ClaimCLAIRE: A Trust-Aware Multi-Component Fact-Checking Agent for Open-World Claims

Xinman Liu **Mayank Sharma**
Stanford University
{xinman, masharma}@stanford.edu

Abstract

Verifying complex real-world claims against diverse and potentially unreliable open-web sources requires balancing evidence comprehensiveness with rigorous source reliability. Current automated fact-checking approaches often fail to address this holistically, losing contextual dependencies and applying trust signals monolithically at the document level. We introduce **ClaimCLAIRE**[1], a multi-component fact-checking agent that integrates four key innovations: (1) iterative component-aware decomposition with exhaustiveness validation, (2) holistic evidence gathering using a ReAct agent that maintains cross-component semantic awareness, (3) trust-modulated retrieval that weights evidence by source credibility to mitigate the influence of misinformation, and (4) adaptive gap-filling to address recall bottlenecks in under-supported sub-claims. Evaluated on the **AVeriTeC** benchmark, ClaimCLAIRE achieves **84.27%** accuracy and a macro-F1 of **0.806**. Our systematic ablations demonstrate that while decomposition alone can degrade performance, its integration with trust-aware retrieval and adaptive gap-filling yields a pipeline where component-level verdicts, source trust ratings, and deterministic AND-logic synthesis together support transparent, accountable fact verification.

1 Introduction

The widespread creation and rapid dissemination of false and misleading information, increasingly amplified by AI-generated content, poses critical challenges to the trustworthiness of information systems (AbuJarour et al., 2024; Vykopal et al., 2024). As manual fact-checking struggles to scale with misinformation volume, automated fact-checking (AFC) systems have become an essential component of trustworthy AI infrastructure (Agunlejika, 2025; Nakov et al., 2021). Early approaches leveraged NLP and machine learning to identify misleading information through language patterns (Saeidnia et al., 2025; Vosoughi et al., 2018), but recent advances in Large Language Models (LLMs) and retrieval-augmented generation (RAG) enable verification against open-web evidence rather than limited knowledge bases (Cao et al., 2023; Quelle and Bovet, 2024). Yet accuracy alone is insufficient for trustworthy AFC: systems must also be *accountable*, providing users with transparent, evidence-backed explanations that expose the reasoning behind verdicts and the reliability of the sources informing them (Nakov et al., 2021).

Claim verification in the open world introduces a fundamental challenge: balancing evidence comprehensiveness with source reliability (Hwang et al., 2025). Consider verifying "Sanders' opulent Vermont mansion cost $2.5 million": this requires checking ownership, property description, and price against diverse sources, where broad retrieval risks unreliable information while narrow retrieval may omit critical evidence. Current approaches address this piecemeal. Decomposition techniques split claims into atomic subclaims (Min et al., 2023; Wei et al., 2024), but verifying components in isolation often misses contextual dependencies. Hu et al. (2025) demonstrate that decomposition quality critically influences performance - appropriate granularity yields gains, while excessive fragmentation degrades accuracy. Complementarily, reliability-aware RAG approaches estimate trust at the document level (Deng et al., 2024; Hwang et al., 2025), but cannot adapt retrieval for different subcomponents. While agentic frameworks show improved reasoning (Ma et al., 2025), dynamically adapting retrieval based on subclaim-specific trust signals remains an open challenge.

We introduce **ClaimCLAIRE**, a trust-aware agentic fact-checking system that holistically integrates component-aware decomposition, reliability-

[1]Web App: https://claimclaire.vercel.app/, Anonymized GitHub: https://github.com/yo-lxmmm/ClaimCLAIRE

Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026), pages 73–91

weighted retrieval, and adaptive evidence recovery. Our **primary contribution is architectural**: we propose a unified pipeline that closes gaps left by prior work in decomposition context-preservation, component-level trust modulation, and evidence gap recovery, and validate each design choice through systematic ablations rather than leaderboard comparison, since existing AFC benchmarks operate under fundamentally different retrieval and label assumptions (Section 4). Building on CLAIRE's agentic architecture (Semnani et al., 2025), we adapt interleaved reasoning and retrieval for external web verification with four key innovations:

- **Iterative component-aware decomposition:** Progressive claim refinement adapting to claim-specific dependencies
- **Agentic holistic investigation:** ReAct agent maintaining component awareness while detecting cross-component connections
- **Trust-modulated retrieval:** Source-level credibility signals integrated into component evaluation
- **Adaptive gap-filling:** Dynamic targeted searches for under-verified components

Through systematic ablations on **AVeriTeC** (Schlichtkrull et al., 2023), ClaimCLAIRE achieves **84.27%** accuracy and **0.806** macro-F1, with each ablation stage demonstrating measurable contribution from individual architectural components. By producing transparent, evidence-backed reports with component-level verdicts and explicit source trust ratings, ClaimCLAIRE contributes to accountable and trustworthy AI systems for AFC.

2 Related Work

We organize prior work around the four major components of the fact-checking pipeline that ClaimCLAIRE advances: (1) decomposition quality and context preservation, (2) trust-aware retrieval, (3) agentic reasoning, and (4) mechanisms for resolving evidence gaps.

2.1 Claim Decomposition

Systems such as FActScore (Min et al., 2023) and VeriScore (Song et al., 2024) extract atomic facts from long-form text independently, while ClaimDecomp (Chen et al., 2022) generates boolean subquestions capturing explicit and implicit context. A persistent limitation is context loss: verifying atomic components in isolation fails to capture inter-dependencies - a price figure may be factually true but misleading without its "opulence" descriptor. While Hu et al. (2025) demonstrate that decomposition has non-monotonic effects across models, most pipelines decouple atomization from retrieval context. We address this through **iterative component-aware decomposition coupled with component-aware holistic investigation**, where an agent maintains global awareness of all components during evidence gathering, preserving contextual relationships without sacrificing granularity.

2.2 Trust-Aware Retrieval and Generation

Incorporating source credibility into RAG systems has become critical for AFC. Early approaches like DeClarE (Popat et al., 2018) integrate source trustworthiness into evidence assessment; more advanced systems integrate trust during generation: CrAM (Deng et al., 2024) scales down low-credibility tokens via attention modification, while RA-RAG (Hwang et al., 2025) estimates reliability through cross-document verification. These methods improve performance on benchmarks like CONFACT (Ge et al., 2025), which evaluates robustness against conflicting evidence. However, all apply credibility scores monolithically to entire claims. ClaimCLAIRE instead **applies trust-weighted evidence pooling at the component level**, enabling component-specific precision/recall trade-offs and exposing which sources informed each sub-verdict, supporting informed human oversight rather than opaque black-box outputs.

2.3 Agentic Architectures for AFC

Beyond static RAG, agentic architectures incorporate interleaved reasoning-action loops and tool-augmented generation for complex queries. ReAct demonstrates that interleaving reasoning traces with action execution significantly reduces hallucination and improves error recovery (Yao et al., 2023). Single-agent systems like SAFE (Wei et al., 2024) and FacTool (Chern et al., 2023) leverage iterative tool use to verify atomic facts, but neither targets standalone open-world claim verification: SAFE decomposes multi-paragraph LLM responses post-hoc for verification via Google Search, while FacTool operates across four specific domains requiring a full model response as input. Multi-agent systems like LoCal (Ma et al., 2025) and DelphiAgent (Xiong et al., 2025) employ adversarial or collaborative deliberation for

cross-checking, but at the cost of computational overhead. Across these paradigms, no existing system integrates component-level trust-aware retrieval, iterative exhaustiveness validation, or adaptive gap-filling for holistic open-world claim verification. Unlike shared-task systems, ClaimCLAIRE targets real-time open-web retrieval: a distinct setting we validate through systematic ablations isolating each component's contribution. ClaimCLAIRE demonstrates that a **single, component-aware ReAct agent can perform parallel evaluation of decomposed components, dynamically retrieve trust-weighted evidence, and efficiently fill gaps** without multi-agent coordination overhead.

3 Methodology

3.1 Overview

ClaimCLAIRE builds upon CLAIRE (Semnani et al., 2025), which combines LLM reasoning with retrieval to identify contradictory facts within Wikipedia. We extend it for open-world fact-checking with three key innovations: component-aware holistic investigation, systematic gap-filling, and trust-evaluated web retrieval.

Figure 1 illustrates our five-stage pipeline. The system decomposes the input claim into atomic components (Stage 1), then a ReAct agent conducts component-aware holistic web investigation, accumulating a shared evidence pool (Stage 2). Components are evaluated in parallel against this pool (Stage 3a); unverified components trigger targeted gap-filling and re-evaluation (Stage 3b). Deterministic AND logic then synthesizes the verdict: all components must be verified for Consistent (Stage 4), and an LLM generates a cited, human-readable report with wording feedback (Stage 5). Prompts are in the Appendix.

3.2 Stages

Stage 1: Claim Decomposition. Our system performs pre-investigation decomposition on user-provided claims, breaking them into atomic components that guide subsequent verification. We use a simplified approach that produces a flat list of atomic components without importance labels or complex logical relationships. Each component must be: (a) a single, indivisible piece of information, (b) independently verifiable, (c) self-contained with sufficient context, and (d) a factual assertion rather than opinion or prediction.

While CLAIRE also extracts atomic facts, it does so post-retrieval from Wikipedia passages and filters them for "worthiness" (excluding common knowledge, opinions, Wikipedia meta-statements). In contrast, we decompose pre-investigation and ensure the decomposed components are exhaustive and capture all factual assertions. To do so, we employ a separate LLM-as-judge model to check whether the proposed components capture all assertions in the original claim (Li et al., 2024). If the validation identifies missing components, they are added to the decomposition and the process repeats (up to 3 iterations) until the decomposition is validated as exhaustive. This addressed our initial observed failure modes where decomposition led to oversimplification where background facts are captured but key assertions are omitted. For illustration, consider the claim: *"Musician Lil Nas X partnered with Nike to create 'Satan shoes' that contain real human blood."* When decomposed into atomic components, this claim can be expressed as three distinct factual assertions: that musician Lil Nas X released a line of shoes called "Satan shoes," that Nike partnered with him on this product, and that the "Satan shoes" contain real human blood. These components capture all verifiable elements of the original claim while isolating each assertion for independent evaluation.

Stage 2: Holistic Evidence Gathering. Following decomposition, we employ a ReAct (Reasoning and Acting) agent architecture (Yao et al., 2023) for evidence gathering, where it is aware of the claim components by investigating the claim holistically through identification of nuanced cross-component connections, overall framing issues, and misleading context that querying the overall claim might gloss over, yet verifying components individually might miss. The agent follows a reasoning-action cycle: (1) reasoning about needed information based on components and evidence gathered, (2) selecting and executing tool actions, (3) observing results, and (4) updating understanding. This continues for up to 100 steps or until sufficient evidence is collected. It has access to the following tools:

- `search_web(query)`: Searches via a web search API, retrieving 20 results with LLM reranking to select top 10, each assigned trust ratings (reliable/mixed/unreliable)
- `explain(topic)`: Generates background explanations for technical terms
- `clarify_entity(entity)`: Performs targeted entity disambiguation

All evidence accumulates in a shared pool with explicit trust ratings assigned via a two-tier hybrid approach. First, domains are checked against curated lists derived from Wikipedia's Perennial Reliable Sources[2], which include 220 reliable, 323 unreliable, and 101 mixed-consensus domains. For domains not present in these lists, we use an LLM to classify the source, with confidence scoring and caching. Each rating includes a confidence score (0.0–1.0): list-based ratings receive a confidence of 0.95, while LLM-generated classifications are capped at 0.75. Trust ratings are computed in parallel for all search results using asynchronous processing.

Stage 3: Parallel Component Evaluation. We then evaluate each of the previously decomposed components against the accumulated evidence through a two-phase approach: (3a) All components are evaluated in parallel by an LLM using the shared evidence pool, producing verdicts of "verified," "refuted," or "unverified." (3b) If any components remain unverified, targeted evidence gathering proceeds in two sub-phases: First, batch evidence collection, which constructs queries combining each unverified component with the full claim and retrieves 15 results per component with LLM reranking to top 5, then adds all new evidence to the shared pool. Second, batch re-evaluation, where all unverified components are re-evaluated against the expanded pool, each benefiting from evidence collected for all unverified components.

We introduced gap-filling here because holistic investigation may miss component-specific evidence. From initial error analysis, we found that without gap-filling, some components would remain "unverified" even when contradictory evidence do exist, leading to cases where the system claims that there is insufficient evidence when the component is actually supported or refuted.

Stage 4: Verdict Synthesis. Given that components are already evaluated with LLM-generated verdicts in Stage 3, the final claim-level verdict is determined by deterministic AND logic rather than additional LLM synthesis: claims are Consistent only if all components are "verified," otherwise Inconsistent. We adopt the terminology Consistent and Inconsistent to reflect the structural nature of our verification task, where the focus is on whether all decomposed components align with available evidence rather than on replicating the multi-class veracity labels typical in fact-checking datasets. This framing emphasizes relational coherence between a claim and its evidence rather than categorical truth labeling. This rule-based approach eliminates LLM sampling stochasticity in verdict determination while also providing better transparency for users, so that they can inspect component-level verdicts and trace exactly how the final verdict was derived.

Stage 5: Report Generation. Finally, synthesizing the verdict determined in Stage 4, the component evaluations with reasoning, and the complete evidence pool, the LLM generates a brief human-readable report, citing specific sources using [n] notation where n refers to unique URLs in order of first appearance. The LLM is instructed to explicitly weight evidence by source reliability (reliable sources prioritized over mixed/unreliable) and to mention source trust ratings in the explanation (e.g., "According to reliable sources [1, 3]..." or "Some unreliable sources [5] claim..."). Additionally, the LLM produces wording feedback flagging misleading phrasing or suggesting improvements to claim clarity.

4 Experimental Results

4.1 Evaluation Dataset

We evaluate ClaimCLAIRE on the dev set of AVeriTeC[3], a fact-checking benchmark for real-world claims (Schlichtkrull et al., 2023). The benchmark contains claims sourced from 50 fact-checking bodies, with every claim annotated using four possible veracity labels: supported ($n = 122$), refuted ($n = 305$), conflicting/cherry-picking ($n = 38$), and not-enough-evidence ($n = 35$), and having substantial inter-annotator agreement ($\kappa = 0.619$). We restricted our evaluation to claims labeled supported, refuted, or conflicting/cherry-picking, as these categories provide definitive judgments appropriate for our evaluation. We encoded supported claims as Consistent, and both refuted and conflicting/cherry-picking claims as Inconsistent, since the presence of contradictory or selectively presented evidence reflects an inconsistency between the claim and the available information. We excluded not-enough-evidence claims ($n = 35$) because they represent ambiguous cases where insufficient information precludes a

[2] https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources/Perennial_sources

[3] https://huggingface.co/chenxwh/AVeriTeC/tree/main/data

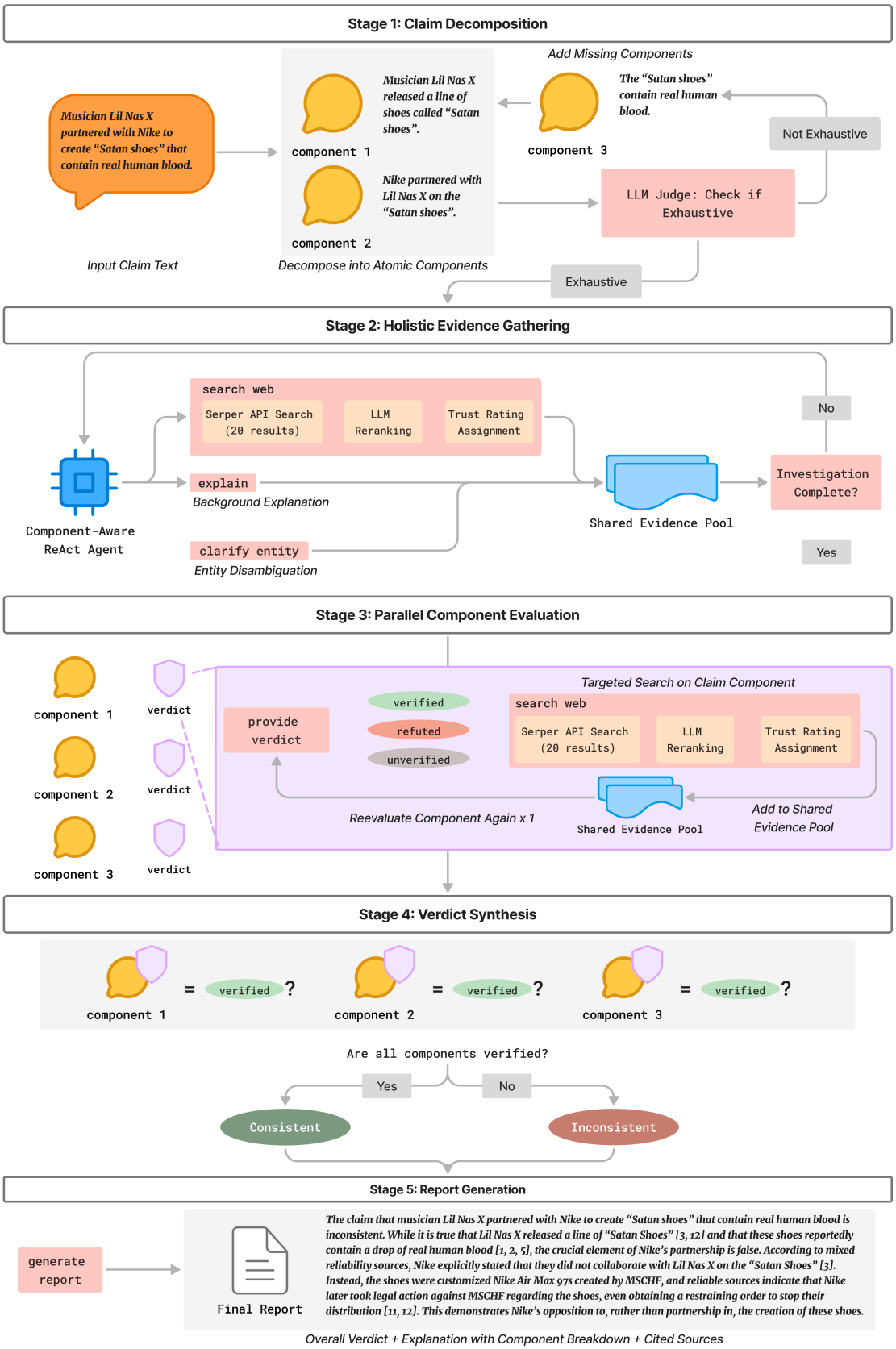


Figure 1: ClaimCLAIRE: Five-stage pipeline for claim verification

definitive verdict; extending ClaimCLAIRE to output an explicit "insufficient evidence" label when components remain unverified after gap-filling is a natural direction for future work. We acknowledge that collapsing Conflicting/Cherrypicking into Inconsistent and excluding NEI simplifies the label space; however, we argue this is principled: cherry-picked claims are by definition inconsistent with the full evidential picture. Extending ClaimCLAIRE to distinguish Conflicting/Cherrypicking from outright Refuted, and to output an explicit "insufficient evidence" verdict when components remain unverified after gap-filling, are natural directions for future work.

4.2 Experimental Setup

We use `claude-sonnet-4-20250514` for claim decomposition and validation (precise instruction-following), `gemini-2.5-flash` for the ReAct agent, component evaluation, and report generation (strong reasoning with low latency), and `gemini-2.5-flash-lite` for search result reranking (lightweight relevance comparison). We adopt such multi-model design because using a single model throughout would either sacrifice reasoning quality at the decomposition and agent stage or substantially increase cost and latency at the reranking stage. All models use temperature $= 0$ (greedy decoding) for consistency. For retrieval, we use the Serper API for web search (with temporal filtering by claim date). The ReAct agent is allowed up to 100 steps (`recursion_limit=100`), retrieving 10 results per `search_web()` call by default. For each search, we retrieve 20 results from Serper before reranking, and `gemini-2.5-flash-lite` reranks these to select the top 10 most relevant results. For gap-filling, we retrieve 15 results before reranking and select the top 5, which are added to the shared evidence pool.

For calibration, we note that recent systems evaluated on the AVeriTeC dev set report the following accuracies: INFACT (Rothermel et al., 2024) achieves 72.4% using a pre-fetched knowledge store with GPT-4o, and HerO (Yoon et al., 2024) achieves 75.2% with Llama-3.1-70B (as reported in Putta et al. 2025). ClaimCheck (Putta et al., 2025), the most directly comparable system to ours as it also uses live Serper web retrieval, achieves 76.4% on the full 500-claim dev set with four-class labels. However, direct numerical comparison with ClaimCLAIRE is not straightforward: our binary label scheme and exclusion of 35 NEI claims simplify the evaluation relative to four-class systems. We therefore rely on our internal ablation baseline (A0), a one-shot RAG system under identical retrieval and label conditions, as our primary reference point for interpreting system gains.

4.3 Ablation Analysis

To assess each component's contribution, we conduct cumulative ablations with consistent retrieval, LLM settings, and metrics.

A0: Baseline RAG. Retrieves 10 web results and directly classifies claims as Consistent or Inconsistent. No decomposition, agent reasoning, trust evaluation, or gap-filling.

A1: + ReAct Agent. Adds iterative reasoning and evidence gathering (up to 100 steps) for thorough investigation.

A2: + Iterative Decomposition. Pre-investigation decomposition with validation (up to 3 iterations) ensures exhaustive component extraction. The ReAct agent becomes component-aware, maintaining global awareness during holistic investigation.

A3: + Trust Rating. Adds source credibility evaluation via curated domain lists and LLM classification with confidence scoring. Evidence is tagged with trust ratings (reliable/mixed/unreliable) informing component-level evaluation.

A4: + Adaptive Gap-Filling (ClaimCLAIRE). Complete system with targeted evidence collection for unverified components, followed by batch re-evaluation where all unverified components benefit from shared evidence.

4.4 Evaluation Metrics

A primary use case of ClaimCLAIRE is serving as a first-pass filter that surfaces suspicious claims for human review, supporting fact-checker prioritization rather than replacing their judgment. False positives waste human effort by incorrectly flagging legitimate claims, while false negatives miss true inconsistencies and allow misinformation to persist. We report accuracy and macro-F1, along with per-class precision, recall, and F1-score for both Consistent and Inconsistent verdicts. Given the class imbalance (approx. 26% Consistent), we additionally report weighted F1 to complement macro-F1. We visualize a confusion matrix to understand the trade-offs in predictions.

4.5 Results

We present our ablation results in Table 1. These results reflect single-run evaluations; differences be-

tween adjacent ablation stages should be treated as indicative rather than definitive given the absence of significance testing. The baseline RAG system (A0) achieved 81.29% accuracy and a macro-F1 of 0.754, with substantially weaker performance on the Consistent class (F1 = 0.633) than the Inconsistent class (F1 = 0.875). This difference indicated that the one-shot retrieval pipeline was better at flagging inconsistencies than verifying support, as also reflected in the low Consistent recall (0.615).

Adding a ReAct-style agent (A1) produced consistent gains across metrics, improving accuracy to 82.6% (+1.3%) and macro-F1 to 0.774 (+2.0%). The agent most strongly benefited the Consistent class, increasing both precision (0.669 vs. 0.652) and recall (0.664 vs. 0.615), suggesting that multi-step retrieval and contextual reasoning enabled the system to surface evidence missed by the one-shot baseline.

Interestingly, iterative decomposition (A2) decreased system performance, reducing macro-F1 to 0.742 and accuracy to 81.5%. While Consistent precision increases (0.688), Consistent recall drops sharply to 0.541, indicating that solely decomposing claims into finer-grained components risks introducing noise. This aligns with Hu et al. (2025)'s findings that show decomposition has non-monotonic effects and it must be calibrated to an appropriate level of granularity.

Trust-rating (A3) yielded the strongest single-stage improvement, raising accuracy to 84.95% (+3.4% over A2) and macro-F1 to 0.795 (+5.3%). Incorporating credibility signals boosted precision for both Consistent (0.750) and Inconsistent (0.878) classes and led to particularly high Inconsistent recall (0.924). However, Consistent recall remained modest (0.639), suggesting that trust filters may be overly conservative, occasionally discarding lower-credibility but relevant supporting evidence.

Finally, the full ClaimCLAIRE system (A4) issued targeted searches for under-supported components. This directly addresses the recall bottleneck introduced by decomposition and trust filtering. A4 achieved the highest overall macro-F1 (0.806) and substantially improved Consistent recall to 0.779 (+14% over A3) while maintaining strong Inconsistent-class performance (F1 = 0.890, precision = 0.916). These results indicate that many true claims failed at earlier stages not because they were difficult to verify, but because key supporting evidence was missing; adaptive gap-filling effectively corrected this failure mode. Overall, A4 provided the best balance between precision and recall (see Appendix B), delivering the strongest performance across both classes. Notably, A3 and A4 achieve identical weighted F1 (0.846), reflecting the dataset's class imbalance (26% Consistent); macro-F1 remains the more informative metric here, as it weights both classes equally and better captures gains on the harder minority Consistent class. Despite the improvements being modest, the stepwise advances achieved by ClaimCLAIRE validate the necessity of a holistic approach that integrates decomposition-aware investigation with trust-weighted retrieval and adaptive gap-filling. We also note that A1 already achieves 82.58% accuracy, within 1.7 points of the full ClaimCLAIRE system, suggesting that practitioners with latency or cost constraints may find the ReAct-only configuration a practical alternative.

4.6 Error Analysis

To better understand ClaimCLAIRE's limitations, we conduct error analysis on the 73 incorrect predictions (15.7% of 464 evaluated claims, 1 rejected by LLM content-filter rejection). As shown in Appendix B, these errors comprise 27 false positives (true claims incorrectly marked Inconsistent) and 46 false negatives (false claims incorrectly marked Consistent), revealing an asymmetry where the system more frequently fails to detect misinformation than incorrectly flags true claims. We provide concrete examples of each error category in Appendix C. Below we detail four main error categories of failure cases.

Insufficient context or vague phrasing in original claim text. A large source of error stems from the claim text itself being phrased vaguely or lacking critical contextual information needed for accurate verification, resulting in subsequent incorrect component extraction and evidence matching. For example, some claims in the AVeriTeC dataset include "For a cumulative 29 of our 60 years of existence as a nation, we have been under military rule," "558 people were killed by the police in 2018, while 201 people died in police custody," and "They [the Democrats] want to ... ban fracking." While AVeriTeC annotators had access to contextual information during labeling, ClaimCLAIRE sees only the claim text and therefore cannot reliably infer the intended subject, location, or context.

Ablation	Accuracy	Macro-F1	Wgt-F1	Per-label F1		Precision		Recall	
				C	I	C	I	C	I
Baseline RAG (A0)	81.29%	0.754	0.811	0.633	0.875	0.652	0.866	0.615	0.883
+ ReAct (A1)	82.58%	0.774	0.826	0.667	0.882	0.669	0.881	0.664	0.883
+ Iterative Decomposition (A2)	81.51%	0.742	0.807	0.606	0.879	0.688	0.848	0.541	0.913
+ Trust-rating (A3)	**84.95%**	0.795	**0.846**	0.690	**0.901**	**0.750**	0.878	0.639	**0.924**
+ Gap filling (A4) = ClaimCLAIRE	84.27%	**0.806**	**0.846**	**0.722**	0.890	0.674	**0.916**	**0.779**	0.866

Table 1: Ablation study results on AVeriTeC dev set ($n = 464$ claims, excluding "not-enough-evidence" labels and 1 LLM-rejected claim). Bold indicates best performance. C = Consistent class, I = Inconsistent class.

Evidence provenance and circular reporting. ClaimCLAIRE sometimes fails to catch erroneous reporting that propagates through the media ecosystem, particularly when the source is high-credibility. For instance, in one case, the system incorrectly verified the claim "Trump Administration claimed songwriter Billie Eilish Is Destroying Our Country In Leaked Documents." According to ground truth justifications on the AVeriTeC dataset, this claim originated from a Washington Post article that wrongly attributed such a statement to the Trump administration. Since the Washington Post is tagged as a reliable source in ClaimCLAIRE's trust rating system, and temporal filtering restricts retrieval to articles before the claim date (before the error was publicly corrected), the system found multiple secondary sources corroborating the Washington Post's reporting, lacking the mechanism to identify when news providers are echoing false reports from credible outlets.

Semantic precision and verdict leniency. A subset of errors arises not from a failure to retrieve or understand evidence, but from a misalignment in verdict thresholds between the model and human annotators. For example, regarding the claim "Duterte has signed order to open nuclear power plant," ClaimCLAIRE correctly noted in its final report that the order was only to study nuclear energy, yet still classified the claim as Consistent because the administration was openly considering the opening. This represents a calibration error where the system prioritizes broad thematic alignment and fails to penalize subtle but critical semantic distinctions that human annotators were trained to flag as Inconsistent.

Fluctuating evidence gathering via web search API. Although the Serper web search API is generally stable, evidence retrieval can fluctuate because we cannot fully control what results are returned at each query. This is particularly noticeable for borderline claims, where changes in the retrieved evidence may impact the final verdict. Furthermore, API rate limiting and content-filter blocks triggered by sensitive topics adds to the fluctuation.

Notably, we do not observe failure cases stemming from temporal ambiguity or evolving guidance, as we explicitly prevented temporal leakage by restricting evidence retrieval via the Serper API to the claim date or earlier. This aligns with the temporal and information constraints given to the AVeriTeC annotators.

5 Insights & Discussion

Our ablation study reveals that claim decomposition alone is insufficient: simply decomposing claims (A2) reduced the accuracy of the ReAct system (A1). This could be attributed to over-fragmentation, where breaking claims into atomic components strips away necessary contextual information that helps disambiguate meaning (Hu et al., 2025); however, given that ClaimCLAIRE retrieves evidence holistically with awareness of all claim components, we hypothesize that such over-fragmentation impacts less on evidence retrieval than on evaluation—where, regardless of their centrality to the claim, each component is independently verified against mixed-credibility evidence and weighted equally for verdict synthesis, which in turn introduces verification noise and amplifies uncertainty. Crucially, our results show that this is mitigated by the integration of trust-aware retrieval (A3) and adaptive gap-filling (A4), which boosted the system performance. As such, we frame decomposition as a structural premise of our architecture enabling granular verification, yet its utility is strictly contingent on the subsequent integration of trust-aware retrieval and adaptive gap-filling.

Furthermore, we observe a complementary relationship between trust-ratings and gap-filling. While the addition of trust ratings provides a critical performance boost by filtering unreliable sources, strict credibility filtering can create an

overly conservative system, reducing recall for Consistent claims where high-credibility coverage is sparse. The full ClaimCLAIRE system resolves this trade-off through adaptive gap-filling, which acts as a corrective mechanism by dynamically recovering evidence for these under-supported components. Ultimately, ClaimCLAIRE reflects the necessity of a holistic approach towards robust open-world fact-checking.

We also acknowledge an inherent bias introduced by AND-logic verdict synthesis: as the number of components grows, the probability that at least one component remains unverified or refuted increases, creating a systematic tendency toward *Inconsistent* verdicts. This is reflected in our ablation results—adding decomposition (A2) sharply drops Consistent recall to 0.541 while simultaneously raising Inconsistent recall to 0.913. Relatedly, we note that our reranker is also optimized for refutation, which may further reduce evidence recall for Consistent claims. However, such a trade-off is an intentional design choice: ClaimCLAIRE's primary use case is as a first-pass filter for human fact-checkers, where false negatives (missed misinformation) are costlier than false positives (flagging true claims for review). The AND-logic and the reranker thus encode a conservative verification standard appropriate for this setting, at the cost of Consistent recall.

Meanwhile, our error analysis reflects fundamental challenges in open-world fact verification. For instance, insufficient context in claim text in the existing benchmark prevents reliable verification, while cases where multiple sources echo single erroneous reports from credible outlets expose weaknesses in static trust-rating approaches. Additionally, misalignment in verdict thresholds causes the system to accept claims as consistent despite identifying contradictory evidence, reflecting different standards and semantic interpretations compared to human annotators. These patterns suggest future systems need discourse context integration, provenance-tracking for echoed misinformation, and dynamic credibility assessment with precise verdict calibration.

6 Conclusion

ClaimCLAIRE demonstrates that combining decomposition-aware holistic investigation with trust-weighted retrieval and adaptive gap-filling meaningfully improves open-world fact verification. While the performance gains are modest, the system provides a transparent framework that surfaces component-level evidence and failure modes, offering a practical step toward more interpretable fact-checking agents.

Limitations

We acknowledge several limitations of this study. First, although AVeriTeC contains high-quality, real-world claims concerning events from around the world, all annotations and evidence are drawn exclusively from English-language fact-checking sources, which limits the generalizability of findings to multilingual or non-English fact-verification settings.

Second, while we explicitly bias the system toward minimizing false negatives for Inconsistent cases through a deterministic AND-logic verdict synthesis, this strategy may inadvertently increase false positives, particularly for claims involving partial truths, mixed evidence, or legitimate uncertainty. Additionally, the end-to-end pipeline exhibits substantial latency (approximately 40 seconds per claim) which limits its practicality for large-scale deployment.

Third, the ablation study reveals only modest performance gains, and without multi-run variance estimates or significance testing, differences between ablation stages should be interpreted with caution; furthermore, while calibration figures from prior systems are provided in Section 4.2, direct numerical comparison remains limited by differences in label schemes and retrieval assumptions. Relatedly, system performance is closely tied to the capabilities and biases of the underlying LLMs (Gemini and Claude) and the external retrieval infrastructure (Serper API); errors in retrieval, reranking, or model reasoning can propagate through the pipeline and degrade overall accuracy. We designed ClaimCLAIRE prioritizing transparency and interpretability, and while specific components provide incremental improvements, deeper challenges related to evidence scarcity, ambiguity in open-web information, and the inherent complexity of claim interpretation remain unresolved in this work.

Finally, while the claim reports ClaimCLAIRE produces can identify partial truth or graded veracity, this explanatory depth is inevitably compressed into a reductive binary Consistent/Inconsistent label for benchmark evaluation, which does not reflect the nuanced,

context-dependent nature of real-world fact-checking. Future work should explore multilingual extensions, evaluation on larger and more diverse datasets with graded veracity labels that better reflect the complex realities of fact-checking, more efficient pipeline designs, open-source LLM alternatives, and methods that better address evidence scarcity.

Ethics Statement

Fact-checking is often framed as an epistemic tool for limiting the spread and influence of misinformation; however, the system presented in this paper is not intended as an authoritative arbiter of truth. The labels and verdicts produced by ClaimCLAIRE reflect evidence surfaced through temporally filtered web search and are constrained by the limitations of the underlying language models, the Serper search API, and our trust-weighting mechanism. Acting on these veracity estimates without human oversight risks epistemic harm, particularly in high-stakes or automated decision-making contexts. In addition, systems of this kind raise concerns about representational fairness and transparency due to the uneven distribution of online evidence. Because the pipeline relies on English-language web content and search-engine retrieval, claims originating from communities, regions, or epistemic traditions that are underrepresented online may be systematically disadvantaged or misclassified. To minimize these risks, any deployment should provide transparency about evidence coverage, clearly communicate uncertainty, and avoid application in settings where misclassification could produce social, political, or institutional harm.

Impact Statement

Automated fact-checking at scale could meaningfully support journalists, researchers, and policymakers in combating misinformation during rapidly evolving news cycles where manual verification cannot keep pace. By producing transparent, component-level verdicts with cited sources, systems like ClaimCLAIRE contribute to more trustworthy and accountable AI-assisted information verification - and model evidence-based reasoning for everyday users, supporting broader information literacy. We discuss deployment risks and limitations in the Ethics Statement above.

References

Safa'a AbuJarour, Ameera Qarariah, Noor Saadeh, and Mojahida Salem. 2024. AI, Misinformation, and Fake News: A Literature Review of Ethical and Technical Approaches. In Nadia Mansour and Lorenzo M. Bujosa Vadell, editors, *Finance and Law in the Metaverse World: Regulation and Financial Innovation in the Virtual World*, pages 641–652. Springer Nature Switzerland, Cham.

Taiwo Agunlejika. 2025. AI-Driven Fact-Checking in Journalism: Enhancing Information Veracity and Combating Misinformation: A Systematic Review.

Han Cao, Lingwei Wei, Mengyang Chen, Wei Zhou, and Songlin Hu. 2023. Are Large Language Models Good Fact Checkers: A Preliminary Study. Publisher: arXiv Version Number: 1.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative ai – a tool augmented framework for multi-task and multi-domain scenarios. *Preprint*, arXiv:2307.13528.

Boyi Deng, Wenjie Wang, Fengbin Zhu, Qifan Wang, and Fuli Feng. 2024. Cram: Credibility-aware attention modification in llms for combating misinformation in rag. *Preprint*, arXiv:2406.11497.

Ziyu Ge, Yuhao Wu, Daniel Wai Kit Chin, Roy Ka-Wei Lee, and Rui Cao. 2025. Resolving conflicting evidence in automated fact-checking: A study on retrieval-augmented llms. *Preprint*, arXiv:2505.17762.

Qisheng Hu, Quanyu Long, and Wenya Wang. 2025. Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6313–6336, Albuquerque, New Mexico. Association for Computational Linguistics.

Jeongyeon Hwang, Junyoung Park, Hyejin Park, Dongwoo Kim, Sangdon Park, and Jungseul Ok. 2025. Retrieval-augmented generation with estimation of source reliability. *Preprint*, arXiv:2410.22954.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *Preprint*, arXiv:2412.05579.

Jiatong Ma, Linmei Hu, Rang Li, and Wenbo Fu. 2025. Local: Logical and causal fact-checking with llm-based multi-agents. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 1614–1625, New York, NY, USA. Association for Computing Machinery.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *Preprint*, arXiv:2305.14251.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated Fact-Checking for Assisting Human Fact-Checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 4551–4558, Montreal, Canada. International Joint Conferences on Artificial Intelligence Organization.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Brussels, Belgium. Association for Computational Linguistics.

Akshith Reddy Putta, Jacob Devasier, and Chengkai Li. 2025. Claimcheck: Real-time fact-checking with small language models. *Preprint*, arXiv:2510.01226.

Dorian Quelle and Alexandre Bovet. 2024. The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence*, 7:1341697.

Mark Rothermel, Tobias Braun, Marcus Rohrbach, and Anna Rohrbach. 2024. InFact: A strong baseline for automated fact-checking. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 108–112, Miami, Florida, USA. Association for Computational Linguistics.

Hamid Reza Saeidnia, Elaheh Hosseini, Brady Lund, Maral Alipour Tehrani, Sanaz Zaker, and Saba Molaei. 2025. Artificial intelligence in the battle against disinformation and misinformation: a systematic review of challenges and approaches. *Knowledge and Information Systems*, 67(4):3139–3158.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Sina Semnani, Jirayu Burapacheep, Arpandeep Khatua, Thanawan Atchariyachanvanit, Zheng Wang, and Monica Lam. 2025. Detecting corpus-level knowledge inconsistencies in Wikipedia with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34827–34854, Suzhou, China. Association for Computational Linguistics.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. Veriscore: Evaluating the factuality of verifiable claims in long-form text generation. *Preprint*, arXiv:2406.19276.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*. Publisher: American Association for the Advancement of Science.

Ivan Vykopal, Matúš Pikuliak, Simon Ostermann, and Marián Šimko. 2024. Generative Large Language Models in Automated Fact-Checking: A Survey. *arXiv preprint*. ArXiv:2407.02351 [cs].

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. *Preprint*, arXiv:2403.18802.

Cheng Xiong, Gengfeng Zheng, Xiao Ma, Chunlin Li, and Jiangfeng Zeng. 2025. Delphiagent: A trustworthy multi-agent verification framework for automated fact verification. *Information Processing Management*, 62(6):104241.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Yejun Yoon, Jaeyoon Jung, Seunghyun Yoon, and Kunwoo Park. 2024. HerO at AVeriTeC: The herd of open large language models for verifying real-world claims. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 130–136, Miami, Florida, USA. Association for Computational Linguistics.

Appendix

A Prompts

A.1 Holistic Investigation Agent Prompt

You will be given a "claim" statement to fact-check, along with its decomposition into atomic components.
Your task is to conduct a thorough holistic investigation across the web and news sources. While you are AWARE of the components below, your investigation should be holistic to catch:

- Cross-component connections
- Overall framing issues
- Misleading context

Components to be aware of:

1. Component 1
2. Component 2
3. ...

As you conduct your investigation, you may come across articles that support the claim. However, you should continue searching for inconsistencies that might exist in other places. Inconsistencies might appear in subtle or indirect ways.
IMPORTANT: When evaluating evidence from search results, pay close attention to source reliability ratings:

- **Reliable sources** (weight: 1.0, marked with ' Reliable Source'): Major news organizations, academic journals, established institutions. Give these sources HIGHEST weight (1.0) in your analysis.
- **Mixed reliability sources** (weight: 0.5, marked with ' Mixed Reliability'): Opinion sites, partisan media, aggregators. Use these with CAUTION (weight: 0.5) and cross-reference with reliable sources.
- **Unreliable sources** (weight: 0.1, marked with ' Unreliable Source'): Tabloids, social media, known misinformation sources. Give these sources LOWEST weight (0.1) or disregard them entirely.

Trust threshold guidance:

- Effective weight = trust weight × confidence.
- **High weight (>0.7):** Strong, decision-quality evidence.
- **Medium weight (0.4–0.7):** Use with caution and corroborate.
- **Low weight (<0.4):** Weak evidence; treat as anecdotal.
- Aim for at least three independent sources with effective weight >0.7 before stating a strong conclusion.

When making your final verdict, prioritize evidence from reliable sources. If reliable sources contradict the claim, the claim is likely inconsistent. If only unreliable sources contradict it, investigate further with reliable sources.
You will conduct your investigation in multiple steps. At each step, you should think about the information you have gathered so far, and choose one of these available tools:

- `explain(topic: str)`: Use this action to understand the basics of a specific term or concept you encounter, for example a technical term or the rules of a sport.
- `clarify_entity(entity_name_and_description: str)`: Use this action to get a report on an entity (person, organization, event etc.) to clarify other entities with similar names. This will help you properly differentiate similar-sounding entities when researching inconsistencies. For example, `clarify_entity("WW III wrestling event")` will explain all potential events with similar names, or the same event in different years.
- `search_web(query: str)`: Use this action to search the web and news sources. Search results will include trust ratings for each source—use these to weight the evidence appropriately.

A.2 Claim Decomposition Prompt

You will be given a claim to fact-check. Your task is to break it down into ALL atomic, verifiable components.
CRITICAL RULES:

1. **Extract EVERY factual assertion** – Don't oversimplify! Most claims have multiple parts.
2. **Quotes and allegations are MANDATORY components** – If the claim says someone "said X" or "did Y", that MUST be a separate component.
3. **Actions, events, and statements are separate from identities** – Don't just verify "who someone is", verify "what they did/said".
4. **Keep components atomic but complete**: Each should be independently verifiable, but don't lose information.
5. **Mixed claims need full breakdown**: Claims with both true and false parts must be split so each can be evaluated.
6. **All components use AND logic**: Every single component must be verified for the claim to be consistent.

Common mistakes to avoid:

- Only extracting background facts (e.g., "X is a senator") while ignoring the main allegation
- Combining multiple assertions into one component
- Skipping quotes, statements, or controversial parts
- Extract both the context AND the main assertion
- Treat each quote, action, or event as a separate component

Examples:
Claim: "A photograph shows Bernie Sanders' opulent Vermont mansion, purchased in 2016 for $2.5 million."
Components:

- "The property is owned by Bernie Sanders"
- "The property is located in Vermont"
- "The property is an opulent mansion"
- "The property was purchased in 2016"
- "The purchase price was $2.5 million"

Claim: "Trump donated his salary and Melania had only 4 staff while Obama donated nothing and Michelle had 23 staff."
Components:

- "Donald Trump donated his presidential salary"
- "Melania Trump had a White House staff of 4"
- "Barack Obama did not donate his presidential salary"
- "Michelle Obama had a White House staff of 23"

Claim: "Kentucky Derby jockey John Velazquez turned down an invitation to the White House and said, 'if I wanted to see a horse's ass I would of came in second.'"
Components:

- "John Velazquez is a Kentucky Derby jockey"
- "John Velazquez received an invitation to the White House"
- "John Velazquez turned down the White House invitation"
- "John Velazquez said 'if I wanted to see a horse's ass I would of came in second'"

A.3 Decomposition Validation Prompt

You will be given an original claim and a proposed decomposition into components.
Your task: Check if the decomposition is EXHAUSTIVE and captures ALL factual assertions in the original claim.
CRITICAL RULES:

1. **Every specific fact, quote, action, or allegation** in the original claim MUST appear in the components
2. **Background context alone is NOT exhaustive** – if the claim makes a specific assertion, it must be decomposed
3. **Quotes and controversial statements** are the MOST IMPORTANT parts to capture
4. If the original claim has N distinct factual assertions, the decomposition should have ∼N components

Ask yourself:

- Are there any quotes, statements, or allegations in the original claim that are NOT in the components?
- Are there any actions or events mentioned in the original claim that are missing?
- Did the decomposition only extract background facts while ignoring the main assertion?
- If I only verified the components, would I have verified the ENTIRE original claim?

Examples:
Original: "Says Kentucky Derby jockey John Velazquez turned down an invitation to the White House and said, 'if I wanted to see a horse's ass I would of came in second.'"
Proposed components:

- "John Velazquez is a Kentucky Derby jockey"

Result: `is_exhaustive=FALSE`
Missing:

- "John Velazquez received an invitation to the White House"
- "John Velazquez turned down the White House invitation"
- "John Velazquez said 'if I wanted to see a horse's ass I would of came in second"'

Explanation: The decomposition only captured background context (that he's a jockey) but missed the main claims about the invitation and the quote.

Original: "Trump donated his salary and Melania had only 4 staff."
Proposed components:

- "Donald Trump donated his presidential salary"
- "Melania Trump had a White House staff of 4"

Result: `is_exhaustive=TRUE`
Missing: []
Explanation: All factual assertions are captured.

Return:

- `is_exhaustive`: true ONLY if ALL assertions are captured, false otherwise
- `missing_components`: list of any missing factual assertions (empty if exhaustive)
- `explanation`: brief explanation of what's missing or why it's complete

A.4 `Explain` Prompt

You will be given a topic or term.
Your task is to write a concise, self-contained explanation of the topic, providing background information for people who are unfamiliar with it.
If a term, event, or concept has multiple interpretations or meanings, briefly list all plausible ones.

Example input 1
Topic: Infanta Amalia

Example output 1
"Infanta Amalia" refers to a title and name in Spanish and Portuguese royal contexts. "Infanta" is a title used in Spain and Portugal for the daughters of a monarch who are not heir apparent, similar to "princess" in English. "Amalia" is a given name. Therefore, "Infanta Amalia" would refer to a princess named Amalia within a Spanish or Portuguese royal family.

Example input 2
Topic: The Great Gatsby

Example output 2
"The Great Gatsby" is a novel by F. Scott Fitzgerald, published in 1925. It is considered a classic of American literature set during the Jazz Age. The term could also refer to film adaptations of the novel (including notable versions from 1974 and 2013), theatrical productions, or an opera adaptation.

A.5 `clarify_entity` Prompt

You will be given an entity name and a list of search results about it.
Your task is to write a concise paragraph explaining the entity and disambiguating it from other similar entities found in the search results.
Entities with similar names might lead to confusion—your goal is to clarify the differences.
Pay attention to people with the same name, events with the same name but different years or locations, organizations with similar names but different purposes or locations, etc.

A.6 Component Evaluation Prompt

You will be given a single atomic claim component to verify and a list of search results.
Your task is to determine if the component is:

- **verified**: Evidence clearly supports it
- **refuted**: Evidence clearly contradicts it
- **unverified**: Insufficient or conflicting evidence

Weight evidence by source reliability: reliable sources (weight 1.0) > mixed sources (weight 0.5) > unreliable sources (weight 0.1). Use the provided effective weights (trust weight $\times$ confidence) to judge strength of each citation.

- Effective weight > 0.7 = strong evidence, < 0.4 = weak evidence.
- Prefer citing at least three high-weight sources when available.

Provide brief reasoning with source citations in [n] format.

A.7 Report Generation Prompt

Your job is to explain the result of an inconsistency detection investigation to a user in simple terms. You will be provided with the original claim, its decomposition into components, and evaluation results for each component.
Return an object with fields: `verdict` (consistent or inconsistent), `wording_feedback` (guidance on improving the claim wording), and `explanation` (1–2 paragraphs citing specific search results using [n] notation).
CRITICAL – Determining the verdict (SIMPLIFIED LOGIC):

- The verdict has already been determined by evaluating ALL components with AND logic.
- ALL components must be "verified" for the claim to be consistent.
- If ANY component is "refuted" or "unverified", the overall claim is "inconsistent".
- Your task is to write a coherent explanation that synthesizes the component evaluations.

IMPORTANT: When writing the explanation:

- **Weight evidence by source reliability**: Prioritize evidence from reliable sources (weight: 1.0) over mixed (weight: 0.5) or unreliable (weight: 0.1) ones.
- **If reliable sources contradict the claim**: The claim is likely inconsistent—state this clearly. Reliable sources have weight 1.0.
- **If only unreliable sources contradict**: Mention this but note that reliable sources should be consulted. Unreliable sources have weight 0.1.
- **If reliable sources support the claim**: The claim is likely consistent, even if unreliable sources contradict it. Reliable sources (weight: 1.0) outweigh unreliable ones (weight: 0.1).
- **In your explanation**: Explicitly mention the reliability of sources you cite, e.g., "According to reliable sources [1, 3]..." or "Some unreliable sources [5] claim..."

A.8 LLM Reranker Prompt

You are an intelligent assistant tasked with ranking passages based on their relevance to a given query, and their usefulness in refuting a false claim. Your goal is to provide an accurate ranking of the passages in descending order of usefulness. To complete this task, follow these steps:

1. Carefully read the claim and all the passages.
2. For each passage, analyze its content and determine its relevance and usefulness for refuting the given claim. Consider the following factors:
    a. Relevance: How closely does the passage relate to the topic of the claim?
    b. Specificity: Does the passage provide specific facts, figures, or details that directly contradict the claim?
3. Rank the passages based on their overall usefulness for fact-checking the claim. The most useful passage should be ranked first, and the least useful passage should be ranked last.
4. Present your final ranking in the given format. For example, [1, 2, 4, 3] would indicate that passage [1] is the most useful, followed by [2], then [4], and finally [3] is the least useful.

Only provide the ranking result. Do not include any explanations.

A.9 Query Rephrase Prompt

You are helping to rephrase a search query that was blocked by a content filter.
The original query was rejected, likely because it contains sensitive terms, profanity, slurs, or other content that violates search API policies.
Your task: Rephrase the query to search for the same information while avoiding sensitive terms.
CRITICAL: You MUST preserve ALL specific details from the original query:

- Keep all names, dates, locations, and specific facts EXACTLY as they appear
- Keep all quoted phrases (except the offensive term itself)
- Maintain the exact context and claim being investigated
- Only replace the offensive/blocked terms with neutral descriptive alternatives

Guidelines:

1. Replace ONLY the explicit/offensive terms with neutral, descriptive alternatives
2. Keep ALL other details, names, dates, and context identical
3. Use professional, academic language for replacements
4. Maintain the exact factual claim and search intent
5. Do NOT summarize, generalize, or lose any specificity

Examples:

- "Judge Amy Barrett said N-word not hostile environment" → "Judge Amy Barrett said racial slur not hostile environment"
- "Photo shows [explicit violence] at protest" → "Photo shows graphic violence at protest"

Original query: {query}
Respond with ONLY the rephrased query, no explanation or quotes.

A.10 Source Trust Rating Prompt

You are a fact-checking expert evaluating source credibility.
Classify the following source into ONE of these categories:

- **"reliable"** (weight: 1.0): Generally trustworthy news sources with strong editorial standards (e.g., major newspapers, established news agencies, peer-reviewed journals)
- **"mixed"** (weight: 0.5): Sources with mixed reliability or potential bias (e.g., opinion-heavy sites, some partisan media, aggregators)
- **"unreliable"** (weight: 0.1): Generally unreliable sources (e.g., tabloids, social media, user-generated content sites, known misinformation sources)

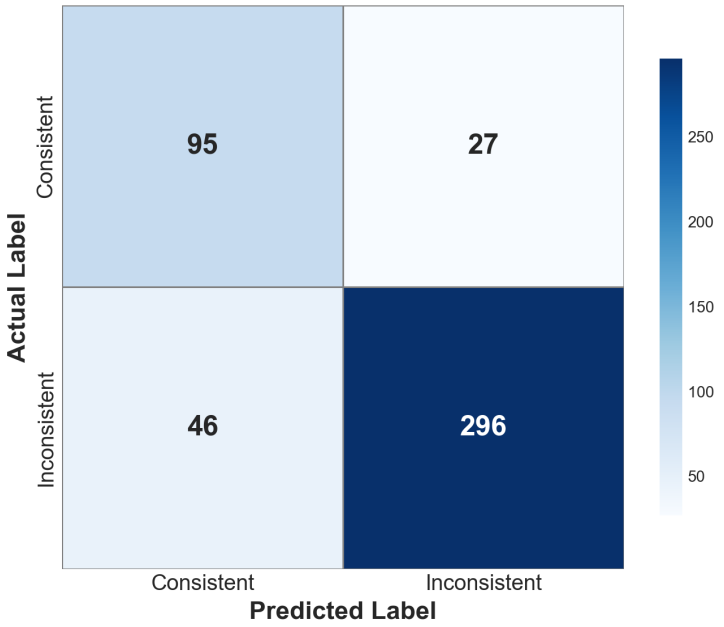
Source to classify:
Domain: {domain}
URL: {url}
{f'Title: {title}' if title else ""}
Consider:

1. Editorial standards and fact-checking processes
2. Track record of accuracy
3. Journalistic ethics and transparency
4. Potential biases or conflicts of interest

Respond ONLY with a JSON object in this exact format:

```
{
    "rating": "reliable" | "mixed" | "unreliable",
    "reason": "Brief 1-2 sentence explanation",
    "confidence": 0.0-1.0
}
```

B Confusion Matrix



Confusion Matrix (A4; ClaimCLAIRE)

C Error Analysis: Qualitative Examples

We present four representative examples from the AVeriTeC dev set illustrating ClaimCLAIRE's behavior across correct predictions and key failure modes identified in our error analysis.

Example 1: Correct Prediction

Claim: "The World Health Organization (WHO) says there is no proof that face masks protect against Covid-19."

Gold label: Refuted **Predicted:** Inconsistent **Correct:** ✓

Component verdicts:
(1) The WHO made a statement about face masks and Covid-19 → *refuted*
(2) The WHO stated there is "no proof" masks protect against Covid-19 → *refuted*

Analysis: ClaimCLAIRE correctly identifies that while a WHO representative cautioned that masks could give a "false sense of protection," this does not constitute a blanket statement that masks provide "no proof" of protection. Reliable sources including the CDC and WHO consistently recommended masks to slow transmission. Both components are refuted, yielding an **Inconsistent** verdict.

Table C1: Correct prediction: system accurately distinguishes a nuanced WHO caution from a sweeping denial of mask efficacy.

Example 2: False Positive — Insufficient Claim Context

Claim: "558 people were killed by the police in 2018, while 201 people died in police custody."

Gold label: Supported **Predicted:** Inconsistent **Correct:** ×

Component verdicts:
(1) 558 people were killed by police in 2018 → *refuted*
(2) 201 people died in police custody in 2018 → *unverified*

Analysis: ClaimCLAIRE retrieves reliable US sources (Washington Post, CNN) reporting approximately 992–1,165 police killings in the US in 2018, refuting component (1). However, the AVeriTeC annotators labeled this Supported based on South African police statistics, where the figures are accurate. Because the claim contains no geographic context, the system defaults to the US—the dominant web search result—and incorrectly flags the claim.

Error category: Insufficient context / vague claim phrasing.

Table C2: False positive: absent geographic context causes the system to retrieve evidence for the wrong country, producing a spurious **Inconsistent** verdict.

Example 3: False Positive — Semantic Precision

Claim: "Donald Trump said: 'Last month, I took on Big Pharma. I signed orders that would massively lower the cost of your prescription drugs.' "

Gold label: Supported **Predicted:** **Inconsistent** **Correct:** ×

Component verdicts:
(1) Trump made this statement → *verified*
(2) Trump took on Big Pharma in the prior month → *verified*
(3) Trump signed orders related to prescription drug costs → *verified*
(4) The orders would massively lower prescription drug costs → *refuted*

Analysis: Components (1)–(3) are verified by reliable sources. However, component (4) is refuted: reliable sources report that most people were unlikely to see drug cost savings from the orders. AND logic yields **Inconsistent**. The gold label is Supported because AVeriTeC annotators evaluate whether Trump *said* it, not whether the content is accurate. ClaimCLAIRE over-decomposes the quote, evaluating the factual accuracy of its claims rather than treating the utterance as the unit of verification.

Error category: Semantic precision / verdict threshold misalignment.

Table C3: False positive: the system correctly identifies a partially inaccurate assertion, but misaligns with annotators' intent to verify whether the quote was uttered rather than whether its content is true.

Example 4: False Negative — Circular Reporting

Claim: "Donald Trump said: 'When asked if he supports cutting police funding, Joe Biden replied, Yes, absolutely.' "

Gold label: Refuted **Predicted:** **Consistent** **Correct:** ×

Component verdicts:
(1) Trump made this statement → *verified*
(2) Biden said "Yes, absolutely" when asked about police funding → *verified*
(3) Trump's characterization of Biden's response is accurate → *verified*

Analysis: Reliable sources confirm Trump made the statement and that Biden did say "Yes, absolutely" in some context, so all three components are verified and the system yields **Consistent**. However, the gold label is Refuted because Biden's full remark concerned redirecting funds toward social services—not defunding police—making Trump's framing a misrepresentation. Multiple reliable sources repeat Trump's framing without correcting it, and the system lacks a mechanism to detect when a quote attribution is accurate but its implied meaning is misleading.

Error category: Circular reporting / semantic precision.

Table C4: False negative: the system verifies the literal quote attribution but misses that the framing misrepresents Biden's original context—a limitation of static trust-rating approaches when misinformation propagates through reliable outlets.