostly true | 0.24 | 0.28 | 0.26 | 910 |
| Llama 3.2 3B | | 0 | Half true | 0.20 | 0.00 | 0.00 | 907 |
| Llama 3.2 3B | | 0 | Mostly false | 0.17 | 0.78 | 0.27 | 929 |
| Llama 3.2 3B | | 0 | False | 0.45 | 0.12 | 0.18 | 1931 |
| Llama 3.2 3B | | 0 | Pants on fire | 0.50 | 0.06 | 0.11 | 821 |

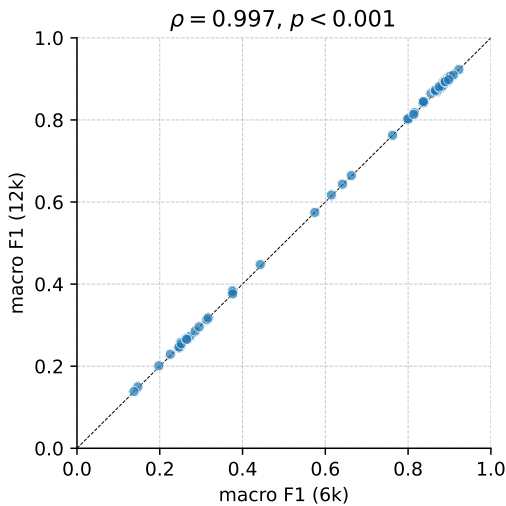


Figure A2: Comparison of fact-checking macro F1 scores on the 6k and 12k subsets. Each dot represents a standard LLM across all test scenarios. The Spearman’s correlation (ρ) and level of significance are included above.

We assessed whether models appropriately signal uncertainty by analyzing their use of NEI responses as a binary classification problem. Since no objective ground truth exists for when models should abstain, we operationalized “sufficient information” as follows: claims whose matching article summary appeared in the top- k retrieval results were labeled as having sufficient evidence, while those without were labeled as insufficient. We then considered a model correct when it provided a response with the NEI label for cases with insufficient evidence (true positive) or provided a veracity judgment when sufficient evidence was available (true negative). Conversely, responses were incorrect when models assigned veracity la-

bels despite insufficient evidence (false positive) or used NEI when sufficient evidence was present (false negative). This formulates a binary classification task, for which we computed precision, recall, and F1 scores to measure how reliably models abstain when evidence is lacking.

Figure A4(a) reports precision, recall, F1, and accuracy for NEI across models and settings. Although accuracy is high, precision, recall, and F1 are uniformly low. Figure A4(b) explains why: retrieval succeeds for 97% of claims, creating substantial class imbalance. Most test cases are negative (sufficient information present), while positive cases are rare. As a result, models achieve high accuracy by defaulting to veracity labels for about 90% of claims, even though they often fail to abstain when they should.

Against our test set, models rarely face situations that require the NEI label, and when relevant evidence is available, they typically give accurate judgments (Figures 2, 3 in main text). However, the low precision and recall scores expose a critical limitation: when uncertainty should be signaled, models perform poorly.

Figure A4(c) shows how often each model uses the NEI label overall, as well as how this varies by claim veracity. As expected, NEI is used rarely—typically on fewer than 10% of claims. Yet, the frequency of NEI responses nearly doubles for “Pants on Fire” claims. Figure A4(d) helps explain this pattern: average top- k retrieval accuracy (across k values) is substantially lower for Pants on Fire claims. This suggests that models may be responsive to signals from failed retrieval. Taken together with their low performance on the NEI task, this highlights an important area for improving how

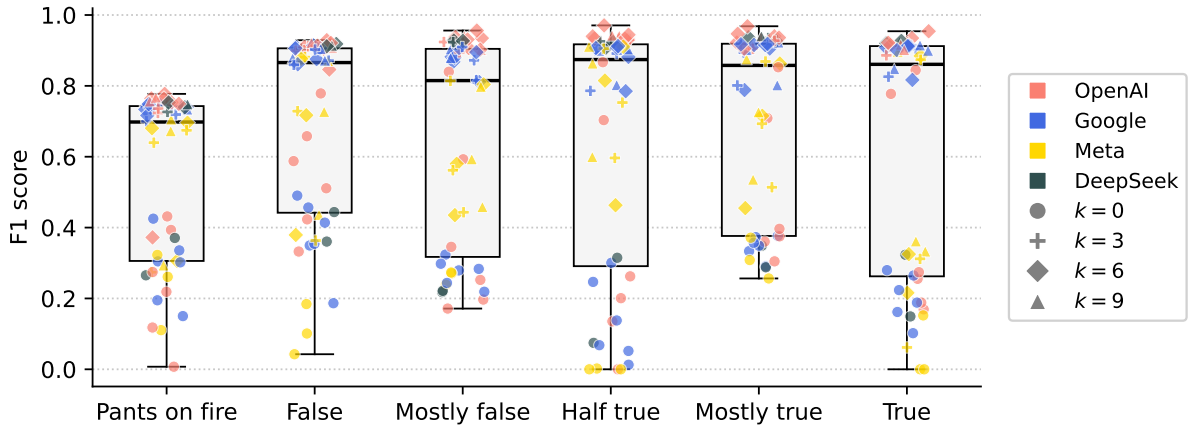


Figure A3: Boxplot distributions of F1 scores for individual PolitiFact verdicts across all tested configurations. The color and shape of each data point represents the model family and number of retrieved articles (k), respectively.

systems detect and act on such signals in order to communicate uncertainty more reliably.

10 Search Performance Conditional on Citation Behavior

Table A5 reports macro F1 scores for search-enabled GPT models across three conditions: responses in which any citation matches the correct PolitiFact article, responses with one or more citations that do not match the original fact-checking article, and responses with no URLs cited. As discussed in the main text, citing irrelevant sources is generally more damaging to performance than citing no sources at all, though both effects are substantially reduced when a curated RAG context is provided.

11 Citation Domain Classification

We classify domains cited by LLMs into seven categories using a hierarchical rule-based system detailed in Table A6. The classification proceeds in order of priority: fact-checking sites are identified first, followed by news organizations, government domains, Wiki-style sites, educational institutions, research organizations, and finally a catch-all “other” category.

12 Performance with Coarser Veracity Categories

We report macro F1 performance under three veracity classifications: binary, ternary, and the full PolitiFact scale. For the binary scheme, labels are mapped to True (True, Mostly True, Half True) or False (Mostly False, False, Pants on Fire). For the

ternary scheme, labels are mapped to True (True, Mostly True), Mixed (Half True, Mostly False), or False (False, Pants on Fire).

Figure A5 illustrates that performance improves as the number of categories decreases, as expected. However, two patterns remain consistent: (1) the curated RAG condition outperforms zero-shot and search-only settings across all classifications; and (2) finer-grained distinctions remain important for meaningful fact-checking.

13 Prompt Robustness

To assess whether our results depend on the relatively structured prompt used in the main experiments, we conducted a robustness analysis on a random sample of 1,000 claims from the 6k set. We evaluated GPT-4o and GPT-4o mini (with and without search, and with and without curated RAG context) using a more minimal prompt: “Please fact check this claim.” This version omits PolitiFact’s Truth Scale definitions, numbered instructions, and other scaffolding. The RAG condition adds only: “Leverage the fact-checking information provided below,” with retrieved items inserted directly below the claim. We also leverage OpenAI’s structured response feature, as in our main experiments. We focus on these models because they were the strongest performers in the standard and search conditions.

Table A7 indicates that performance is highly similar across these prompting styles, with mean changes of ≤ 0.03 macro F1 when averaged across models within each retrieval setting. Retrieval augmentation continues to yield substantial improvements in macro F1, given the highly similar perfor-

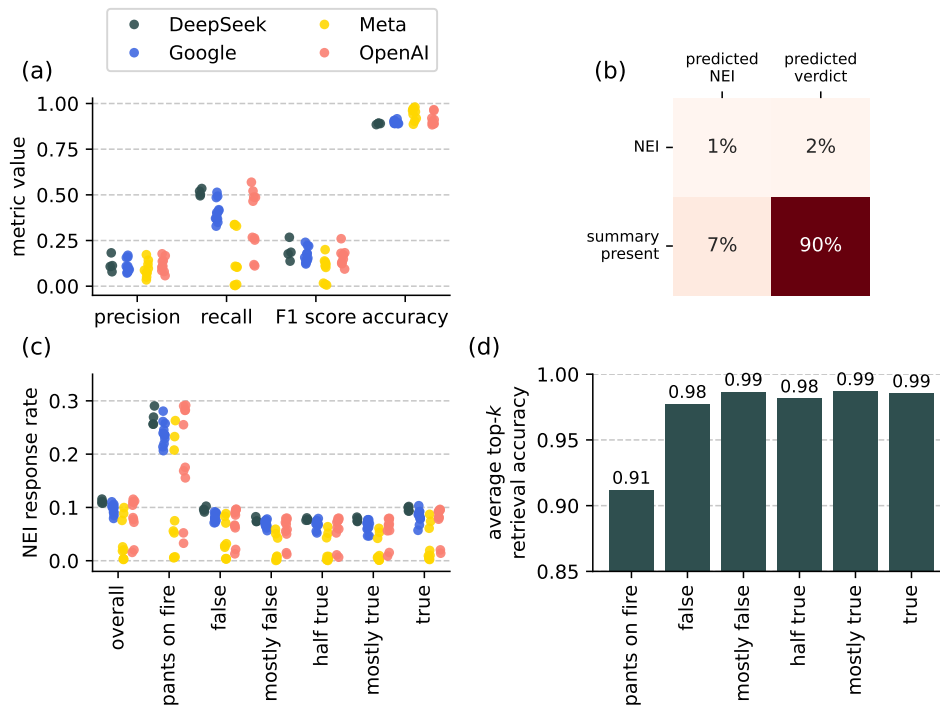


Figure A4: Models correctly refrain from using the “Not Enough Information” (NEI) label in most cases, but when they do use it, they frequently apply it incorrectly. (a) Precision, recall, F1, and accuracy for NEI predictions across all models and retrieval settings. (b) Average confusion matrix across tests. (c) NEI usage rates by veracity label. (d) Top- k retrieval accuracy by veracity. Dots in (a,c) represent all Curated RAG tests.

Table A5: Fact-checking model performance (macro F1) in search models, depending on whether the search model returned the correct URL, another URL, or no URL at all.

| Model | k | Correct Article F1 | Other URL F1 | No URL F1 |
|--------------------|-----|--------------------|--------------|-----------|
| GPT-4o Search | 0 | 0.90 | 0.50 | 0.68 |
| GPT-4o Search | 6 | 0.96 | 0.80 | 0.89 |
| GPT-4o mini Search | 0 | 0.76 | 0.40 | 0.24 |
| GPT-4o mini Search | 6 | 0.86 | 0.69 | 0.77 |

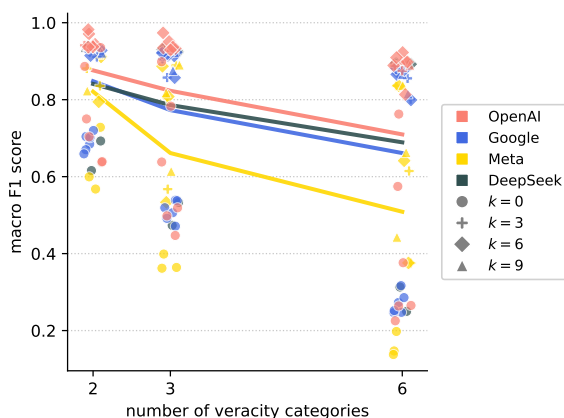


Figure A5: Macro F1 by model family as the number of veracity categories increases. Points represent individual model/configuration results, colored by family with markers indicating RAG k -value. Lines connect per-family means.

mance.

These findings increase our confidence that the core conclusions are not an artifact of prompt engineering.

14 Domain Reliability Sensitivity Analysis

We assess whether our findings regarding domain reliability and political leaning depend on NewsGuard’s proprietary reliability scores. To do so, we reanalyze the data using an alternative set of domain quality scores from Lin et al. (2023). In that work, the authors use an ensemble approach to construct aggregate reliability scores for over 11k domains by combining multiple well-studied domain-quality lists along a single principal component of reliability ratings.

Figure A6 presents the results of this analysis and reveals patterns that are largely consistent with

Table A6: Domain classification patterns and examples.

| Category | Key Patterns and Examples |
|---------------|--|
| Fact-checking | Specific organizations: snopes, politifact, factcheck.org, truthorfiction, check-yourfact |
| News | Curated list plus patterns: cnn.com, pbs.org, reuters.com, news., wusf.org |
| Government | Official extensions (.gov, .mil), federal agencies (whitehouse.gov, senate.gov, cdc.gov), international (.gov.uk, .gov.ca) |
| Wiki | Wiki variants: wiki (anywhere in domain) |
| Educational | Academic extensions (.edu, .ac.uk), institutional terms (university, college, .k12.) |
| Research | Think tanks (brookings.edu, heritage.org, rand.org), policy centers (pewresearch.org, ballotpedia.org), international organizations (oecd.org, un.org) |
| Other | All domains not matching above patterns |

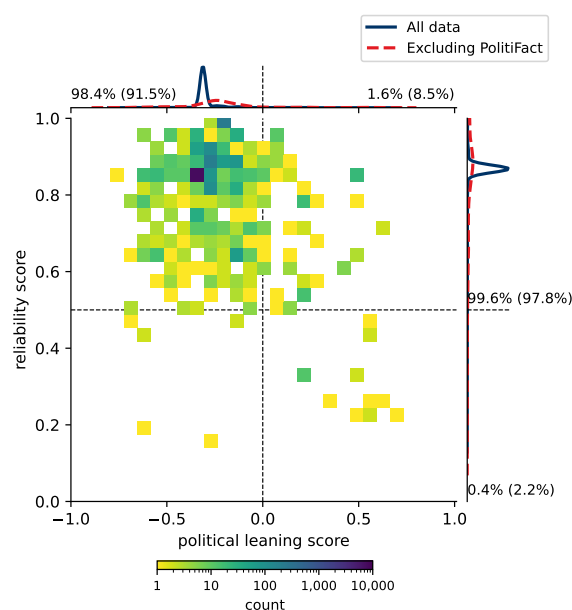


Figure A6: Reanalysis of reliability and political leaning patterns leveraging Lin et al. (2023) reliability scores. Joint distribution of reliability and political leaning scores for sources cited by search-enhanced GPT models. Marginal distributions are shown in the top and right panels for all citations (blue) and for citations excluding [politifact.com](https://www.politifact.com) (red). Black dashed lines separate different regions, and annotated percentages indicate the share of sources falling above or below each line (0.5 as the threshold for the domain-quality scores); values in parentheses report the same percentages with PolitiFact excluded.

those reported in the main text.

15 Temporal Analysis

Figure A7 shows model performance by fact-check year. The results suggest that performance is largely stable over time, with information retrieval playing a much greater role than claim date.

16 Details of the Tested LLMs

Table A8 summarizes the tested LLMs, detailing their providers, model names, reasoning and search capabilities, sizes, release dates, and knowledge cutoff dates (year/month). Model sizes for Google and OpenAI are not publicly disclosed. For some models, initial access was limited to select users before public release; we have listed the dates when models became broadly available. We could not locate the official documentation on the knowledge cutoff dates for two Google models: `gemini-2.0-flash-thinking-exp-01-21` and `gemini-2.0-pro-exp-02-05`. The table includes information (underlined) for these models obtained from a third-party repository.¹⁵

17 Example PolitiFact Claims

The full list of PolitiFact claims can be found on [politifact.com/factchecks](https://www.politifact.com/factchecks). Table A9 shows a few examples.

¹⁵<https://github.com/Hao00Wang/llm-knowledge-cutoff-dates> (accessed on 2026-02-19)

Table A7: Macro F1 performance of our main prompt, which provided clear instructions to the model (“Instr.”), as compared with a “Simple” prompt (“Please fact check this claim.”).

| Model | $k = 0$ | | | $k = 6$ | | |
|--------------------|---------|--------|----------|---------|--------|----------|
| | Instr. | Simple | Δ | Instr. | Simple | Δ |
| GPT-4o | 0.26 | 0.26 | -0.004 | 0.90 | 0.95 | +0.05 |
| GPT-4o mini | 0.23 | 0.23 | +0.001 | 0.89 | 0.96 | +0.07 |
| GPT-4o Search | 0.76 | 0.73 | -0.04 | 0.92 | 0.93 | +0.006 |
| GPT-4o mini Search | 0.57 | 0.55 | -0.02 | 0.81 | 0.80 | -0.01 |
| Average | | | -0.01 | | | +0.03 |

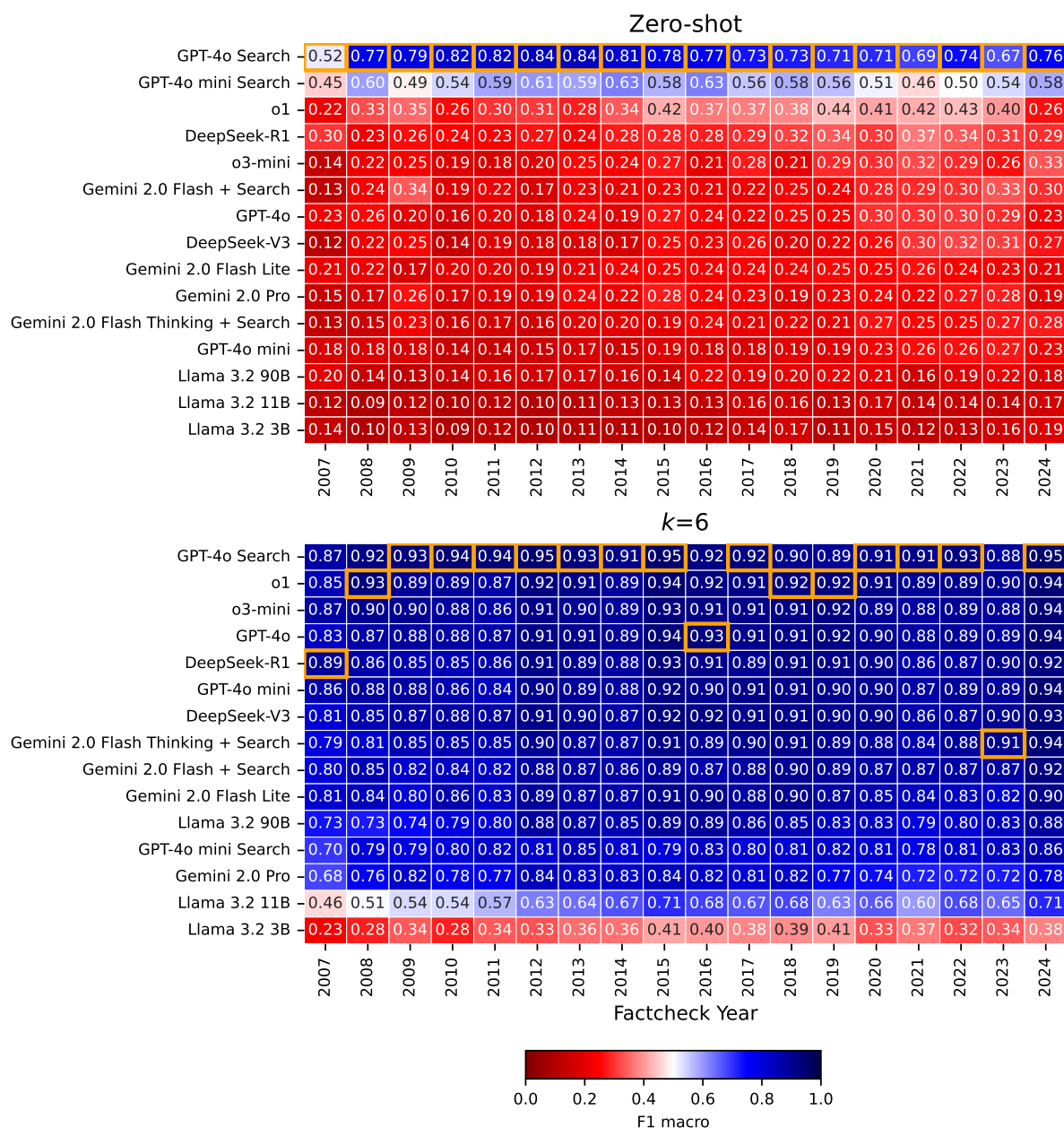


Figure A7: Macro F1 scores by fact-check year. Results are shown for all models in the zero-shot setting (top) and with Curated RAG at $k = 6$ (bottom). For Gemini models tested both with and without search (Gemini 2.0 Flash and Flash Thinking), only the search-enabled results are reported. Models are ordered by mean yearly performance, and orange boxes highlight the best performer in each year.

Table A8: Details of the tested LLMs, including their providers, models, reasoning and search capabilities, sizes, release dates, and knowledge cutoff dates (year-month). Underscored entries indicate models with uncertain knowledge cutoff dates; see the text for details.

| Provider | Model | Reasoning | Search | Size | Release | Knowledge cutoff |
|----------|---------------------------------------|-----------|--------|-------|------------|------------------|
| DeepSeek | DeepSeek-V3 | | | 671B | 2024-12-26 | 2024-07 |
| DeepSeek | DeepSeek-R1 | ✓ | | 671B | 2025-01-20 | 2024-07 |
| Meta | Llama-3.2-3B-Instruct-Turbo | | | 3.21B | 2024-09-25 | 2023-12 |
| Meta | Llama-3.2-11B-Vision-Instruct-Turbo | | | 10.6B | 2024-09-25 | 2023-12 |
| Meta | Llama-3.2-90B-Vision-Instruct-Turbo | | | 88.8B | 2024-09-25 | 2023-12 |
| Google | gemini-2.0-flash | | ✓ | N/A | 2025-02-05 | 2024-08 |
| Google | gemini-2.0-flash-lite | | | N/A | 2025-02-05 | 2024-08 |
| Google | gemini-2.0-flash-thinking-exp-01-21 | ✓ | ✓ | N/A | 2025-01-21 | <u>2024-05</u> |
| Google | gemini-2.0-pro-exp-02-05 | | | N/A | 2025-02-05 | <u>2025-01</u> |
| OpenAI | gpt-4o-2024-11-20 | | | N/A | 2024-11-20 | 2023-10 |
| OpenAI | gpt-4o-mini-2024-07-18 | | | N/A | 2024-07-18 | 2023-10 |
| OpenAI | gpt-4o-mini-search-preview-2025-03-11 | | ✓ | N/A | 2025-03-01 | 2023-10 |
| OpenAI | gpt-4o-search-preview-2025-03-11 | | ✓ | N/A | 2025-03-01 | 2023-10 |
| OpenAI | o1-2024-12-17 | ✓ | | N/A | 2024-12-17 | 2023-10 |
| OpenAI | o3-mini-2025-01-31 | ✓ | | N/A | 2025-01-31 | 2023-10 |

Table A9: Illustrative examples of PolitiFact claims, their sources, and verdicts.

| Statement | Source | Verdict |
|--|--|----------------|
| “The fact is, today abortion law in the United States is more aligned with China and North Korea than with Western nations in Europe.” | Mike Pence, Former Vice President of the United States | Mostly False |
| Every single McDonald’s french fry comes exclusively from potatoes grown on Bill Gates-owned farmland. | Facebook Posts | False |
| Government shutdowns in 2013 and 2018 “cost our economy billions of dollars each.” | Sarah Jacobs, United States Congresswoman | True |
| Video shows Elon Musk saying Oreos are “satanic.” | Viral Image | Pants on Fire! |