Contradiction to Consensus: Dual-Perspective, Multi-Source Fact Verification with Source-Level Disagreement using LLM

Anonymous ACL submission

Abstract

The rapid spread of misinformation across digital platforms poses significant societal risks. Yet most of the automated fact-checking systems depend on a single knowledge source and prioritize only supporting evidence without exposing disagreement among sources, limiting coverage and transparency. To address these limitations, we present a complete system for open-domain fact verification (ODFV) that leverages large language models (LLMs), multi-perspective evidence retrieval, and cross-source disagreement analysis. Our approach introduces a novel retrieval strategy that collects evidence for both the original and the negated forms of a claim, enabling the system to capture supporting and contradicting information from diverse sources Wikipedia, PubMed, and Google. These evidence sets are filtered, deduplicated, and aggregated across sources to form a unified and enriched knowledge base that better reflects the complexity of real-world information. This aggregated evidence is then used for veracity classification using LLMs. We further enhance interpretability by analyzing model confidence scores to quantify and visualize inter-source disagreement. Through extensive evaluation on four benchmark datasets with five LLMs, we showed that knowledge aggregation not only improves claim classification performance but also reveals differences in source-specific reasoning. Our findings underscore the importance of embracing diversity, contradiction, and aggregation in evidence for building reliable and transparent fact-checking systems. Our full code is available on GitHub [1]

1 Introduction

In an age where information travels faster than ever (Vosoughi et al., 2018), the rise of misinformation (Barve et al., 2023) and disinformation (Ghosal et al., 2020) has emerged as one of the most pressing challenges for society. With just a few clicks, misleading claims and fabricated narratives can cascade across digital platforms, shaping public opinion and, in some cases, threatening lives and livelihoods (Arcos et al., 2022). Nowhere is this danger more pronounced than in domains such as healthcare, finance, and public safety areas where trust in accurate information is huge, and where the consequences of misinformation can be threatening (Sarrouti et al., 2021; Rangapur et al., 2025; Addy, 2020).

Recognizing this urgent problem, the natural language processing (NLP) community has rallied to develop automated systems that can prevent the tide of falsehoods online. Over the past few years, researchers have proposed increasingly advanced methods for detecting and verifying claims (Augenstein et al., 2024; Eldifrawi et al., 2024; Wolfe and Mitra, 2024). Yet, despite this progress, real-world solutions remain impractical. Most current systems focus narrowly on isolated pieces of evidence, overlooking the complex reality that information is often distributed across multiple sources rather than contained within a single repository.

Moreover, the vast landscape of available knowledge is rarely leveraged to its full potential. Fact-checking systems tend to rely on a single, primary knowledge source such as Wikipedia while ignoring the wealth of information contained in secondary or tertiary sources, like scientific literature or web search results. This limitation not only constrains the system's ability to make well-rounded decisions but also hinders interpretability for end-users, who are left unaware of the underlying disagreements among sources.

These challenges raise two pivotal questions at the heart of our work: How can we build fact-checking systems that embrace, rather than ignore, the diversity and disagreement inherent in real-

[1] https://anonymous.4open.science/r/Automated-Fact-Verification-system–0BF7/

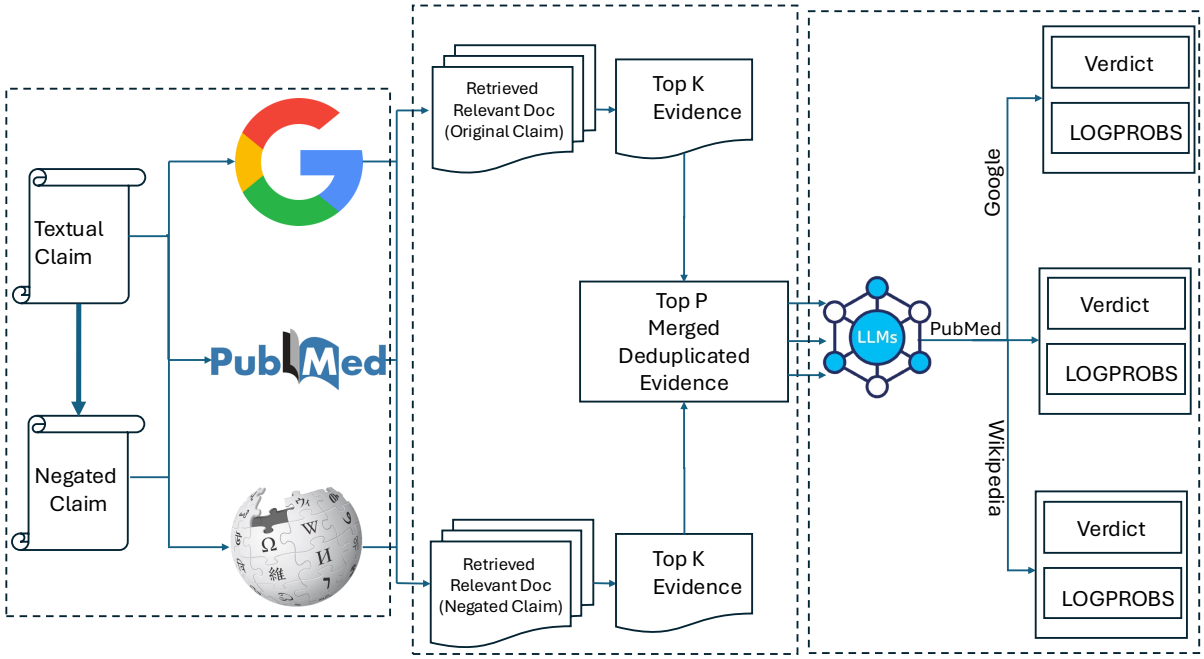

Figure 1: Workflow of the proposed system. A claim and its negation are used to retrieve evidence from Wikipedia, PubMed, and Google; candidate sentences are ranked and deduplicated, keeping the top-p per source. The per-source sets are aggregated into a compact evidence set E_i, which a zero-shot LLM uses to predict the verdict; per-source log-probabilities are reported and visualized to convey agreement and uncertainty.

world evidence? And how can we transparently convey this complexity to users, empowering them to make informed decisions based on a full spectrum of perspectives? This study addresses these questions directly by introducing a novel system dedicated to open-domain fact verification. It effectively captures, measures, and conveys the complexity and contentious nature of knowledge in the digital era. Our methodology excels in retrieving context-specific evidence while also maintaining a diversity of viewpoints by including both corroborative and opposing information. Our main contributions include:

- We introduce a novel document retrieval approach that considers both supporting and opposing evidence of a claim, preserving contrasting and corroborating information. Our method's effectiveness is shown through evaluations on four benchmark datasets, using evidence from three major knowledge bases.
- We demonstrate the effectiveness of aggregating evidence from multiple knowledge sources, as opposed to relying on a single source, for claim verification, mirroring the human approach of consulting diverse sources to assess the veracity of information.
- Finally, we showed how the verdict for each claim can vary across different knowledge sources. We quantified the level of disagreement among them, even when the sources reach the same verdict, using confidence scores.

To summarize, we present a new open-domain fact-verification pipeline that gathers evidence from multiple angles. For each claim, we retrieve documents that both support it and challenge it by also searching for a negated or opposing formulation, then filter and deduplicate the combined results. We broaden coverage by merging multiple knowledge sources beyond a single primary repository. Five state-of-the-art zero-shot large language models reason over this evidence and provide confidence signals that we aggregate to quantify how strongly the sources disagree. The result is better document recall, stronger claim-classification performance, and an interpretable view of where knowledge converges or conflicts so users can trust the outcome.

2 Related Work

Fact verification has received increasing attention in the computational linguistics and AI com-

munities, driven by the urgent need to address the proliferation of online misinformation (Patwa et al., 2022; Guo et al., 2022). Open-domain fact verification (ODFV) has matured through significant advances in evidence retrieval, model reasoning, and the integration of diverse knowledge sources (Dmonte et al., 2024). While remarkable progress has been made, important challenges remain regarding evidence selection, handling contradictory information, and ensuring interpretability.

2.1 Dual-Perspective Evidence Retrieval

Most ODFV pipelines retrieve only evidence supporting the claim, risking confirmation bias (Zhou et al., 2019; Hanselowski et al., 2018; Wang et al., 2017; Jiang et al., 2020). CONFLICTINGQA shows that retrieval-augmented LMs often prioritize relevance over stylistic or credibility cues, diverging from human judgments (Wan et al., 2024). While HCI and argumentation research study how people handle conflicting information (Fogg et al., 2003; Kakol et al., 2017; Gretz et al., 2020; Toledo et al., 2019), how AI systems reconcile contradictions remains underexplored (Wan et al., 2024). Thorne et al. (Thorne and Vlachos, 2018) advocate considering both supporting and refuting evidence, which improves transparency but complicates reconciliation and verdict aggregation. Samarinas et al. (Samarinas et al., 2021) increase explainability by distinguishing support vs. refute, yet retrieve only from the original claim. We extend this paradigm by retrieving with both the original claim and its explicit negation, forming a dual-perspective evidence pool that better captures complementary support and contradiction for ranking and verdicts.

2.2 Multi-Source Evidence Aggregation

The choice of knowledge source and retrieval method strongly affects ODFV performance. (Vladika and Matthes, 2024) shows that both source (PubMed, Wikipedia, Google) and retrieval technique (BM25 vs. semantic search) significantly impact accuracy PubMed is strongest for specialized biomedical claims, while Wikipedia better serves everyday health queries. Other work largely adopts single-source pipelines: (Santos and Pardo, 2020) uses a Portuguese Wikipedia based knowledge graph and Google snippets; many systems rely on Wikipedia and cast verification as NLI (Nie et al., 2019; Si et al., 2021; Thorne et al., 2018; Yoneda et al., 2018; Ma et al., 2019); DrQA uses Wikipedia exclusively for open-domain QA (Chen et al., 2017); MultiFC retrieves web evidence via a search API (Augenstein et al., 2019); and Cao et al. (Cao et al., 2024) incorporate external multimodal signals with a heterogeneous graph. Despite this progress, there has been little to no systematic study of aggregating evidence across multiple independent sources for a single claim. We address this gap with a multi-source pipeline that retrieves from Wikipedia, PubMed, and Google, then deduplicates, merges, and ranks sentences to form a unified evidence set per claim, capturing support and contradiction that any single source may miss.

2.3 Measuring Disagreement Across Evidence Sources

Disagreement across sources is both common and informative in ODFV. Prior work quantifies uncertainty via inter-source (or annotator) agreement (Kavtaradze, 2024), models ambiguity and annotator disagreement with soft labels (AMBIFC) (Glockner et al., 2024), promotes representational diversity via disagreement regularization in attention (Li et al., 2018), and argues for preserving divergent judgments rather than collapsing to majority vote (Leonardelli et al., 2023). Token-level uncertainty methods such as CCP further isolate uncertainty tied to factual content (Fadeeva et al., 2024). Yet, most systems still do not explicitly expose source-level disagreement to users. In our approach, for each claim, we compute per-source confidence scores (log-probabilities) for the predicted label, then quantify dispersion across sources. Low dissemination indicates agreement; high dissemination flags disagreement and potential uncertainty. We visualize these per-source logprobs to show confident agreements and ambiguous cases, enabling transparent, multi-source verdict interpretation to the end user.

3 Automated Fact-Checking Pipeline

Our comprehensive methodology for Open-Domain Fact Verification (ODFV) is designed to systematically generate, retrieve, select, and evaluate evidence to accurately classify textual claims. The figure 1 describes the whole architecture of our pipeline. The methodology consists of three key components: (1) generation of negated claims, (2) evidence retrieval and selection, and (3) claim verification using Large Language Models (LLMs). We describe each component in detail in the fol-

Dataset	Model	Knowledge Source	Original Claim				Original + Negated Claim			
			A	P	R	F1	A	P	R	F1
SCIFACT	Llama 70B	Wikipedia	0.430	0.468	0.407	0.415	0.230	0.376	0.335	0.203
		Pubmed	0.597	0.588	0.597	0.584	0.617	0.609	0.626	0.605
		Google	0.550	0.543	0.558	0.530	0.607	0.615	0.620	0.573
	Llama 405B	Wikipedia	0.447	0.489	0.420	0.425	0.443	0.475	0.421	0.427
		Pubmed	0.593	0.583	0.592	0.576	0.617	0.606	0.622	0.599
		Google	0.580	0.573	0.577	0.550	0.597	0.591	0.602	0.562
	Phi-4	Wikipedia	0.410	0.474	0.391	0.379	0.413	0.473	0.401	0.400
		Pubmed	0.583	0.574	0.583	0.578	0.587	0.573	0.599	0.579
		Google	0.590	0.579	0.583	0.574	0.593	0.584	0.608	0.568
	Qwen 2.5	Wikipedia	0.437	0.504	0.410	0.401	0.423	0.478	0.403	0.398
		Pubmed	0.583	0.573	0.575	0.573	0.597	0.588	0.603	0.591
		Google	0.587	0.578	0.577	0.565	0.590	0.567	0.595	0.563
	Mistral	Wikipedia	0.393	0.447	0.368	0.358	0.393	0.438	0.375	0.368
		Pubmed	0.573	0.565	0.571	0.566	0.590	0.585	0.595	0.585
		Google	0.583	0.576	0.582	0.568	0.603	0.591	0.620	0.586
Averitec	Llama 70B	Wikipedia	0.259	0.417	0.355	0.229	0.230	0.376	0.335	0.203
		Pubmed	0.183	0.444	0.298	0.163	0.196	0.438	0.307	0.176
		Google	0.375	0.351	0.354	0.288	0.383	0.384	0.387	0.311
	Llama 405B	Wikipedia	0.379	0.367	0.372	0.278	0.340	0.376	0.360	0.260
		Pubmed	0.267	0.394	0.330	0.221	0.273	0.404	0.329	0.229
		Google	0.434	0.383	0.395	0.321	0.444	0.363	0.396	0.317
	Phi-4	Wikipedia	0.424	0.416	0.402	0.306	0.463	0.438	0.408	0.325
		Pubmed	0.308	0.500	0.333	0.230	0.326	0.553	0.330	0.242
		Google	0.453	0.385	0.376	0.334	0.495	0.389	0.402	0.360
	Qwen 2.5	Wikipedia	0.238	0.552	0.347	0.228	0.214	0.433	0.321	0.193
		Pubmed	0.153	0.404	0.283	0.119	0.173	0.559	0.295	0.141
		Google	0.358	0.393	0.383	0.301	0.389	0.407	0.413	0.329
	Mistral	Wikipedia	0.320	0.530	0.379	0.281	0.322	0.516	0.382	0.283
		Pubmed	0.196	0.476	0.298	0.154	0.202	0.474	0.295	0.160
		Google	0.399	0.394	0.374	0.331	0.440	0.409	0.401	0.359
Liar	Llama 70B	Wikipedia	0.210	0.234	0.185	0.114	0.219	0.573	0.196	0.130
		Pubmed	0.207	0.353	0.183	0.100	0.200	0.319	0.175	0.087
		Google	0.394	0.657	0.388	0.382	0.402	0.659	0.397	0.393
	Llama 405B	Wikipedia	0.243	0.241	0.235	0.207	0.256	0.245	0.247	0.217
		Pubmed	0.222	0.212	0.219	0.170	0.203	0.188	0.198	0.143
		Google	0.396	0.514	0.398	0.401	0.415	0.541	0.413	0.419
	Phi-4	Wikipedia	0.226	0.237	0.212	0.164	0.230	0.258	0.212	0.167
		Pubmed	0.202	0.208	0.183	0.108	0.200	0.193	0.183	0.106
		Google	0.387	0.456	0.391	0.386	0.385	0.454	0.390	0.387
	Qwen 2.5	Wikipedia	0.210	0.334	0.184	0.096	0.211	0.323	0.185	0.097
		Pubmed	0.201	0.090	0.174	0.073	0.200	0.085	0.172	0.071
		Google	0.402	0.737	0.391	0.382	0.408	0.728	0.397	0.389

Table 1: Original-only vs. original+negated evidence in zero-shot evaluation. We report Accuracy (A), Precision (P), Recall (R), and macro-F_1 across datasets, models, and knowledge sources (Wikipedia, PubMed, Google).

Dataset	Model	Knowledge Source	Original Claim				Original + Negated Claim			
			A	P	R	F1	A	P	R	F1
Liar	Mistral	Wikipedia	0.210	0.334	0.184	0.096	0.211	0.323	0.185	0.097
		Pubmed	0.214	0.134	0.184	0.105	0.206	0.116	0.176	0.095
		Google	0.375	0.608	0.361	0.356	0.381	0.631	0.365	0.362
PubHealth	Llama 70B	Wikipedia	0.245	0.333	0.313	0.197	0.241	0.325	0.318	0.195
		Pubmed	0.199	0.313	0.315	0.163	0.207	0.341	0.331	0.169
		Google	0.451	0.377	0.356	0.266	0.467	0.391	0.393	0.292
	Llama 405B	Wikipedia	0.317	0.319	0.346	0.240	0.353	0.480	0.386	0.270
		Pubmed	0.323	0.323	0.380	0.247	0.324	0.329	0.354	0.243
		Google	0.535	0.631	0.384	0.316	0.559	0.392	0.398	0.329
	Phi-4	Wikipedia	0.305	0.364	0.314	0.255	0.330	0.383	0.329	0.275
		Pubmed	0.261	0.307	0.283	0.213	0.284	0.328	0.288	0.230
		Google	0.492	0.421	0.355	0.336	0.517	0.413	0.365	0.356
	Qwen 2.5	Wikipedia	0.183	0.343	0.314	0.158	0.194	0.345	0.322	0.170
		Pubmed	0.165	0.303	0.308	0.139	0.176	0.322	0.304	0.151
		Google	0.463	0.473	0.387	0.281	0.476	0.391	0.386	0.295
	Mistral	Wikipedia	0.424	0.297	0.266	0.229	0.417	0.274	0.247	0.212
		Pubmed	0.425	0.262	0.223	0.184	0.440	0.302	0.239	0.198
		Google	0.496	0.349	0.304	0.278	0.488	0.341	0.283	0.258

Table 2: Extension of Table 1: Comparison of Original Claims vs. Original + Negated Claims

Dataset	Model	Knowledge Source			
		Wikipedia	Pubmed	Google	Merged(W+P+G)
SciFact	Llama 70B	0.430	0.597	0.550	0.610
	Llama 405B	0.447	0.593	0.580	0.597
	Phi-4	0.410	0.583	0.590	0.583
	Qwen 2.5	0.437	0.583	0.587	0.607
	Mistral	0.393	0.573	0.583	0.617
Averitec	Llama 70B	0.230	0.196	0.383	0.384
	Llama 405B	0.340	0.273	0.444	0.574
	Phi-4	0.463	0.326	0.495	0.515
	Qwen 2.5	0.214	0.173	0.389	0.387
	Mistral	0.322	0.202	0.440	0.521
Liar	Llama 70B	0.219	0.200	0.402	0.300
	Llama 405B	0.256	0.203	0.415	0.320
	Phi-4	0.230	0.200	0.385	0.289
	Qwen 2.5	0.211	0.200	0.408	0.285
	Mistral	0.211	0.206	0.381	0.294
PubHealth	Llama 70B	0.241	0.207	0.467	0.449
	Llama 405B	0.353	0.324	0.559	0.557
	Phi-4	0.330	0.284	0.517	0.553
	Qwen 2.5	0.194	0.176	0.476	0.467
	Mistral	0.417	0.440	0.488	0.490

Table 3: Comparative Accuracy of Individual vs. Aggregated Knowledge Sources (Wikipedia+PubMed+Google)

lowing subsections.

3.1 Datasets

We evaluate our ODFV system on four benchmarks ranging scientific, health, sociopolitical, and political claims: **SciFact** (Wadden et al., 2020) (biomedical claims with sentence-level evidence from abstracts), **PubHealth** (Kotonya and Toni, 2020) (health claims with expert justifications), **Averitec** (Schlichtkrull et al., 2023) (controversial/ambiguous sociopolitical claims emphasizing uncertainty), and **LIAR** (Wang, 2017) (large-scale political claims from speeches, social media, and news). Table 4 in the appendix summarizes domains, sources, label sets, and claim counts.

3.2 Generation of Negated Claims

For each claim set $\mathcal{C} = \{c_1, \ldots, c_n\}$, we generate a negated counterpart $\bar{c}_i$ for every c_i using Mistral AI[2], producing informative contrasts (including numerical reframing). For example: *"A deficiency of vitamin B12 increases homocysteine"* $\rightarrow$ *"A surplus of vitamin B12 decreases homocysteine"*; and *"5% of perinatal mortality is due to low birth weight"* $\rightarrow$ *"95% of perinatal mortality is not due to low birth weight"*. Pairing $(c_i, \bar{c}_i)$ ensures both supportive and contradictory perspectives for subsequent retrieval and verification.

3.3 Evidence Retrieval

We use three major knowledge sources, Wikipedia, PubMed, and Google, guided by (Vladika and Matthes, 2024), which finds Wikipedia stronger on popular/trending claims and PubMed more precise for technical/scientific queries. To ensure coverage and domain adaptability, we utilized all three.

Let $\mathcal{K} = \{k_1, k_2, k_3\}$ denote Wikipedia, PubMed, and Google. For each claim $c_i \in \mathcal{C}$ and source $k_j \in \mathcal{K}$, we retrieve the top-k documents

$$R(c_i, k_j) = \{d_1^{(i,j)}, \ldots, d_k^{(i,j)}\},$$

and do so for both c_i and its negation $\bar{c}_i$ to gather supportive, neutral, and potentially contradictory evidence.

Source-specific pipelines:

- **Wikipedia:** We used English dumps[3] ($\sim$7M articles) cleaned with WikiExtractor (Attardi, 2015) and indexed in Elasticsearch[4] for scalable retrieval.
- **PubMed:** 23.6M abstracts[5] preprocessed and encoded with transformer-based sentence embeddings for dense retrieval, enhanced by BM25 (Amati, 2009) for lexical ranking.
- **Google:** We leveraged Google Custom Search API[6] queries each claim and returns ranked results (title, snippet, URL) as web evidence.

3.4 Evidence Selection

After retrieval, we filter sentences using SPICED embeddings (Shushkevich et al., 2023). For each claim $x \in \{c, \bar{c}\}$, we embed the claim and every sentence from the top M retrieved documents and compute their cosine similarity. From each document we keep the top-k sentence(s) by similarity and unite them across documents to form the evidence set. We apply the same procedure to c and $\bar{c}$ to capture both supportive and contradictory context.

3.5 Evidence Deduplication and Final Selection

The core idea is to remove overlapping (duplicate) evidence sentences and merge the rest. For claim c_i and source k_j, let $E_{ij}^{+} = E(c_i, k_j)$ and $E_{ij}^{-} = E(\bar{c}_i, k_j)$. After normalization (lowercasing, punctuation stripping), we form candidates via symmetric difference with light merging (to fuse split segments, e.g., `[SEP]`):

$$E_{ij}^{\text{cand}} = \text{Merge}\big(\tilde{E}_{ij}^{+} \triangle \tilde{E}_{ij}^{-}\big).$$

We then rank E_{ij}^{cand} by SPICED (Shushkevich et al., 2023) similarity to c_i and keep the top-p as E_{ij}^{final}. Finally, we aggregate per-claim evidence across sources as

$$E_i = \bigcup_{j=1}^{|\mathcal{K}|} E_{ij}^{\text{final}}.$$

3.6 Veracity Prediction

For final classification, we employ a large language model (LLM), denoted L, to predict the veracity of each claim c_i given E_i. We evaluate five state-of-the-art LLMs (open or widely available): Llama 3.3 (70B) and Llama 3.1 (405B) (Grattafiori et al., 2024), Mistral-Large, Qwen 2.5 (Team, 2024),

[2] https://docs.mistral.ai/api/
[3] https://dumps.wikimedia.org/
[4] https://www.elastic.co/elasticsearch
[5] https://pubmed.ncbi.nlm.nih.gov/download/
[6] https://developers.google.com/custom-search/v1/overview

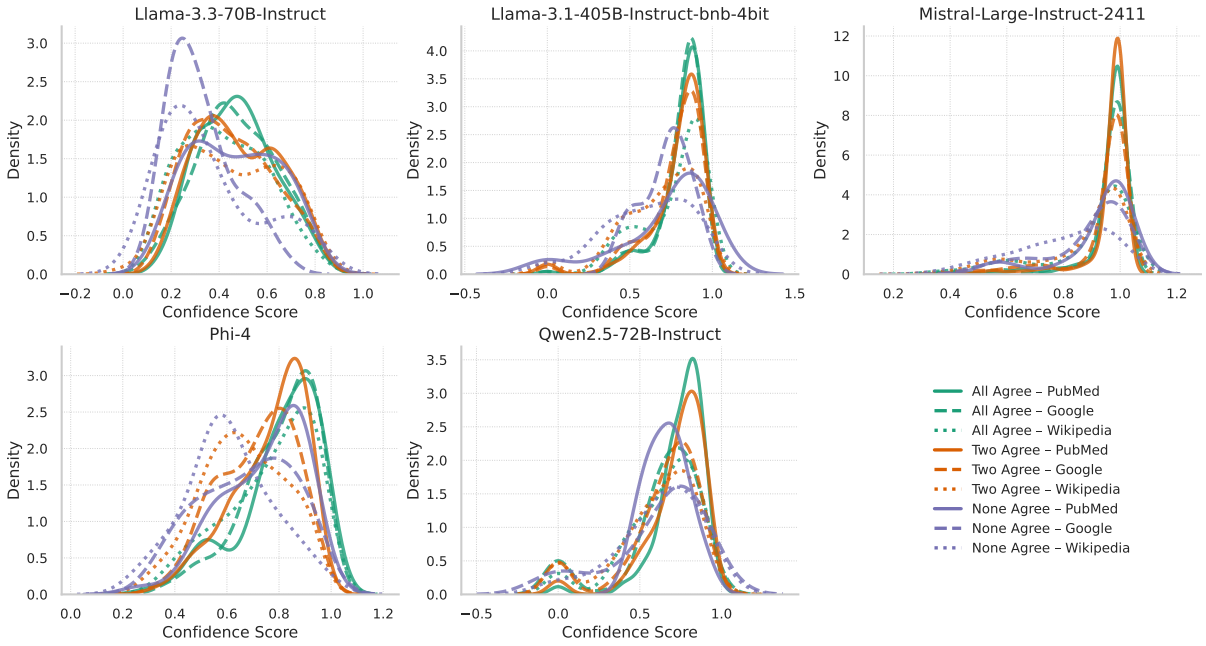

Figure 2: Confidence distribution (KDE) across different knowledge sources for the Averitec dataset, illustrating variation in model certainty and inter-source disagreement during fact verification

and Phi-4 (Abdin et al., 2024), selected for strong benchmark performance and accessibility (Fourrier et al., 2024). In a zero-shot setting, each model is prompted with the claim and its evidence using a direct fact-checking prompt (Appendix Figure 6) adapted to each model and dataset label schema.

We present veracity as m-class prediction over $Y = \{y_1, \ldots, y_m\}$ (e.g., $m = 3$: *Refuted*, *Supported*, *Not Enough Evidence*). For each claim evidence pair, the LLM outputs

$$\hat{y}_i = F(c_i, E_i; L) \in Y,$$

via a single-token choice (e.g., A–C) deterministically mapped to dataset classes.

3.7 Quantifying Source-Level Disagreement

Recent work shows that token-level log-probabilities ("logprobs") help interpret LLM decisions (Kauf et al., 2024). We treat each veracity label as a discrete token and use the logprob of the chosen label as the confidence score. For claim c_i with evidence E_i, letting $\mathbf{z}$ be the logits over labels $\mathcal{Y}$ and $\hat{y}_i = \arg\max \mathrm{softmax}(\mathbf{z})$, confidence is

$$\mathrm{conf}(c_i, E_i) = \log\big(\mathrm{softmax}(\mathbf{z})\big)_{\hat{y}_i}.$$

We compute this per source (Wikipedia, PubMed, Google) and compare the resulting logprobs across sources: low dispersion indicates agreement, while high dispersion signals disagreement and uncertainty; visualizations make these patterns explicit.

4 Results

We purposefully evaluate in a zero-shot setting (no fine-tuning) to isolate the contribution of our method rather than optimize absolute scores. Across SciFact, Averitec, LIAR, and PubHealth with five LLMs (Llama 70B, Llama 405B, Phi-4, Qwen 2.5, Mistral), two robust trends emerge (Tables 1, 2, 3).

First, augmenting each claim with its explicit negation generally improves accuracy and macro-F_1 over using the original claim alone, with typical relative gains of about +2–10% (accuracy) and +2–8% (F_1). Representative examples include SciFact with Llama 70B+Google (+10.4% accuracy, +8.1% F_1; 0.550→0.607, 0.530→0.573), Averitec with Phi-4+Wikipedia (+9.2%, +6.2%), LIAR with Llama 405B+Google (+4.8%, +4.5%), and PubHealth with Phi-4+Google (+5.1%, +6.0%). While a few model source pairs show neutral or slight decreases (e.g., SciFact with Phi-4+Google F_1), the overall effect is consistently positive.

Second, aggregating evidence from Wikipedia, PubMed, and Google typically boosts performance beyond any single source, especially relative to weaker sources: on SciFact, Llama 70B's merged F_1 exceeds Wikipedia by +41.9% and Google by

+10.9%; on Averitec, Llama 405B's merged F_1 is +68.8% over Wikipedia and +29.3% over Google; on PubHealth, Phi-4's merged F_1 is +67.6% over Wikipedia and +7.0% over Google. For LIAR, where Google alone is already strong, merged performance remains above Wikipedia and PubMed but is below Google (e.g., Llama 405B: 0.320 vs. 0.256/0.203/0.415). Collectively, dual-perspective retrieval and multi-source aggregation provide complementary, often double-digit relative gains across models and datasets in zero-shot conditions, demonstrating the robustness and practical effectiveness of the proposed approach.

4.1 Consensus and Conflict Across Sources

We visualize per-source confidence (token log-probability of the predicted label with KDEs formed by source agreement among PubMed, Google, and Wikipedia (*all*, *two*, *none*). **Averitec** (Figure 2) shows the expected ordering: unanimity yields sharper, higher-confidence peaks; partial agreement is broader and lower; and no agreement is lowest and most dispersed, though curves are flatter than in structured domains. Appendix figures for **LIAR**, **PubHealth**, and **SciFact** (Figure 3, 4, 5) confirm the same trend, with clearer separation in SciFact/PubHealth and greater dissemination in LIAR. Confidence magnitudes are sometimes model-dependent: **Llama** separates agreement regimes most distinctly, **Mistral** is similar but broader, **Phi-4** spreads more under disagreement, and **Qwen 2.5** shows tight peaks under unanimity. PubHealth violin plots (Appendix Figure 7) support inter-model shifts in central tendency and dispersion. Reporting per-source log-probabilities and their dispersion alongside the verdict thus quantifies source-level disagreement (e.g., Google agrees while Wikipedia does not) and makes residual uncertainty transparent.

5 Discussion

Dual-perspective retrieval, considering both the original claim and its negation together with aggregation across sources, yields consistent gains in our zero-shot results. Kernel Density Estimations (KDEs) of per-source log-probabilities indicate a strong correlation between agreement and certainty: complete agreement across sources like PubMed, Google, and Wikipedia results in sharp peaks of high confidence, whereas partial or no consensus produces lower and wider distributions, more so in open-domain data sets (e.g., LIAR, Averitec) than in structured ones (e.g., SciFact, PubHealth). Explicitly negating claims systematically enhances the verification process by retrieving evidence that both supports and refutes, and combining information from multiple sources further improves performance while decreasing uncertainty when sources are in agreement. Since no single source predominates, the utility of a source is dependent on the claim and domain, justifying aggregation. In practice, displaying per-source confidence and its distribution offers a model-agnostic indicator of reliability and reveals disagreements transparently for end users. However, because raw confidence levels vary between models, comparisons across models require calibration or intra-model baselines. Remaining challenges include time-sensitive evidence and handling long contexts, which may hinder certainty even when aggregation is employed.

6 Future Work

Validated in zero-shot settings, our next steps are to add some advanced and useful techniques. We will add temporal reasoning with time-aware retrieval and alignment of claims and evidence. We will extend to multilingual verification and aim for comparable performance across languages. We will develop context-aware retrieval that adapts to user context under neutrality constraints. We will improve web evidence quality by grading content and estimating source reliability to prioritize stronger evidence. We will mitigate hallucinations through faithfulness checks, confidence calibration, and enforcing consistency between evidence and verdicts.

7 Conclusions

We presented an open-domain fact verification system that retrieves with both original and negated claims to capture support and refutation, aggregates evidence from Wikipedia, PubMed, and Google, and quantifies uncertainty via label log-probabilities and KDE-based visualizations. Through a series of experiments, we showed that negated-claim retrieval and multi-source aggregation yield consistent, complementary gains without fine-tuning, improving both performance and interpretability. Llama and Mistral showed consistently strong performance across knowledge sources and datasets. By exposing source-level agreement and confidence score, the system strengthens reliability and transparency in automated fact-checking.

Limitations

Our study has several limitations. First, constrained context windows can truncate or underweight relevant passages when verification requires long, multi-document evidence; although emerging long-context models (e.g., $\geq$32K tokens) may help in this, we did not exploit them here, and hierarchical selection/ordering effects ("lost in the middle") may depress performance. Second, open-domain datasets such as *LIAR* include noisy or adversarial claims and minority labels; in zero-shot classification, this combination yields lower macro-F_1 even when accuracy is moderate, and models struggle to resolve incomplete or genuinely conflicting evidence. Third, we do not impose time-aware retrieval or reasoning, so evidence that is outdated or post-dates the claim can lead to inconsistent verdicts for time-sensitive statements. Finally, like other LLM-based systems, our approach remains sensitive to distribution shifts, misleading inputs, and biases in source corpora and pre-training, which can limit generalization and reliability in real-world settings.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Jamie M Addy. 2020. The art of the real: Fact checking as information literacy instruction. *Reference Services Review*, 48(1):19–31.

Giambattista Amati. 2009. Bm25. In *Encyclopedia of database systems*, pages 257–260. Springer.

Rubén Arcos, Manuel Gertrudix, Cristina Arribas, and Monica Cardarilli. 2022. Responses to digital disinformation as part of hybrid threats: a systematic review on the effects of disinformation and the effectiveness of fact-checking/debunking. *Open Research Europe*, 2:8.

Giusepppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):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. *arXiv preprint arXiv:1909.03242*.

Yashoda Barve, Jatinderkumar R Saini, Rutuja Rathod, and Hema Gaikwad. 2023. Multi-modal misinformation detection: An exhaustive review. In *2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pages 1–5. IEEE.

Han Cao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2024. Multi-source knowledge enhanced graph attention networks for multimodal fact verification. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2024. Claim verification in the age of large language models: A survey. *arXiv preprint arXiv:2408.14317*.

Islam Eldifrawi, Shengrui Wang, and Amine Trabelsi. 2024. Automated justification production for claim veracity in fact checking: A survey on architectures and approaches. *arXiv preprint arXiv:2407.12853*.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, et al. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. *arXiv preprint arXiv:2403.04696*.

Brian J Fogg, Cathy Soohoo, David R Danielson, Leslie Marable, Julianne Stanford, and Ellen R Tauber. 2003. How do users evaluate the credibility of web sites? a study with over 2,500 participants. In *Proceedings of the 2003 conference on Designing for user experiences*, pages 1–15.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

Soumya Suvra Ghosal, Deepak P, and Anna Jurek-Loughrey. 2020. Resco-cc: Unsupervised identification of key disinformation sentences. In *Proceedings of the 22nd International Conference on Information Integration and Web-based Applications & Services*, pages 47–54.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. Ambifc: Fact-checking ambiguous claims with evidence. *Transactions of the Association for Computational Linguistics*, 12:1–18.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813.

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

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. Ukp-athene: Multi-sentence textual entailment for claim verification. *arXiv preprint arXiv:1809.01479*.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. Hover: A dataset for many-hop fact extraction and claim verification. *arXiv preprint arXiv:2011.03088*.

Michal Kakol, Radoslaw Nielek, and Adam Wierzbicki. 2017. Understanding and predicting web content credibility using the content credibility corpus. *Information Processing & Management*, 53(5):1043–1061.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A Ivanova. 2024. Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models. *arXiv preprint arXiv:2403.14859*.

Lasha Kavtaradze. 2024. Dominant disciplinary and thematic approaches to automated fact-checking: A scoping review and reflection. *Digital Journalism*, pages 1–26.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. *arXiv preprint arXiv:2010.09926*.

Elisa Leonardelli, Alexandra Uma, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, and Massimo Poesio. 2023. Semeval-2023 task 11: Learning with disagreements (lewidi). *arXiv preprint arXiv:2304.14803*.

Jian Li, Zhaopeng Tu, Baosong Yang, Michael R Lyu, and Tong Zhang. 2018. Multi-head attention with disagreement regularization. *arXiv preprint arXiv:1810.10183*.

Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2019. Sentence-level evidence embedding for claim verification with hierarchical attention networks. Association for Computational Linguistics.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6859–6866.

Parth Patwa, Shreyash Mishra, S Suryavardan, Amrit Bhaskar, Parul Chopra, Aishwarya Reganti, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. 2022. Benchmarking multi-modal entailment for fact verification. In *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*.

Aman Rangapur, Haoran Wang, Ling Jian, and Kai Shu. 2025. Fin-fact: A benchmark dataset for multimodal financial fact-checking and explanation generation. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 785–788.

Chris Samarinas, Wynne Hsu, and Mong-Li Lee. 2021. Improving evidence retrieval for automated explainable fact-checking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 84–91.

Roney Lira de Sales Santos and Thiago Alexandre Salgueiro Pardo. 2020. Fact-checking for portuguese: Knowledge graph and google search-based methods. In *International Conference on Computational Processing of the Portuguese Language*, pages 195–205. Springer.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems*, 36:65128–65167.

Elena Shushkevich, Long Mai, Manuel V Loureiro, Steven Derby, and Tri Kurniawan Wijaya. 2023. Spiced: News similarity detection dataset with multiple topics and complexity levels. *arXiv preprint arXiv:2309.13080*.

Jiasheng Si, Deyu Zhou, Tongzhe Li, Xingyu Shi, and Yulan He. 2021. Topic-aware evidence reasoning and stance-aware aggregation for fact verification. *arXiv preprint arXiv:2106.01191*.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. *arXiv preprint arXiv:1806.07687*.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and verification (fever) shared task. *arXiv preprint arXiv:1811.10971*.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment–new datasets and methods. *arXiv preprint arXiv:1909.01007*.

Juraj Vladika and Florian Matthes. 2024. Comparing knowledge sources for open-domain scientific claim verification. *arXiv preprint arXiv:2402.02844*.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

Alexander Wan, Eric Wallace, and Dan Klein. 2024. What evidence do language models find convincing? *arXiv preprint arXiv:2402.11782*.

Shuohang Wang, Mo Yu, Jing Jiang, Wei Zhang, Xiaoxiao Guo, Shiyu Chang, Zhiguo Wang, Tim Klinger, Gerald Tesauro, and Murray Campbell. 2017. Evidence aggregation for answer re-ranking in open-domain question answering. *arXiv preprint arXiv:1711.05116*.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Robert Wolfe and Tanushree Mitra. 2024. The impact and opportunities of generative ai in fact-checking. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1531–1543.

Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Ucl machine reading group: Four factor framework for fact finding (hexaf). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 97–102.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. Gear: Graph-based evidence aggregating and reasoning for fact verification. *arXiv preprint arXiv:1908.01843*.

A Appendix

Dataset	Source	Domain	Label	Claims
SCIFACT	Science	Scientific	Supported Refuted Not Enough Info	1400
Averitec	Factcheck	Mixed	Supported Refuted Conflicting evidence/cherrypicking Not Enough Info	4568
LIAR	POLITIFACT.COM	Fake News	Pants on Fire False Barely True Half True Mostly True True	12,836
PUBHEALTH	Factcheck	Biomedical	True False Mixture Unproven	11,832

Table 4: Overview of Benchmark Datasets

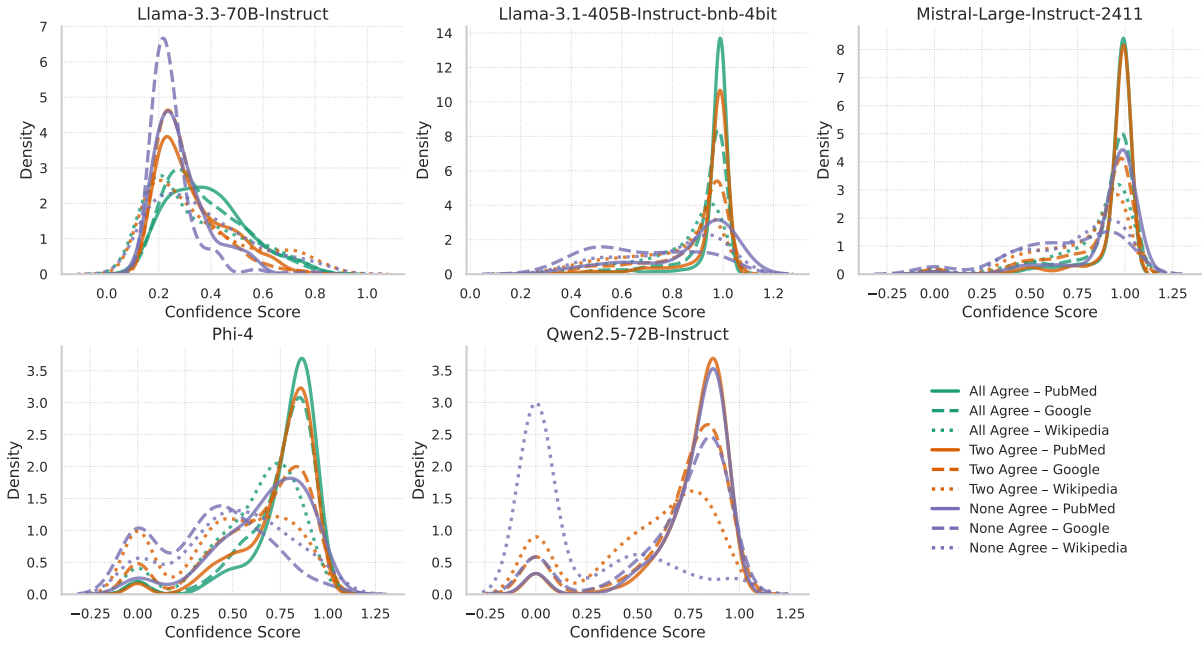

Figure 3: Confidence distribution (KDE) across different knowledge sources for the LIAR dataset, illustrating variation in model certainty and inter-source disagreement during fact verification

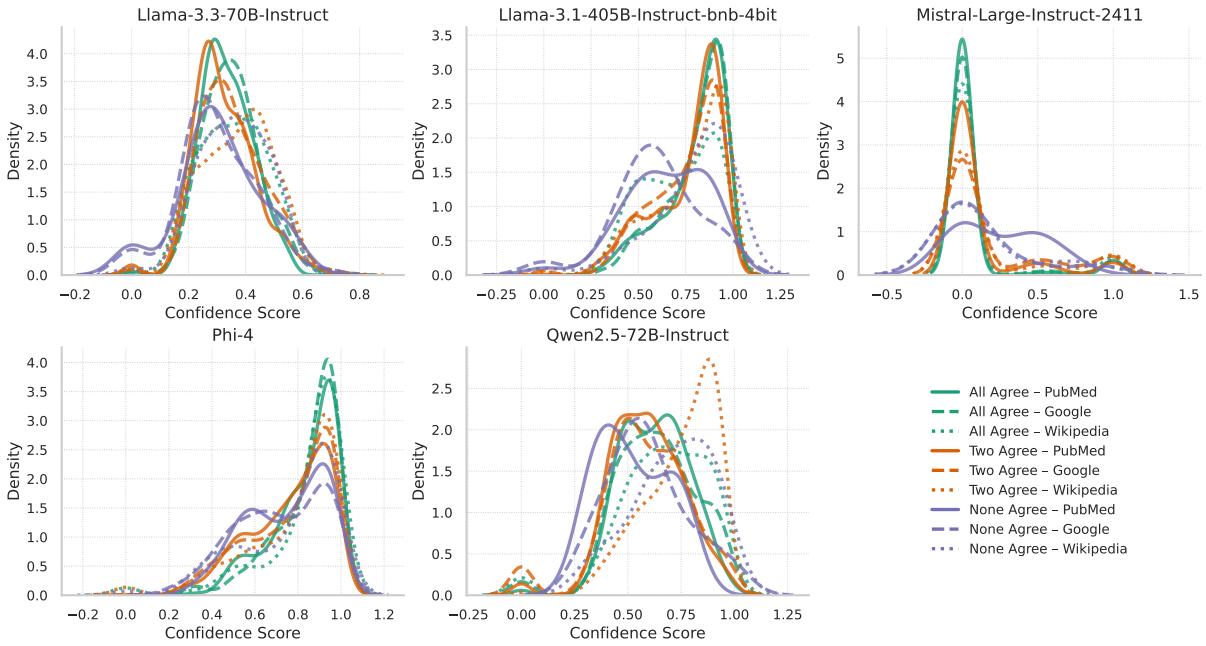

Figure 4: Confidence distribution (KDE) across different knowledge sources for the Pubhealth dataset, illustrating variation in model certainty and inter-source disagreement during fact verification

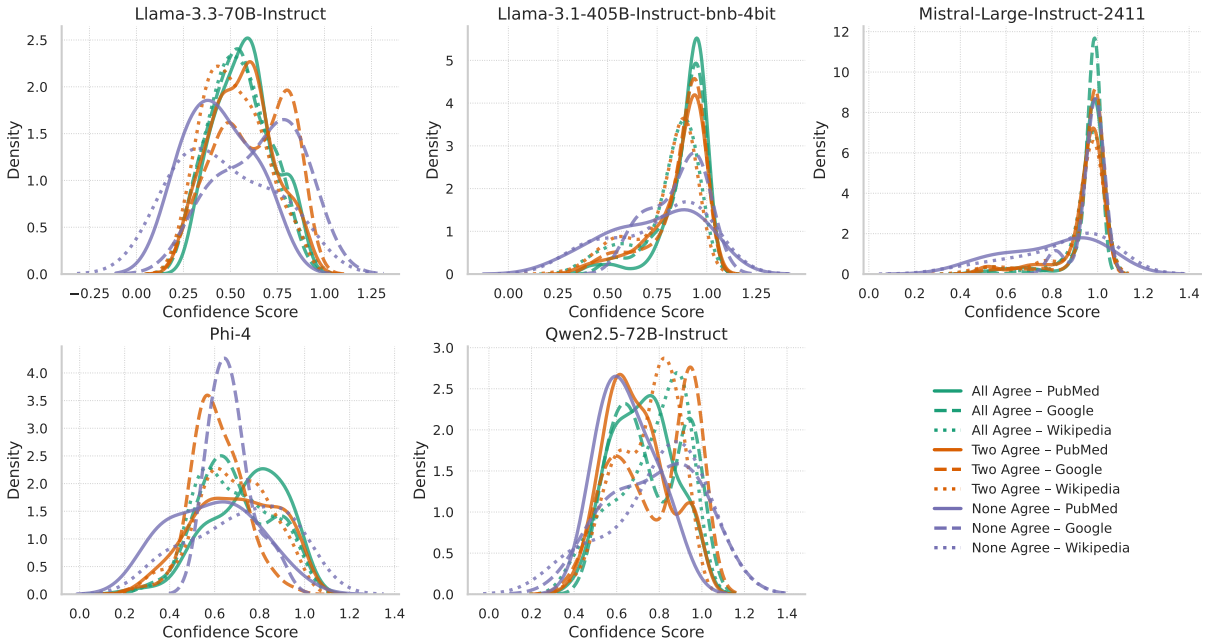

Figure 5: Confidence distribution (KDE) across different knowledge sources for the SCIFact dataset, illustrating variation in model certainty and inter-source disagreement during fact verification

Prompt

(Task) Answer the following question:

(Input) Facts: {evidence}

Statement: {claim}

Is the statement entailed by the given facts?

(A) Verdict X (B) Verdict Y (C) Verdict Z

Please answer just A or B or C:

Figure 6: Prompt Template for Verdict Prediction

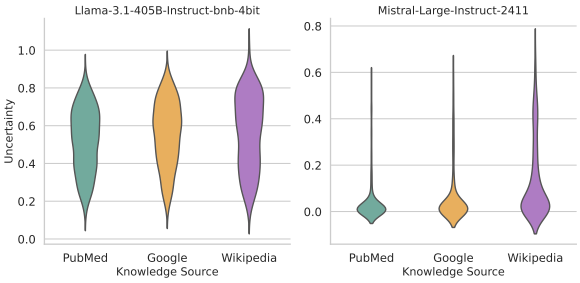

Figure 7: Violin Plot for Pubhealth