POLITIBENCH: A STRUCTURED EVIDENCE BENCHMARK FOR POST-TRAINING LLM IN FACT-CHECKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated fact-checking with large language models (LLMs) is dominated by frontier closed-source models, and prior attempts to specialize smaller models have not closed the gap yet. We ask whether the articles written by professional fact-checkers can serve as post-training data that teaches smaller LLMs to fact-check. We collect, process and convert 26,101 real-world PolitiFact fact-checking data into training instances that pair a contextualized claim with the fact-checker's analysis, stripped of verdict-revealing sentences, and with evidence extracted from fact-checking articles and organized into supporting facts, weakening facts, and missing context. Our post-training of Qwen3-4B and Qwen3.5-9B, with Group Relative Policy Optimization (GRPO) and our proposed reward function, raises six-class verdict accuracy over the base checkpoints by 5.6% and 7.4%, respectively, with particularly large accuracy gains on difficult ambiguous classes (e.g., *Half True* from 6.51% to 44.79% and *Mostly True* from 11.02% to 29.02%). In the setting that mirrors retrieval-augmented fact-checking in practical deployment, where only the claim and formatted evidence are available, the post-trained 9B model exceeds the best among the 5 frontier models, GPT-5.6 Luna, by 1.3%. When sufficient evidence is available, our post-trained models improve the accuracy of their base models by 7.1% and 5.8%, respectively, once again reaching the same level as GPT-5.6 Luna. The gains transfer to 11 external fact-checking evaluation cases, where our post-trained models outperform their base checkpoints in 80% of cases, and match or surpass frontier models on 8 of the 11 benchmarks under zero-shot settings. This verifies that the post-training on our datasets boosts the general fact-checking ability of small models. We release **PolitiBench**, the processed corpus, as a resource for both evaluating and specializing LLMs for evidence-grounded fact-checking.

1 INTRODUCTION

We increasingly ask large language models (LLMs) what is true, and the best automated fact-checkers are frontier models run by a few companies (Carvalho & Goldschmidt, 2026). That concentration is itself an epistemic risk since these models are highly persuasive and can raise belief in false claims when prompted to mislead (Costello et al., 2026; Hackenburg et al., 2025), they cannot be independently audited, and as they become the infrastructure through which societies reason, their errors propagate at scale (Casper et al., 2026; Marchal et al., 2026). Small open-weight models are a natural counterweight since they can be inspected, run locally, and deployed by newsrooms and civil-society groups at modest cost (Putta et al., 2025), and even a weaker independent model can materially blunt adversarial persuasion by a stronger one (Wachowiak et al., 2026). But a counterweight only works if it is nearly as good, and today small models are not. Frontier LLMs are the strongest verifiers across benchmarks, while smaller fine-tuned models lag on claims that require reasoning over evidence (Seo et al., 2025). We ask whether a small open model can be specialized to match frontier models at evidence-grounded fact-checking.

Prior attempts to specialize small models have trained on verdict labels attached to claims and evidence, whether from existing benchmarks or written by a larger LLM (Section 2.3). None uses what professional fact-checkers actually produce. A PolitiFact article is a worked example of verification that states the claim in context, gathers evidence from cited sources, weighs what supports and what

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

undercuts the claim, and explains the verdict. We ask whether such articles can serve as post-training data that teaches a small open model to fact-check as well as a frontier model.

To this end, we collect 26,101 English PolitiFact[1] articles from May 2007 to June 2026, each rated on the six-point Truth-O-Meter scale, and convert each into a training instance in three steps. We remove the ruling section and verdict-revealing sentences, following Politi-Fact-Only (Shukla et al., 2025). We contextualize each claim: PolitiFact-derived datasets usually keep only the short quoted statement, and many of these are ambiguous or unverifiable alone, rated false because of an accompanying video or true only for a given speaker and date (Pelrine et al., 2023; Thibault et al., 2025), so we prepend speaker, date, and venue metadata and resolve remaining references against the article, for claim decontextualization (Deng et al., 2024). Finally, we extract the article's evidence into supporting/weakening facts, and missing context, each with an importance score and, for supporting facts, a full or partial support label. A retrieval-augmented fact-checker (Tian et al., 2024) would hand its verifier evidence in this form and supplying it directly isolates the verification step.

We post-train Qwen3-4B and Qwen3.5-9B on this corpus with GRPO and our designed reward function, and compare against five frontier and large open models. In the retrieval-augmented setting, where the model sees only the contextualized claim and the structured evidence, post-training raises six-class accuracy by 5.6% and 7.4%, and the 9B model exceeds the best frontier model, GPT-5.6 Luna, by 1.3% (Table 2). With the fact-checker's article also available, accuracy gains are 7.1% and 5.8% and both models reach GPT-5.6 Luna's level. On ten external datasets in 11 zero-shot evaluation configurations (Table 3), post-training improves over the base checkpoints in 35 of 44 comparisons and matches or surpasses the best frontier model in eight, surpassing it outright on four. Since most of these benchmarks predate the frontier models' knowledge cutoff and may have been partially memorized during training (Thibault et al., 2025; He et al., 2026), the improvement over the base checkpoint is a more reliable signal of ability transfer. Nonetheless, a post-trained 9B model overtakes frontier models on several datasets is still notable. In addition, ablations over four training variants show that the full article with structured evidence outperforms structured evidence alone, the article with verdict cues removed, and the raw article with all cited sources: this indicates that what the model learns from is the fact-checkers' own reasoning and evidence organization. Our main contributions are as follows:

- We show that articles written by professional fact-checkers are an effective post-training signal. RL post-training on PolitiFact articles improves six-class accuracy of 4B and 9B models by 5.6% and 7.4% in the retrieval-augmented setting and by 7.1% and 5.8% with the article available; across four training variants, the full article works best, indicating that ***human fact-checkers' reasoning and evidence organization are the best signal***.
- We show that small post-trained models can match or surpass frontier models. The post-trained 9B model reaches 51.40% accuracy and 43.18% balanced accuracy in the retrieval-augmented setting, exceeding GPT-5.6 Luna by 1.30% and 0.66%, and matches or surpasses frontier models in eight of 11 external evaluation configurations, despite shifts in domain and label set.
- We release **PolitiBench**: 26,101 PolitiFact fact-checks (2007–2026) with six-way verdicts, metadata, contextualized claims, leakage-filtered analyses, and evidence organized into supporting facts, weakening facts, and missing context, with training and evaluation code, for evaluating and specializing LLMs in evidence-grounded fact-checking.

2 Related Works

We review three lines of work: the datasets used to evaluate fact-checking, the LLM pipelines built on them, and attempts to specialize smaller models for the task.

2.1 Fact-Checking Datasets

Benchmarks differ in where claims come from, what evidence is assumed, and how fine the labels are (Table 1). LIAR collected naturally occurring political claims from PolitiFact with speaker metadata and six-way verdicts (Wang, 2017), and LIAR-PLUS added the journalist's justification paragraph from each article (Alhindi et al., 2018). These datasets keep the fine-grained verdicts of professional fact-checking but stop at the quoted statement: they neither supply the evidence the fact-checker

[1] https://politifact.com/

Table 1: Comparison of fact-checking benchmarks. PolitiBench provides the most recent naturally occurring claims, contextualized to be verifiable on their own, together with the evidence needed to verify them, with verdict cues removed.

Dataset	Date coverage	# Claims	Claim aug.	Meta data	Evidence provenance	Evidence format	leakage control	Labels
LIAR	2007-2016	∼12.8k	×	✓	None	None	×	6
LIAR-PLUS	2007-2016	∼12.8k	×	✓	PolitiFact articles	Extracted justifications	✓	6
FEVER	Synthetic	185.4k	×	×	Wikipedia	Evidence sentences	×	3
HoVer	Synthetic	∼26k	×	×	Wikipedia	Multi-hop sentences	×	2
MultiFC	1995-2018	∼34.9k	×	✓	Retrieved web sources	Documents and snippets	×	Var.
AVeriTeC	2016-2025	4,568	✓	✓	Open web	QA pairs and justifications	✓	4
PolitiBench	**2007-2026**	26,101	✓	✓	PolitiFact articles	*Extracted Categorized facts*	✓	6

used nor resolve the context a statement depends on. A substantial share of claims in LIAR-style datasets are ambiguous or unverifiable on their own, and a statement can be rated false because of the video it accompanied, or true only for a particular speaker and date (Pelrine et al., 2023; Thibault et al., 2025). Therefore a model judged against the bare quotation is partly being scored on information it was never given. FEVER and HoVer instead construct claims against Wikipedia with three- or two-way labels (Thorne et al., 2018; Jiang et al., 2020); their evidence is controlled and reproducible but their claims are synthetic and their label sets coarse. MultiFC (Augenstein et al., 2019) and AVeriTeC (Schlichtkrull et al., 2023) move back toward real-world claims from many fact-checking organizations, with web evidence as question–answer pairs and controlling for evidence published after the claim. This is the closest to PolitiBench but much smaller and not sufficient for post training. A further concern for any benchmark older than the models evaluated on it is contamination. Most of these datasets predate current frontier models, and their claims and verdicts may be in the pretraining data (Thibault et al., 2025; He et al., 2026). PolitiBench extends coverage to June 2026, so a portion of the claims post-date the training data of models we evaluate.

2.2 LLM-Based Fact-Checking Pipelines

Recent systems use LLMs at every stage of verification. AVeriTeC shared-task entries reformulate claims into search queries or verification questions and reason over retrieved web evidence (Schlichtkrull et al., 2024); HerO, Fathom, and SANCTUARY generate hypothetical documents or questions to improve retrieval before filtering noisy web results (Yoon et al., 2024; Rashid & Hakak, 2025; Dharmavaram & Hakak, 2025), and Graph-of-Thoughts refines the information need iteratively (Rolinger & Liu, 2025). Systems differ mainly in how retrieved evidence reaches the verifier: as long documents in a two-stage RAG pipeline (Ullrich & Drchal, 2025), as semantically filtered passages (Upravitelev et al., 2025), as question–answer pairs, or as claim–evidence relations modeled explicitly (Tan et al., 2025); agentic systems further decompose the claim and track which components remain unresolved (Liu & Sharma, 2026). Retrieval agents such as Tian et al. (2024) show that with good evidence, the verifier is the bottleneck, and DeVerna et al. (2025) find that even frontier models with web search need curated context to be reliable on political claims. When evidence is drawn from professional fact-check articles, leakage becomes a concern since the article's ruling and analysis can reveal the label, and removing them, as Politi-Fact-Only does, lowers scores accordingly (Shukla et al., 2025). Here, we focus on the verifier when the retrieval pipeline works perfectly and provides the required evidences in the structured form, and also follow Politi-Fact-Only in stripping verdict cues from the article text. This allows us to isolate and improve the verification/reasoning step.

2.3 Specializing Models for Fact-Checking

Fine-tuned encoders were the standard approach before LLMs, and fine-tuned transformers set the state of the art on most misinformation benchmarks at the time, outperforming more elaborate architectures (Pelrine et al., 2021). On PolitiFact-derived data, classifiers trained on LIAR or LIAR-PLUS

have long been the reference baselines (Wang, 2017; Alhindi et al., 2018), and the most recent such study trains ModernBERT (Warner et al., 2025) on 18k PolitiFact claims with speaker context and retrieved web evidence, only to find that larger LLMs used zero-shot consistently surpass it (Sahitaj et al., 2025). Seo et al. (2025) reach a similar conclusion across 14 benchmarks where frontier LLMs with few-shot prompting are the strongest verifiers, and small fine-tuned models are weakest on claims that require multi-hop reasoning over evidence, improving only when trained on synthetic multi-hop data.

Fine-tuning generative models has relied on either existing benchmark labels or supervision written by a larger model. FactLLaMA instruction-tunes LLaMA on LIAR and RAWFC with retrieved evidence (Cheung & Lam, 2023); MiniCheck trains small verifiers on GPT-4-generated examples of grounding errors (Tang et al., 2024). More recently, reinforcement learning has been applied to fact verification with rewards for retrieval quality and verdict correctness. ProFact trains a single policy end-to-end over decomposition, retrieval, and verdict on AVeriTeC (Yang et al., 2026), and R3Check learns iterative retrieval and structured reasoning for complex claims (Qi et al., 2026). In all of these, the training signal is a verdict label attached to a claim and some evidence; none uses the reasoning a professional fact-checker wrote when reaching that verdict.

Fact-check articles themselves have been used in two ways. JustiLM distills fact-check articles into a retrieval-augmented T5 model so that it can generate justifications without the article at test time (Zeng & Gao, 2024), and PrimeFacts decomposes 13k PolitiFact articles into decontextualized evidence premises that improve zero-shot retrieval and verification (Sahitaj et al., 2026). Both treat the article as a source of evidence or justification text. To our knowledge, no prior work uses the article, with its analysis, evidence, and verdict, as post-training data for a generative model that must predict the verdict, nor asks whether a small model trained this way can match frontier models.

3 POLITIBENCH

Our pipeline transforms PolitiFact records into contextualized claims and formatted evidence for verdict prediction. Figure 2 presents its principal stages. We first collect fact-check records and associated metadata from PolitiFact, then retrieve and extract content from the external sources cited in each article. Claims are augmented using their available metadata, while the extracted source content is organized into supporting facts, weakening facts, and missing context. Finally, the resulting claim–evidence representation is used to predict claim veracity under multiple label granularities. In the following subsection, we will introduce the details of each component.

3.1 DATA COLLECTION

We crawled articles accessible through PolitiFact's "Latest" listing up to June 29, 2026. We retained English-language fact-checks evaluated using PolitiFact's six-category Truth-O-oMeter scale, excluding Spanish-language articles and articles using other rating systems to maintain a consistent task definition and label space. For each retained article, we stored the original HTML and used Python-based parsing scripts to extract the claim, verdict, and associated metadata, including the speaker, publication date, statement context, and article description. We also extracted the quick summary, fact-checker's analysis, and final ruling as separate fields. The preservation of these components separately allows the original records to be retained, while controlling which information is supplied to downstream models.

To reduce explicit verdict leakage, we applied two stages of filtering. First, we excluded the final ruling section from model inputs. Because article structures vary and rating statements may appear outside this section, we additionally removed sentences containing both predefined rating-related keywords and the article's gold verdict label. The original article and gold verdict were retained separately for documentation and evaluation. This procedure targets explicit disclosures of the rating but does not guarantee the removal of indirect verdict cues.

3.2 CLAIM AUGMENTATION

Naturally occurring claims are often short and context-dependent rather than fully standalone. They may contain ambiguous references to people, organizations, policies, events, or time periods that are only identifiable from the surrounding article context. We therefore augment each claim us-

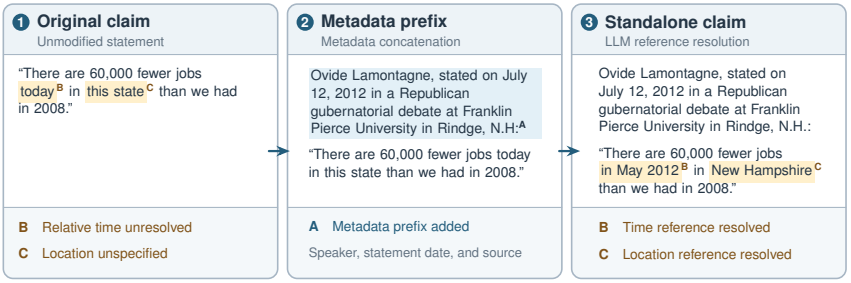

Figure 1: Overview of the claim augmentation example.

ing the associated PolitiFact metadata and contextual information to produce a more self-contained statement for downstream evidence extraction and verdict prediction.

The augmentation procedure consists of two stages. First, available metadata including the speaker, statement date, and source or setting of the statement are concatenation to the original claim. This stage provides attribution and contextual information without modifying the original claim's wording. Second, GPT-5.6 Luna receives the metadata-prefixed claim associated article context and is instructed to minimally resolve any remaining context-dependent expressions. An expression is replaced only when its referent is unclear within the prefixed claim and uniquely established by the supplied context. We refined the augmentation instructions through preliminary development experiments to address recurring cases of unnecessary rewriting and changes to the original assertion. The resulting prompt emphasizes resolving contextual ambiguity while preserving the claimant's meaning, even when the assertion is misleading or factually incorrect.

We used few-shot demonstrations adapted from Snopes[2] fact-checks to illustrate the intended editing behavior. Four demonstrate resolving contextual references, while four preserve claims that are already self-contained. These examples guide the model to clarify ambiguity without correcting the assertion or adding information relevant only to its factual assessment.

The model is instructed not to introduce evidence, explanations, qualifications, verdicts, or factual corrections. Its output consists solely of the resulting metadata-prefixed claim, without commentary or formatting. If the input is already standalone, or the context does not establish a unique referent with sufficient confidence, the model is instructed to return the input exactly unchanged. These constraints are intended to limit semantic drift, unnecessary augmentation, and verdict leakage. The complete prompt template and eight demonstrations are provided in the Appendix B.

3.3 Structured Evidence Extraction

The evidence-extraction stage transforms the available textual material into a compact, structured representation for verdict prediction. The extractor receives the augmented proposition, defined as the claim text after augmentation, excluding the metadata prefix added during the first augmentation step. This input is provided alongside the preprocessed PolitiFact analysis text, from which ruling-related material has been removed to prevent leakage of the target verdict.

We use few-shot prompting with two synthetic claim–article pairs and their corresponding structured evidence outputs. The first demonstrates the distinction between weakening evidence and missing context, including an empty supporting evidence category. The second illustrates partial support, where evidence supports the direction of a claimed change while challenging its magnitude and causal attribution. Both examples demonstrate the required output structure and importance scores. The prompt and template along with figures and examples can be found in Appendix C.

The extracted evidence is divided into three categories. *Supporting facts* are evidence showing that the claim is accurate, whereas *weakening facts* contradict, disprove, or seriously weaken it. *Missing context* identifies omitted information that materially changes how the claim should be interpreted. Each evidence item contains a textual fact and an importance score assigned by the same extractor during extraction. The importance score is scored from 1 to 5, where 1 indicates minor relevance, 2 indicates useful but weak evidence, 3 indicates important evidence, 4 indicates strong evidence,

[2]https://www.snopes.com/

and 5 indicates evidence considered decisive for assessing the claim. The facts are further sorted in decreasing order of importance within each category

Supporting facts are additionally assigned a support type. A fact is classified as *full* only when it independently supports the complete central claim. A *partial* designation is used when the evidence supports only an individual number, quotation, date, person, premise, or subclaim. This distinction prevents evidence supporting a single claim component from being interpreted as verification of the entire statement.

Finally, we format the extracted evidence into a standardized textual representation for downstream verdict prediction. Facts are grouped by category and ordered by decreasing importance score within each category. Support types are retained explicitly, while importance scores are used only to determine the ordering and are omitted from the formatted text. The complete prompt template and examples are provided in Appendix C

We evaluate evidence extraction limits of $k \in 2, 3, 4$, where k specifies the maximum number of facts the model is instructed to extract per evidence category. The model may return fewer than k facts when insufficient relevant evidence is available. Comparing these settings allows us to examine whether a compact evidence set is sufficient for verdict prediction or whether a larger extraction budget improves downstream performance. The results are shown in the following section.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Preliminary Experiments We generated all model-generated benchmark components using GPT-5.6 Luna. We then conducted a set of preliminary experiments on a validation set of 2,004 claims to compare claim representations and evidence extraction limits k using the regenerated components. The results in Table 5 in Appendix C.1, guided the configuration selected for subsequent training experiments.

A further analysis was done on the augmentation process. The first augmentation step adds a metadata prefix to every claim, yielding 100% coverage. In the second step, approximately 40% of claims undergo further editing to solve contextual reference and improve standalone interpretability, while the remaining 60% retain their prefixed form unchanged. This editing rate describes how frequently the augmentation model modifies the claim beyond prefix addition.

Table 5 compares different combinations of claim and evidence configuration for six-class verdict prediction on a validation set of 2,004 examples. We consider two claim representations: the original claim and the augmented standalone claim. Both representations are paired with the formatted evidence extracted during benchmark construction. We vary $k \in \{2, 3, 4\}$, where k specifies the maximum number of evidence items extracted per category. Preliminary experiments favor an extraction limit of $k = 2$: retaining more evidence items does not consistently improve verdict prediction. Claim augmentation benefits GPT across the evaluated settings, while its effects on Qwen are small and mixed. We therefore use the original claim with formatted evidence extracted at $k = 2$ for subsequent Qwen post-training, following its validation performance. Detailed results and evidence-count analyses appear in Appendix C.

Post-Training We post-train Qwen3-4B-Thinking-2507[3] and Qwen3.5-9B[4] with Group Relative Policy Optimization (GRPO) (Shao et al., 2024), starting from the released checkpoints downloaded from Hugging Face, without an additional supervised fine-tuning stage. We train for two to three epochs and observe little further improvement after the second epoch. The best checkpoints are used for all experiments. The reward prioritizes verdict correctness: predictions that match the gold label receive scores between 0.60 and 1.00, while incorrect predictions receive at most 0.20. Within these ranges, auxiliary terms account for valid answer and analysis blocks, analysis length, keyword overlap with the claim and evidence, and the occurrence of predefined reasoning markers. The length term gives full credit to analyses of 20–180 words. These terms encourage the requested

[3]https://huggingface.co/Qwen/Qwen3-4B-Thinking-2507

[4]https://huggingface.co/Qwen/Qwen3.5-9B

response structure while ensuring that an incorrect verdict always scores below a correct one. See more details in Appendix D.4.

Unfiltered RL (U-RL) uses the training examples without preprocessing, retaining any sentences, regardless of whether it might reveal the final verdict. In Appendix D.1, we also consider filtered RL (F-RL), which removes verdict-revealing sentences from the training inputs. For all training variants, verdict-revealing sentences are removed from the validation inputs before evaluation. Thus, "filtered" and "unfiltered" refer only to the preprocessing of the training data. The formatted-evidence variant uses the same training inputs as U-RL and adds a field containing supporting facts, weakening facts, and missing context. All original input fields are retained unchanged, and the model is trained independently from the released checkpoint. We randomly split the PolitiFact data into 23512 training examples[5] and 2,612[6] test examples. We perform full-parameter training, both models use the AdamW (Loshchilov & Hutter, 2017) optimizer with a constant learning rate of 1×10^{-6} and no warmup. The training batch size is 256 prompts, with four responses sampled per prompt, yielding 1024 generated sequences per batch. We set the policy-update mini-batch size to 32 and the per-GPU micro-batch size to 1, with dynamic batching enabled. Training responses are sampled with temperature 1.0, top-$p = 1.0$, and no top-k restriction. The maximum input and response lengths are 2,048 and 4,096 tokens, respectively. We use a random seed of 42 for both training runs.

4.2 Results and Validation

Abbreviation and Labels We use the following input labels in Tables 2. **Claim** denotes the original claim, while **Augmented Claim** denotes its context-enriched version. **Article** comprises the sanitized article analysis, description, and short summary, with explicit verdict-revealing sentences removed. **Formatted Ev.** refers exclusively to the formatted evidence extracted from the article. Accordingly, **Claim + Article** provides the original claim and the three article fields; **Claim + Article + Formatted Ev.** provides the original claim, the three article fields, and the formatted evidence. In contrast, **Aug. Claim + Formatted Ev. (retrieval-augmented setting)** provides only the augmented claim and formatted evidence, excluding the article analysis, description, and short summary. Thus, the last two configurations differ in both claim representation and access to the article fields. The task uses the six PolitiFact labels: True, Mostly True, Half True, Mostly False, False, and Pants on Fire. Each result is reported as **Accuracy / Balanced Accuracy** (%), where balanced accuracy is the macro-average recall across the six classes. $k = 2$ was used for all cases related to formatted evidence.

Baseline Results on PolitiBench We evaluate all baseline models in Table 2 in a zero-shot setting, without web access or external retrieval. All models use a shared verdict-prediction prompt template, with the claim and evidence fields adjusted according to the input setting. We retain each model's default temperature. Models are instructed to output one of the six PolitiFact verdict labels within an `<answer>...</answer>` block, from which the prediction is extracted. We report accuracy and balanced accuracy, with the latter defined as the macro-average recall across the six classes.

Table 2 shows that providing article content together with formatted evidence yields the strongest performance across all evaluated baselines. Accuracy ranges from 57.88% to 62.56% in this setting, compared with 25.92%–36.91% for claim-only prediction, while balanced accuracy increases from 24.14%–31.66% to 53.45%–57.31%. Claim augmentation provides more modest gains, improving balanced accuracy for all five models and accuracy for four, with the Qwen baseline being the exception. GPT-5.6 Luna achieves the highest accuracy and balanced accuracy in all five input settings, reaching 62.56% and 57.31%, respectively, with article content and formatted evidence. JEV 1.13 performs closely in this setting, achieving 62.29% accuracy and 57.08% balanced accuracy. These results suggest that resolving contextual ambiguity can assist verdict prediction, while supplying substantive evidence provides substantially greater benefits.

Post-trained Models on PolitiBench *U-RLF (Unfiltered RL + formatted evidence)* is independently trained from the base model using the claim, formatted evidence, article analysis, article description, and short summary. The 4B and 9B prefixes indicate the approximate parameter scale of the corresponding model backbone.

[5] 23474 for formatted evidence dataset

[6] 2607 for formatted evidence dataset

Table 2: PolitiFact six-class results: accuracy / balanced accuracy (%). With augmented claims and formatted evidence, 9B Unfiltered RL exceeds the best available frontier scores by **1.30% / 0.66%**. Δ denotes the metric-wise best 4B/9B score minus the best available frontier score (pp).

Model	Claim	Claim + Article	Augmented Claim	Claim + Article + Formatted Ev.	Aug. Claim + Formatted Ev.
Frontier baselines					
JEV 1.13	31.97 / 28.86	54.48 / **49.93**	33.46 / 31.59	62.29 / 57.08	48.98 / 42.47
GPT-4o-mini	27.07 / 26.45	46.67 / 43.24	29.79 / 28.66	57.88 / 53.45	47.37 / **42.52**
Qwen 3.8 27B	33.73 / 25.45	50.77 / 46.34	31.58 / 26.81	58.42 / 53.99	46.99 / 40.28
GPT-5.6 Luna	**36.91 / 31.66**	**54.67** / 49.06	**40.89 / 37.64**	**62.56 / 57.31**	**50.10** / 41.77
GPT-6 Luna Pro	25.92 / 24.14	52.79 / 47.35	34.53 / 30.37	59.15 / 54.86	48.91 / 42.23
4B/9B checkpoints: Base and post-trained					
4B Base (no RL)	33.77 / 19.92	46.55 / 36.91	32.73 / 20.18	53.59 / 45.37	42.08 / 31.18
4B Unfiltered RL	32.47 / 25.72	51.72 / 45.24	30.63 / 27.37	60.64 / 54.36	47.64 / 39.25
4B U-RLF	31.38 / 25.55	49.98 / 44.23	31.45 / 26.27	60.72 / 55.76	46.53 / 38.74
9B Base (no RL)	30.49 / 23.81	47.99 / 41.48	32.68 / 26.06	56.92 / 50.90	43.96 / 36.75
9B Unfiltered RL	**35.64 / 27.94**	**52.68 / 45.78**	**36.33 / 29.02**	62.56 / 56.73	**51.40 / 43.18**
9B U-RLF	34.48 / 24.95	50.44 / 43.26	35.71 / 28.72	**62.72 / 57.06**	48.37 / 40.26
Δ (pp)	$-1.27/-3.72$	$-1.99/-4.15$	$-4.56/-8.62$	$+0.16/-0.25$	$+1.30/+0.66$

Providing richer evidence allows the post-trained smaller models to approach the performance of larger models. Under the Article + Formatted Ev. setting, the formatted-evidence-trained 4B model achieves 60.72% accuracy and 55.76% balanced accuracy, compared with 62.72% and 57.06% for its 9B counterpart. Their accuracy gap decreases from 3.10% under claim-only evaluation to 2.00% with article content and formatted evidence. Both models also approach the GPT-5.6 Luna baseline reported in Table 2, which achieves 62.56% accuracy and 57.31% balanced accuracy in the corresponding input setting. These results suggest that access to relevant evidence can partially compensate for differences in model capacity. When information is insufficient, we observe that smaller models tend to favor the *False* label. The low balanced accuracy in claim-only evaluation further indicates uneven performance across verdict classes. Supplying article content and formatted evidence substantially improves balanced accuracy, suggesting more effective discrimination across the six verdict categories. Also, one can note that compared with claim-only prediction, providing article content together with extracted evidence improves accuracy by approximately 20% for the base models and nearly 28% for the post-trained models, highlighting the greater benefit of evidence-rich inputs after task-specific adaptation. Moreover, when expert-written analyses are unavailable, Augmented Claim + Formatted Evidence represents a more realistic input setting. In this setting, our post-trained 9B model outperforms the best evaluated frontier baseline by 1.30% in accuracy, further supporting the effectiveness of formatted information organization.

Table 3: External evaluation of six checkpoints and four reference models. Each cell reports accuracy / balanced accuracy (%). Δ reports the metric-wise best score among the six 4B/9B checkpoints minus the best reference-model score, in percentage points (pp).

Dataset	4B			9B			Reference models				Δ (pp)
	Base	U-RL	U-RLF	Base	U-RL	U-RLF	Jev 1.13	GPT-6 Luna	GPT-5.6 Luna	Qwen 27B	Acc. / Bal. Acc.
Snopes (C)	27.70/27.70	30.20/30.20	**31.00/31.00**	32.10/32.10	**34.70/34.70**	28.40/28.40	34.60/34.60	47.00/47.00	**50.80/50.80**	32.80/32.80	$-16.10/-16.10$
Snopes (C+E)	46.40/46.40	54.80/54.80	**57.10/57.10**	50.50/50.50	**60.50/60.50**	55.10/55.10	60.50/60.50	**62.90/62.90**	60.10/60.10	54.20/54.20	$-2.40/-2.40$
ClaimReview 2025Q4	42.06/39.28	32.19/**40.47**	37.51/34.52	54.05/35.17	56.83/33.06	**58.60/36.12**	37.18/37.05	44.17/**50.32**	40.51/46.71	**48.72**/35.34	$+9.88/-9.85$
SciFact	80.24/81.73	**82.89/82.36**	79.65/80.47	76.40/78.20	82.01/82.94	**83.78/84.92**	**82.89/83.09**	68.14/67.44	72.57/72.68	77.29/78.26	$+0.89/+1.83$
ClimateCheck	65.91/64.76	**70.54**/66.44	67.96/**68.59**	65.86/64.26	65.81/70.06	**67.02/70.37**	**66.91/63.16**	58.82/48.81	62.50/55.60	65.28/62.01	$+3.63/+7.21$
RealFactBench	67.71/58.68	**69.43/65.80**	66.71/62.50	69.43/59.55	**72.14/64.31**	71.00/60.74	73.71/64.45	**84.00/83.98**	83.00/82.91	70.71/60.84	$-11.86/-18.18$
SciClaimHunt	78.20/69.81	78.10/72.99	77.30/**76.81**	81.50/71.64	**84.30/78.84**	83.80/77.04	**81.70/69.54**	80.50/66.47	80.00/68.25	81.60/68.91	$+2.60/+9.30$
QuanTemp++	62.92/45.03	55.32/**47.67**	52.81/47.28	50.98/44.98	59.46/**49.93**	**62.85**/49.36	57.97/48.56	62.17/50.70	61.56/**50.80**	**62.92**/46.78	$0.00/-0.87$
AVeriTeC dev	76.03/59.35	77.17/**70.15**	**80.14**/65.37	78.77/61.06	**86.30/65.35**	81.96/59.53	**83.79/71.53**	63.47/65.26	70.09/65.94	72.83/62.77	$+2.51/-1.38$
TSVer	69.11/50.04	**73.17/50.12**	69.92/49.04	64.23/45.17	70.73/49.41	**73.98/52.24**	69.92/51.52	65.04/**58.74**	**73.17**/58.26	69.92/49.53	$+0.81/-6.50$
FACTors 2024+	28.00/23.67	27.75/**25.21**	26.12/22.29	29.00/24.84	**29.75/25.93**	29.00/24.66	32.88/29.17	**37.50/34.96**	36.38/34.88	32.12/26.41	$-7.75/-9.03$

Notes. Black bold values with pale shading mark non-Base results attaining the metric-wise group maximum: **pale blue for 4B**, **pale orange for 9B**, and **pale purple for reference models**. Within each backbone, Base, U-RL, and U-RLF form one comparison group; all non-Base ties at the maximum are highlighted. For Δ, the six-checkpoint comparison includes Base. Accuracy and balanced accuracy are maximized independently and may therefore select different models. C: claim only; C+E: claim with supplied evidence. Dataset configurations appear in Table 10.

Data Preprocessing for Fair Comparison on External Datasets The following summarizes our evaluation subsets and the transfer behavior observed in our experiments. Further details on the datasets are provided in Section E.1.

The training on PolitiFact focuses on the prediction of label truthfulness. However, on other fact-checking datasets, the meanings of labels can have large mismatches with PolitiFact, which can harm the fair comparison of fact checking ability by introducing confounding ability variables, such as general language understanding. For example, some datasets use *SUPPORTED, REFUTED*, and *NEI* to describe the relations between a claim and the supplied evidence instead of truthfulness. The latter labels do not represent truthfulness levels and therefore pose undesirable difficulties for our model. In addition, SciClaimHunt's *NEGATIVE* includes both contradiction and insufficient evidence, while QuanTemp++'s *CONFLICTING* maps its native Half True/False label and should not be assumed identical to conflicting-evidence categories elsewhere. Therefore, we design a modified prompt to mitigate the confounding ability variables by adding an explanation of the labels, and better support truthfulness prediction for our post-trained model.

Following error analysis, we tested a common modified prompt in E.2 on SciFact and ClimateCheck that explicitly checks entities, quantities, dates, negation, and scope, and distinguishes missing information from explicit contradiction. It also clarifies the output format. The prompt was selected on separate development examples and applied uniformly to all models, with evaluation examples and labels fixed. Its mixed effects suggest that decision instructions can mitigate some transfer errors, but do not establish label misunderstanding as their sole cause.

Evaluation on Difficult Classes and Transferability As shown in Table 3, across 11 external zero-shot evaluation settings, post-training improves the mean of accuracy and balanced accuracy in 35 out of 44 comparisons against the corresponding base checkpoints. For example, 9B U-RL increases accuracy from 50.50% to 60.50% on Snopes with supplied evidence, and from 50.98% to 59.46% on the non-PolitiFact subset of QuanTemp++. On SciClaimHunt, the same model improves accuracy from 81.50% to 84.30% and balanced accuracy from 71.64% to 78.84%. These gains across different sources, domains, and label schemes suggest that PolitiFact post-training develops transferable fact checking capabilities.

Post-training yields particularly large improvements on ambiguous categories. On PolitiFact, with augmented claims and formatted evidence, 4B U-RL increases Half True recall from 6.51% to 44.79%, surpassing frontier models on this class. These gains extend to several difficult labels in external datasets, defined as those with baseline class-wise recall below 50%. On Snopes with supplied evidence, 9B U-RL improves Mostly True recall from 16.40% to 50.00% and Mostly False recall from 8.80% to 26.00%. Beyond Snopes, 4B U-RL raises SUPPORTED recall on RealFact-Bench from 30.93% to 54.66% under claim-only evaluation, while 9B U-RL improves POSITIVE recall on SciClaimHunt from 49.64% to 66.67% with supplied evidence. Each highlighted comparison also improves class F1 and overall accuracy, supporting transfer to selected difficult categories beyond PolitiFact. However, gains vary across labels and models, and recall declines for extreme verdicts such as Pants on Fire in some cases.

5 Conclusions

We present PolitiBench, a benchmark combining explicit claim provenance, resolved contextual references, filtered fact-checking analyses, and evidence organized into supporting facts, weakening facts, and missing context. Task-specific post-training substantially improves 4B and 9B models over their base checkpoints, including on ambiguous labels such as Mostly True and Mostly False. With informative evidence, the 9B model achieves performance comparable to GPT-5.6 Luna on six-class PolitiFact evaluation, while the 4B model remains competitive. Gains transfer across datasets, with smaller models surpassing frontier baselines in several settings, including when only augmented claims and formatted evidence are provided. These findings highlight the importance of evidence availability and organization alongside model capacity.

AI USE STATEMENT

Large language models were used for claim augmentation and structured evidence extraction, as described in Section 3, and as verifiers in the reported experiments. AI assistance was also used for manuscript drafting, language editing, and reference checking. Model-generated benchmark content may contain errors and should not be treated as independently verified evidence. The authors are responsible for the manuscript, experimental claims, and final verification of references.

ETHICS STATEMENT

PolitiBench is derived from publicly accessible fact-checking articles and retains the source organizations' verdicts as task labels rather than treating them as universally authoritative judgments. Source selection and annotation practices may introduce political, geographic, and temporal biases, while generated claim augmentations and evidence summaries may omit context or distort the original material. Incorrect automated verdicts can mislead readers or harm the reputations of people and organizations discussed in the claims. The benchmark is intended to support research and human review; its results do not establish that the models are suitable for autonomous high-stakes decisions. Public accessibility does not by itself authorize unrestricted redistribution: any release of source-derived material should respect applicable permissions, attribution requirements, and privacy considerations.

REPRODUCIBILITY STATEMENT

Section 3 describes data collection, filtering, claim augmentation, and evidence extraction. The appendices provide augmentation and extraction prompts and examples (Appendices B and C), the fixed reward definition (Appendix D.4), and external-evaluation subsets, decoding settings, and prompt variants (Appendices E.1 and E.2). Training hyperparameters and the random seed are reported in the experimental settings. Results are point estimates for the reported runs; they do not quantify variation across independent training seeds. Reproduction involving hosted models or changing web sources may be affected by service updates and source availability.

REFERENCES

Raia Abu Ahmad, Aida Usmanova, and Georg Rehm. The ClimateCheck dataset: Mapping social media claims about climate change to corresponding scholarly articles. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pp. 42–56. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.sdp-1.5. URL https://aclanthology.org/2025.sdp-1.5/.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. Where is your evidence: Improving fact-checking by justification modeling. In James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal (eds.), *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pp. 85–90, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5513. URL https://aclanthology.org/W18-5513/.

Enes Altuncu, Can Başkent, Sanjay Bhattacherjee, Shujun Li, and Dwaipayan Roy. FACTors: A new dataset for studying the fact-checking ecosystem. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2025. doi: 10.1145/3726302.3730339. URL https://doi.org/10.1145/3726302.3730339.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4685–4697, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1475. URL https://aclanthology.org/D19-1475/.

Luis Henrique M Carvalho and Ronaldo Ribeiro Goldschmidt. Large language models for automated fact-checking: A systematic literature review. *IEEE Access*, 2026.

Stephen Casper, Jonathan Stray, Anna Gausen, Cameron Jones, Brian Christian, Bálint Gyevnár, Siao Si Looi, Joshua Levy, Elizabeth Seger, Hannah Kirk, Kobi Hackenburg, Matt Kowal, Michelle ., Dan Zhao, Zhonghao He, Luke Hewitt, Hause Lin, Maarten Sap, Dylan Hadfield-Menell, Atoosa Kasirzadeh, Gordon Pennycook, Jean-François Godbout, Kellin Pelrine, Reihaneh Rabbany, Natasha Jaques, Jasmine Li, Yoshua Bengio, David Rand, Thomas Costello, and Mick Yang. Ai epistemic risks: Emerging mechanisms & evidence. 2026. URL https://ssrn.com/abstract=6867358.

Tsun-Hin Cheung and Kin-Man Lam. Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 846–853. IEEE, 2023.

Thomas H. Costello, Kellin Pelrine, Matthew Kowal, Jasper Timm, Antonio A. Arechar, Jean-François Godbout, Adam Gleave, David Rand, and Gordon Pennycook. Large language models can effectively convince people to believe conspiracies, 2026. URL https://arxiv.org/abs/2601.05050.

Zhenyun Deng, Michael Schlichtkrull, and Andreas Vlachos. Document-level claim extraction and decontextualisation for fact-checking. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11943–11954, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.645. URL https://aclanthology.org/2024.acl-long.645/.

Matthew R. DeVerna, Kai-Cheng Yang, Harry Yaojun Yan, and Filippo Menczer. Large language models require curated context for reliable political fact-checking – even with reasoning and web search, 2025. URL https://arxiv.org/abs/2511.18749.

Arbaaz Dharmavaram and Saqib Hakak. SANCTUARY: An efficient evidence-based automated fact checking system. In Mubashara Akhtar, Rami Aly, Christos Christodoulopoulos, Oana Cocarascu, Zhijiang Guo, Arpit Mittal, Michael Schlichtkrull, James Thorne, and Andreas Vlachos (eds.), *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pp. 247–257, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 978-1-959429-53-1. doi: 10.18653/v1/2025.fever-1.19. URL https://aclanthology.org/2025.fever-1.19/.

Kobi Hackenburg, Ben M. Tappin, Luke Hewitt, Ed Saunders, Sid Black, Hause Lin, Catherine Fist, Helen Margetts, David G. Rand, and Christopher Summerfield. The levers of political persuasion with conversational ai, 2025. URL https://arxiv.org/abs/2507.13919.

Haorui He, Xinwen Chen, Dacheng Wen, Reynold Cheng, Francis C. M. Lau, and Yupeng Li. Novel claim or déjà vu? rethinking "contamination-free" dynamic evaluation for multimodal automated fact-checking. In *Proceedings of the ACM International Conference on Multimedia*, 2026. URL https://arxiv.org/abs/2607.23514.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. HoVer: A dataset for many-hop fact extraction and claim verification. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3441–3460, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.309. URL https://aclanthology.org/2020.findings-emnlp.309/.

Sujit Kumar, Anshul Sharma, Siddharth Hemant Khincha, Gargi Shroff, Sanasam Ranbir Singh, and Rahul Mishra. SciClaimHunt: A large dataset for evidence-based scientific claim verification, 2025. URL https://arxiv.org/abs/2502.10003.

Xinman Liu and Mayank Sharma. ClaimCLAIRE: A trust-aware multi-component fact-checking agent for open-world claims. In Kai-Wei Chang, Ninareh Mehrabi, Satyapriya Krishna,

Anubrata Das, Jwala Dhamala, Yang Trista Cao, Tharindu Kumarage, Anil Ramakrishna, Christos Christodoulopoulos, Yixin Wan, Aram Galystan, Anoop Kumar, and Rahul Gupta (eds.), *Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026)*, pp. 73–91, San Diego, California, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-418-7. doi: 10.18653/v1/2026.trustnlp-main.6. URL https://aclanthology.org/2026.trustnlp-main.6/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Nahema Marchal, Stephanie Chan, Matija Franklin, Manon Revel, Geoff Keeling, Roberta Fischli, Bilva Chandra, and Iason Gabriel. Architecting trust in artificial epistemic agents. *arXiv preprint arXiv:2603.02960*, 2026.

Kellin Pelrine, Jacob Danovitch, and Reihaneh Rabbany. The surprising performance of simple baselines for misinformation detection. In *Proceedings of the Web Conference 2021*, WWW '21, pp. 3432–3441, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3450111. URL https://doi.org/10.1145/3442381.3450111.

Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6399–6429, 2023.

Akshith Reddy Putta, Jacob Devasier, and Chengkai Li. ClaimCheck: Automatic fact-checking of textual claims using web evidence. In Weijia Shi, Wenhao Yu, Akari Asai, Meng Jiang, Greg Durrett, Hannaneh Hajishirzi, and Luke Zettlemoyer (eds.), *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pp. 303–316, Albuquerque, New Mexico, USA, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-229-9. doi: 10.18653/v1/2025.knowledgenlp-1.26. URL https://aclanthology.org/2025.knowledgenlp-1.26/.

Peng Qi, Yuyang Zhao, Wynne Hsu, and Mong Li Lee. R3check: Reinforcement learning for iterative retrieval and structured reasoning in complex fact checking. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1473–1484, 2026.

Farrukh Bin Rashid and Saqib Hakak. Fathom: A fast and modular RAG pipeline for fact-checking. In Mubashara Akhtar, Rami Aly, Christos Christodoulopoulos, Oana Cocarascu, Zhijiang Guo, Arpit Mittal, Michael Schlichtkrull, James Thorne, and Andreas Vlachos (eds.), *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pp. 258–265, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 978-1-959429-53-1. doi: 10.18653/v1/2025.fever-1.20. URL https://aclanthology.org/2025.fever-1.20/.

Sascha Rolinger and Jin Liu. Graph-of-thoughts for fact-checking with large language models. In Mubashara Akhtar, Rami Aly, Christos Christodoulopoulos, Oana Cocarascu, Zhijiang Guo, Arpit Mittal, Michael Schlichtkrull, James Thorne, and Andreas Vlachos (eds.), *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pp. 266–273, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 978-1-959429-53-1. doi: 10.18653/v1/2025.fever-1.21. URL https://aclanthology.org/2025.fever-1.21/.

Premtim Sahitaj, Iffat Maab, Junichi Yamagishi, Jawan Kolanowski, Sebastian Möller, and Vera Schmitt. Towards automated fact-checking of real-world claims: Exploring task formulation and assessment with LLMs. In *Proceedings of the 5th Workshop on Reducing Online Misinformation through Credible Information Retrieval (ROMCIR 2025), co-located with ECIR 2025*, 2025. URL https://arxiv.org/abs/2502.08909.

Premtim Sahitaj, Jawan Kolanowski, Ariana Sahitaj, Veronika Solopova, Max Upravitelev, Daniel Röder, Iffat Maab, Junichi Yamagishi, Sebastian Möller, and Vera Schmitt. From articles to

premises: Building PrimeFacts, an extraction methodology and resource for fact-checking evidence. In Stelios Piperidis, Núria Bel, Henk van den Heuvel, Nancy Ide, Simon Krek, and Antonio Toral (eds.), *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 7719–7731, Palma de Mallorca, Spain, May 2026. ELRA Language Resource Association. doi: 10.63317/453datkp6z9s. URL https://aclanthology.org/2026.lrec-1.613/.

Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos. The automated verification of textual claims (AVeriTeC) shared task. In Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos (eds.), *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pp. 1–26, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.fever-1.1. URL https://aclanthology.org/2024.fever-1.1/.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 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*, 2023. URL https://openreview.net/forum?id=fKzSz0oyaI.

Wooseok Seo, Seungju Han, Jaehun Jung, Benjamin Newman, Seungwon Lim, Seungbeen Lee, Ximing Lu, Yejin Choi, and Youngjae Yu. Verifying the verifiers: Unveiling pitfalls and potentials in fact verifiers. 2025. URL https://arxiv.org/abs/2506.13342.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Satyam Shukla, Himanshu Dutta, and Pushpak Bhattacharyya. Recon, answer, verify: Agents in search of truth. In Saloni Potdar, Lina Rojas-Barahona, and Sebastien Montella (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 2429–2448, Suzhou (China), November 2025. Association for Computational Linguistics. ISBN 979-8-89176-333-3. doi: 10.18653/v1/2025.emnlp-industry.167. URL https://aclanthology.org/2025.emnlp-industry.167/.

Snopes. Snopes. https://www.snopes.com/. Source website for the Snopes evaluation subset.

Marek Strong and Andreas Vlachos. TSVer: A benchmark for fact verification against time-series evidence. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29906–29926. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1519. URL https://aclanthology.org/2025.emnlp-main.1519/.

Xin Tan, Bowei Zou, and Ai Ti Aw. Improving explainable fact-checking with claim-evidence correlations. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 1600–1612, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.108/.

Liyan Tang, Philippe Laban, and Greg Durrett. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8818–8847, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.499. URL https://aclanthology.org/2024.emnlp-main.499/.

Camille Thibault, Jacob-Junqi Tian, Gabrielle Péloquin-Skulski, Taylor Lynn Curtis, James Zhou, Florence Laflamme, Luke Yuxiang Guan, Reihaneh Rabbany, Jean-François Godbout, and Kellin Pelrine. A guide to misinformation detection data and evaluation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5801–5809, 2025.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Jacob-Junqi Tian, Hao Yu, Yury Orlovskiy, Tyler Vergho, Mauricio Rivera, Mayank Goel, Zachary Yang, Jean-François Godbout, Reihaneh Rabbany, and Kellin Pelrine. Web retrieval agents for evidence-based misinformation detection. *arXiv preprint arXiv:2409.00009*, 2024.

Herbert Ullrich and Jan Drchal. AIC CTU@FEVER 8: On-premise fact checking through long context RAG. In Mubashara Akhtar, Rami Aly, Christos Christodoulopoulos, Oana Cocarascu, Zhijiang Guo, Arpit Mittal, Michael Schlichtkrull, James Thorne, and Andreas Vlachos (eds.), *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pp. 274–280, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 978-1-959429-53-1. doi: 10.18653/v1/2025.fever-1.22. URL https://aclanthology.org/2025.fever-1.22/.

Max Upravitelev, Premtim Sahitaj, Arthur Hilbert, Veronika Solopova, Jing Yang, Nils Feldhus, Tatiana Anikina, Simon Ostermann, and Vera Schmitt. Exploring semantic filtering heuristics for efficient claim verification. In Mubashara Akhtar, Rami Aly, Christos Christodoulopoulos, Oana Cocarascu, Zhijiang Guo, Arpit Mittal, Michael Schlichtkrull, James Thorne, and Andreas Vlachos (eds.), *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pp. 229–237, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 978-1-959429-53-1. doi: 10.18653/v1/2025.fever-1.17. URL https://aclanthology.org/2025.fever-1.17/.

Venktesh V, Deepali Prabhu, and Avishek Anand. A benchmark for open-domain numerical fact-checking enhanced by claim decomposition, 2025. URL https://arxiv.org/abs/2510.22055.

Lennart Wachowiak, Scott D Blain, David Williams-King, and Samuele Marro. Llm wardens: Mitigating adversarial persuasion with third-party conversational oversight. *arXiv preprint arXiv:2605.08321*, 2026.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 422–426, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2067. URL https://aclanthology.org/P17-2067/.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 2526–2547, 2025.

Rongxin Yang, Shenghong He, Siyuan Zhu, and Chao Yu. From verdict to process: Agentic reinforcement learning for multi-stage fact verification. *arXiv preprint arXiv:2606.13262*, 2026.

Shuo Yang, Yuqin Dai, Guoqing Wang, Xinran Zheng, Jinfeng Xu, Jinze Li, Zhenzhe Ying, Weiqiang Wang, and Edith C. H. Ngai. RealFactBench: A benchmark for evaluating large language models in real-world fact-checking. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025. URL https://arxiv.org/abs/2506.12538.

Yejun Yoon, Jaeyoon Jung, Seunghyun Yoon, and Kunwoo Park. HerO at AVeriTeC: The herd of open large language models for verifying real-world claims. In Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos (eds.), *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pp. 130–136, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.fever-1.15. URL https://aclanthology.org/2024.fever-1.15/.

Fengzhu Zeng and Wei Gao. JustiLM: Few-shot justification generation for explainable fact-checking of real-world claims. *Transactions of the Association for Computational Linguistics*, 12: 334–354, 2024. doi: 10.1162/tacl_a_00649. URL https://aclanthology.org/2024.tacl-1.19/.

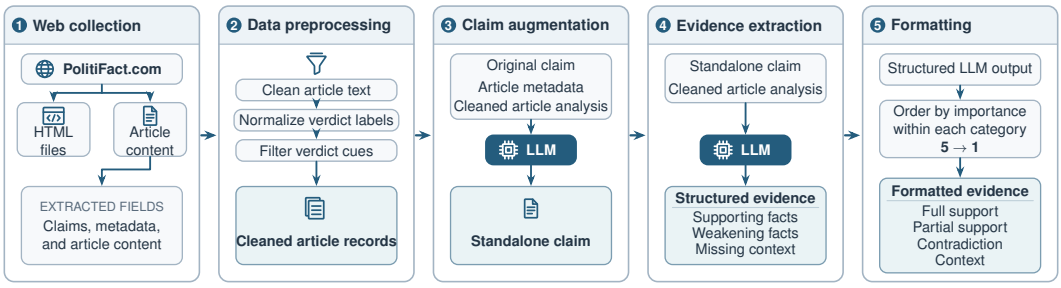

Figure 2: The following figure shows the process of generation of the different benchmark components. From web collection to claim augmentation and end with evidence extraction and formatting.

A Future Work

Building on these results, an important next step is to improve how specialized models handle incomplete evidence and unfamiliar verification settings. The observed tendency to favor False when information is limited highlights the need to distinguish insufficient support from evidence of falsity. Our cross-dataset results further show that transfer depends on domain, language, and label semantics, motivating training objectives that explicitly address evidence sufficiency and uncertainty. Although the additional gains from formatted-evidence training over ordinary RL are modest in the current experiments, component-level ablations could clarify which aspects of evidence organization provide the most useful training signals. Future work should also extend the approach to multilingual settings and integrate evidence retrieval and source assessment, moving toward complete fact-checking systems that can acquire and evaluate their own evidence.

B Claim Augmentation

The following prompt was used to augment the claims.

```
System :
You are a claim-augmentation editor. Resolve contextual ambiguities so the claim can be
↪  understood without the surrounding article, while preserving exactly the proposition
↪  being asserted.

User :
Augment the CLAIM only when information from the CONTEXT
is necessary to make the claim standalone and unambiguous.

WHEN AUGMENTATION IS APPROPRIATE

Resolve a context-dependent expression when its referent
is uniquely established by the CONTEXT. This includes:

- pronouns or demonstratives without a clear referent,
  such as they, them, it, this or those;
- vague noun phrases, such as this bill, the company,
  the policy, the decision or the incident;
- unnamed people, organizations, laws, programs,
  documents, events, places or comparison targets;
- relative time expressions, such as today, yesterday,
  last year or two years ago;
- shortened or incomplete names that do not uniquely
  identify their referent without the article.

Make an edit only when a specific expression has no clear referent
within the CLAIM itself. Replace that expression with its uniquely
supported referent from the CONTEXT and add no other specificity.
Otherwise, return the CLAIM unchanged.

PRESERVATION RULES

- Use only information contained in the CLAIM or explicitly supported
by the CONTEXT.
- Preserve the proposition's meaning exactly.
- Preserve quantities, polarity, certainty, modality, causality,
```

```
comparisons, attribution and time scope.
- Preserve the claimant's wording and characterization, even if it is
informal, awkward, exaggerated, misleading or factually incorrect.
- When resolving an ambiguity inside quoted text, replace only the
context-dependent expression and preserve all other quoted wording.
- Do not correct the claim using the article's evidence or conclusion.
- Do not add explanations, evidence, qualifications, exceptions,
verdicts or background information.
- Do not expand a shortened name or abbreviation merely because the
CONTEXT provides a fuller or more technical name. Expand it only
when the original expression is genuinely ambiguous.
- Do not replace a pronoun when its referent is already clear within
the same sentence. Do not correct spelling.
- First-person references such as I, me and my are normally clear when
the attribution prefix identifies the speaker and should remain
unchanged.
- Do not replace what a claim says an image, video, quotation or
document depicts with the article's conclusion about its
authenticity.
- If the claim is already standalone, return it exactly unchanged.
- If the CONTEXT does not establish exactly one referent with sufficient
confidence, return the claim exactly unchanged.

OUTPUT

Return only one complete, continuous claim. Do not include a label,
explanation, notes or formatting around it.

{examples_text}

NOW PROCESS THE FOLLOWING INPUT

CLAIM:
<claim>
{statement}
</claim>

CONTEXT:
<context>
{context}
</context>

RESULT:
```

We evaluate whether claim augmentation makes statements sufficiently self-contained for fact-checking using a model-assigned feasibility rating. The evaluator assesses whether a statement can be evaluated given only its text and access to web search, without determining its truthfulness. Ratings range from 0 (not assessable due to irrecoverable ambiguity, missing information, or the absence of a factual claim) to 1 (ambiguous but potentially recoverable through search) and 2 (sufficiently clear to assess).

Table 4 shows that GPT-Luna augmentation increases the proportion of statements rated sufficiently clear to assess from 57.04% to 79.19%, a gain of 22.15 percentage points. The proportion rated unassessable decreases from 21.00% to 7.88%, while the mean rating increases from 1.360 to 1.714. These results highlight a key benefit of augmentation: supplying context that makes the intended proposition easier to identify and evaluate. This improvement in model-assessed feasibility complements the downstream verdict-prediction results, although it does not by itself establish factual correctness or preservation of the original claim's meaning.

Table 4: Feasibility ratings for original claims and claims augmented by GPT-Luna. Percentages and mean ratings are calculated over valid ratings.

Rating	**Original**	**Augmented**
0: Not assessable	546 (21.00%)	205 (7.88%)
1: Ambiguous	571 (21.96%)	334 (12.85%)
2: Assessable	1,483 (57.04%)	2,059 (79.19%)
Total examples	2,600	2,600
Valid ratings	2,600	2,598
Mean rating	1.360	1.714

```
System: You assess whether a statement contains enough
information for its veracity to be evaluated.
You are not determining whether the statement
is true or false. You are determining whether
a valid truth assessment is possible from the
statement alone, assisted by web search.

Follow the requested output format exactly:
brief explanation | rating

The rating must be exactly 0, 1, or 2.
Do not add any text after the rating.

User: The following statement is going to be given
to an AI system to determine if it is true or
false and write an explanation why.

Statement: "{text}"

The only thing the AI will be given is the
statement itself, as written above|no context,
visuals, or any other information. Your task
is to assess whether the AI could possibly
give a valid answer. This is not about how
likely the AI is to give the correct answer,
but whether it is possible to evaluate the
veracity of the statement from the
information given.

The AI will have access to web search to find
primary and secondary sources, but evaluation
might still be impossible if too much
ambiguity or mandatory context is missing.

Information whose absence may make evaluation
difficult includes:

1. The identity of a key person, such as an
unidentified speaker or ambiguously
referenced person.
2. A location, when the statement's veracity
depends on the location.
3. Textual information or evidence mentioned
in the statement but not supplied.
4. Visual or audio evidence mentioned in the
statement. The AI receives only the statement
text.
5. Temporal information. The date on which
the statement was made is unknown. This does
not matter if the statement can be evaluated
regardless of when it was made.
6. An actual factual claim for which
evaluating veracity makes sense.

Rate the statement using this scale:

2: There may be some room for interpretation
or contextual influence, but the statement
is still sufficiently clear and unambiguous.
For example, "The Earth is round" is not
perfectly precise, but it can reasonably be
evaluated.

1: There is clear ambiguity, missing context,
or more than one potential interpretation.
However, there seems to be roughly a one-half
chance of recovering the intended meaning or
context through a strong knowledge base or
web search.

0: It is impossible to evaluate. Multiple
valid interpretations would substantially
change the verdict, mandatory and
irrecoverable information is missing, or the
statement contains no factual claim or is
nonsensical.

First give a brief explanation. Then write
a vertical bar "|", followed by the rating
as a number alone.

Required format:
```

```
brief explanation | rating
```

The following eight few-shot demonstrations illustrate contextual reference resolution and cases where the input should remain unchanged. Each demonstration contains a metadata-prefixed claim, supplied context, and expected output. The text is reproduced from the demonstration configuration; the contexts should not be interpreted as verbatim excerpts from the linked articles.

B.1 Example 1: Policy Reference

Claim. Social media users, stated on July 23, 2026 in posts: "198 Democrats voted against this bill."

Context. Posts circulated after a July 22 vote in the U.S. House of Representatives. The House passed H.R. 7008, formally named the Stop Insider Trading Act, by a vote of 232-198. All 198 opposing votes came from Democratic representatives.

The legislation concerned stock trading by members of Congress and their spouses. The article also discussed disagreements about the bill's restrictions, exemptions and other provisions. In the quoted post, "this bill" referred specifically to the Stop Insider Trading Act.

Expected output. Social media users, stated on July 23, 2026 in posts: "198 Democrats voted against the Stop Insider Trading Act."

B.2 Example 2: Pronoun References

Claim. Donald Trump, stated on January 21, 2026 in a World Economic Forum speech: "Until the last few days when I told them about Iceland, they loved me."

Context. President Donald Trump delivered the statement during an address at the World Economic Forum in Davos. The surrounding remarks concerned the United States, Greenland and its relationship with the North Atlantic Treaty Organization.

Immediately before and after the quoted sentence, Trump discussed NATO members and their reaction to his position. In this sentence, both "them" and "they" referred to NATO. The article separately examined whether Trump mistakenly said Iceland when discussing Greenland, but the spoken word in the quotation was Iceland.

Expected output. Donald Trump, stated on January 21, 2026 in a World Economic Forum speech: "Until the last few days when I told NATO about Iceland, NATO loved me."

B.3 Example 3: Relative Time

Claim. Social media users, stated on August 17, 2026 in posts: "Donald Trump announced today that sailors aboard the USS Abraham Lincoln must sign nondisclosure agreements before leaving the ship."

Context. Social media posts circulating that day claimed President Donald Trump had announced a restriction affecting sailors aboard the USS Abraham Lincoln. According to the posts, the sailors would not be allowed to leave the aircraft carrier until they signed legally binding nondisclosure agreements.

The article investigated the alleged announcement using publicly available statements, reporting about the ship and information attributed to Navy officials. Regardless of whether the alleged announcement was genuine, "today" in the circulating claim referred to August 17, 2026.

Expected output. Social media users, stated on August 17, 2026 in posts: "Donald Trump announced on August 17, 2026 that sailors aboard the USS Abraham Lincoln must sign nondisclosure agreements before leaving the ship."

B.4 Example 4: Unnamed Entity

Claim. Social media users, stated on August 20, 2026 in posts: "The company started distributing $5 billion in cash to its automobile-insurance customers."

Context. Posts and news reports discussed a large distribution being made to automobile-insurance customers. The company at the center of the reports was State Farm Mutual Automobile Insurance Company, commonly known as State Farm.

The article explained that State Farm had announced a distribution totaling approximately $5 billion and described which customers could receive money and how amounts would be calculated. It also distinguished the distribution from an ordinary refund. In the quoted claim, "the company" referred to State Farm.

Expected output. Social media users, stated on August 20, 2026 in posts: "State Farm started distributing $5 billion in cash to its automobile-insurance customers."

B.5 Example 5: Unchanged Despite Additional Qualifications

Claim. Social media users, stated on August 11, 2026 in posts: Massachusetts legalized abortion after 24 weeks of pregnancy.

Context. Posts claimed that Massachusetts had legalized abortion after 24 weeks of pregnancy. The article examined legislation enacted in August 2026 and compared it with the state's previous abortion law.

The reporting described the circumstances under which the law permitted an abortion after 24 weeks and explained how the legislation changed existing language. It also discussed disagreements over descriptions such as "until birth." Although those details may affect a verdict, the original claim already identifies the jurisdiction, action, subject and relevant time period without relying on an unresolved reference.

Expected output. Social media users, stated on August 11, 2026 in posts: Massachusetts legalized abortion after 24 weeks of pregnancy.

B.6 Example 6: Unchanged Despite Contradictory Evidence

Claim. Social media users, stated on January 12, 2026 in posts: A leaked NASA document called "Project Anchor" says Earth will lose gravity for seven seconds on August 12, 2026.

Context. Online posts described an allegedly secret NASA project named Project Anchor. They claimed that Earth would temporarily lose gravity for seven seconds on August 12, 2026 and that NASA knew about the approaching event.

The article searched NASA materials, examined the supposed document and discussed a solar eclipse scheduled for that date. It found no scientific mechanism or authentic NASA announcement supporting the gravity claim. That evidence is relevant to fact-checking, but the claim itself already names the organization, alleged document, event, duration and exact date. It therefore requires no augmentation.

Expected output. Social media users, stated on January 12, 2026 in posts: A leaked NASA document called "Project Anchor" says Earth will lose gravity for seven seconds on August 12, 2026.

B.7 Example 7: Unchanged When References Are Already Clear

Claim. Social media users, stated on August 21, 2026 in posts: U.S. Sen. Ron Johnson once said, "Senior citizens have misused Social Security and that's why I want it to sunset."

Context. Social media posts attributed a statement about Social Security to Wisconsin Sen. Ron Johnson. The posts claimed that Johnson said senior citizens had misused Social Security and that he therefore wanted the program to sunset.

The article examined speeches, interviews, congressional records and other publicly available material to determine whether Johnson had made the statement. It also reviewed remarks Johnson had made about changing how federal programs such as Social Security are funded and reconsidered.

Within the attributed quotation, "it" refers directly to Social Security, which is named earlier in the same sentence. The first-person pronoun "I" refers to Johnson, who is identified in the attribution.

Expected output. Social media users, stated on August 21, 2026 in posts: U.S. Sen. Ron Johnson once said, "Senior citizens have misused Social Security and that's why I want it to sunset."

B.8 Example 8: Unchanged Image-Authenticity Assertion

Claim. Social media users, stated in early 2026 in posts: An image authentically shows Trump smiling in a white tank top while young girls kneel before him.

Context. Social media users circulated an alleged black-and-white photograph in early 2026. Captions accompanying the image claimed it had appeared among files connected with the federal investigation of Jeffrey Epstein.

The image purported to show U.S. President Donald Trump smiling and wearing a white tank top while several young girls kneeled in front of him. The article examined the image's provenance, searched for an authentic archival version and considered visual indications that it had been artificially generated.

The article's findings concerned whether the image was authentic. Throughout the article and the accompanying posts, the surname "Trump" referred to Donald Trump.

Expected output. Social media users, stated in early 2026 in posts: An image authentically shows Trump smiling in a white tank top while young girls kneel before him.

C EVIDENCE EXTRACTION

The following prompt was used for evidence extraction.

```
System :
Extract concise factual evidence from the article for fact-checking.
Use only the article. Do not invent facts. Do not give a verdict label.
Return valid JSON only.

User :
Statement:
{statement}

Article:
{article}

Return JSON with exactly these keys:
{
  "supporting_facts": [],
  "weakening_facts": [],
  "missing_context": []
}

Each supporting_facts item must be:
{ "fact": "", "importance": 1, "support_type": "full" }

Each weakening_facts and missing_context item must be:
{ "fact": "", "importance": 1 }

Return at most {k} facts per category. Prefer fewer decisive facts.
Each fact must be one short sentence, maximum 25 words.

Definitions:
- supporting_facts: independent article facts that show the statement is accurate, not
↪  merely that the statement was said.
- weakening_facts: article facts that contradict, disprove, or seriously weaken the
↪  statement.
- missing_context: omitted article facts that change interpretation.

Importance: 1=minor, 3=important but not decisive, 5=decisive.

"Rules:
- support_type is only for supporting_facts and must be 'full' or 'partial'.
- Use 'full' only if the fact independently supports the entire statement's main
↪  relationship.
- Use 'partial' only for support of a number, quote, date, person, premise, or subclaim.
- Never include quoted repetitions of the claim in supporting_facts.
- If the article says the speaker made the exact statement being checked, that is not
↪  supporting evidence. Omit it unless the claim is only about whether the person said
↪  those words.
- A fact is support only if it verifies the claim independently of the speaker's own
↪  statement.
- If a fact says the article, evidence, pledge, data, or record does NOT mention or does
↪  NOT show the relationship claimed in the statement, put it in weakening_facts, not
↪  supporting_facts.
- Do not include 'X said/claimed/stated...' in supporting_facts unless the checked
↪  statement is only about whether X said those words.
- If the speaker merely states, repeats, defends, alleges, or explains their own claim,
↪  omit it.
- Only include a quote as support if the statement is specifically about whether that quote
↪  was said.
- Evidence about long-term troop presence does not support a claim about long-term war
↪  unless the article says it is combat or war.
```

```
- Do not fill categories just to reach the limit.
- Omit background about who made the claim, social media spread, or fact-check flags unless
↪  essential.
- Sort each list from highest importance to lowest.

Return strictly valid JSON only. Do not omit commas between fields.
```

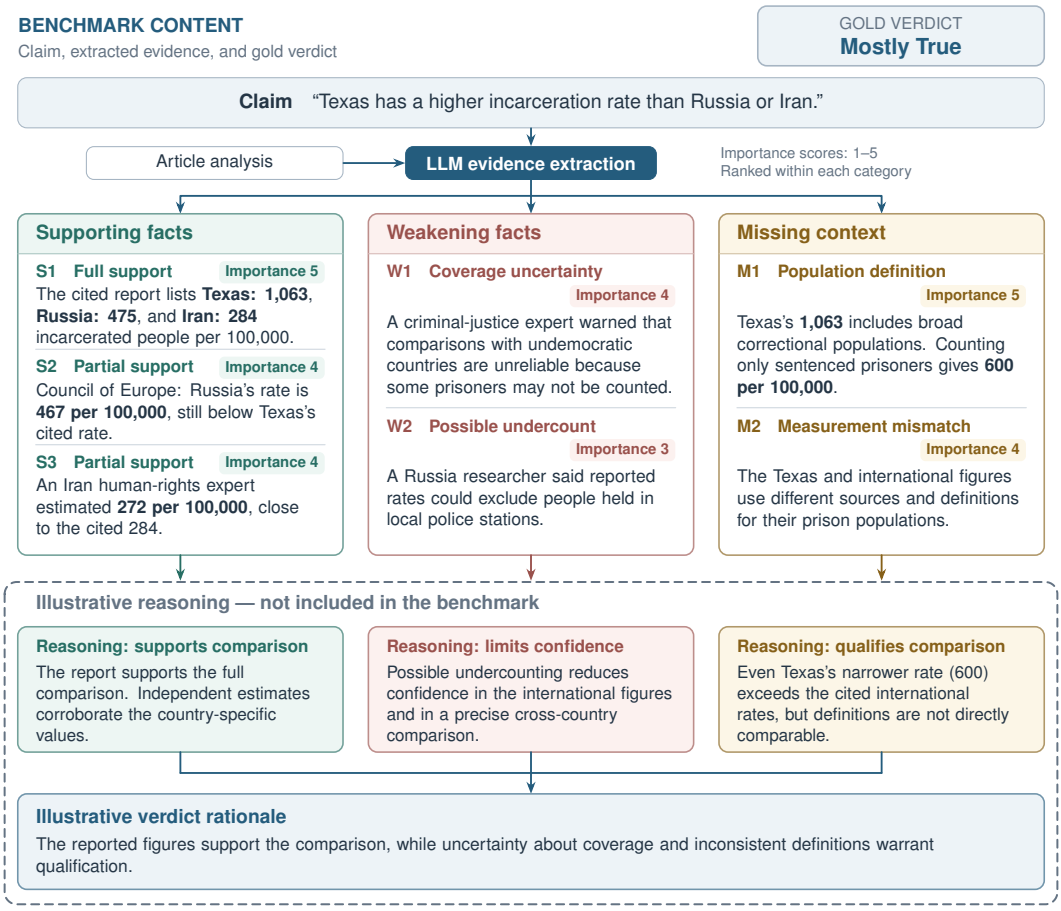

Figure 3: Overview of an evidence example.

C.1 Few-Shot Evidence Extraction Examples

The evidence extraction prompt includes the following two synthetic demonstrations. Each demonstration consists of a claim, an accompanying article, and the expected structured output. These examples were constructed for illustration and were not taken from published fact-checking articles. Importance scores range from 1 (minor) to 5 (decisive).

C.1.1 Example 1: Weakening Evidence and Missing Context

This example distinguishes evidence contradicting the claim from context explaining the reported spending figure. It also demonstrates an empty supporting evidence category.

Input

```
Claim:
The state spent $50 million on a benefits website that never worked.

Article:
During a campaign speech, Jordan Lee said the state spent $50 million on a benefits website
↪  that never worked. The original contract authorized spending of up to $50 million, but
↪  state invoices show that $8.7 million had been paid. The website experienced several
↪  outages during its first month. State records show that it nevertheless processed more
↪  than 120,000 benefit applications during its first year. Officials later approved
↪  additional repairs to improve its reliability.
```

Expected output

```
{
  "supporting_facts": [],
  "weakening_facts": [
    {
      "fact": "State records show the website processed more than 120,000 benefit applications
      ↪  during its first year.",
      "importance": 5
    },
    {
      "fact": "State invoices show that $8.7 million had been paid, not $50 million.",
      "importance": 5
    }
  ],
  "missing_context": [
    {
      "fact": "The $50 million figure was the contract's maximum authorized amount rather than
      ↪  actual spending.",
      "importance": 4
    }
  ]
}
```

C.1.2 Example 2: Partial Support and Weakening Evidence

This example illustrates how evidence can support one aspect of a claim while weakening others. The observed unemployment decline provides partial support for its direction, but does not establish the claimed magnitude or causal attribution.

Input

```
Claim:
The policy reduced unemployment by 20% in one year.

Article:
The governor repeated the claim that the policy reduced unemployment by 20%. State labor data
↪  show that unemployment fell from 6.0% to 5.4% during the year after the policy began. That
↪  represents a 10% reduction, not 20%. Economists interviewed for the article said the
↪  available data could not attribute the entire decline to the policy.
```

Expected output

```
{
  "supporting_facts": [
    {
      "fact": "State labor data show unemployment fell after the policy began.",
      "importance": 3,
      "support_type": "partial"
    }
  ],
  "weakening_facts": [
    {
      "fact": "Unemployment fell by 10%, not the claimed 20%.",
      "importance": 5
    },
    {
      "fact": "Economists could not attribute the entire unemployment decline to the policy.",
      "importance": 4
    }
  ],
  "missing_context": []
}
```

This example illustrates the structured and textual representations of the same extracted evidence. We use a claim from the test set, evaluated by PolitiFact in March 2016:[7]

> "Texas has a higher incarceration rate than Russia or Iran."

The dataset's gold verdict is *Mostly True*. It is reported here for reference and is not included in either evidence representation below.

[7] https://www.politifact.com/factchecks/2016/mar/23/gina-hinojosa/gina-hinojosa-vowing-reforms-says-texas-has-higher/

Structured evidence. At $k = 3$, the stored extraction contains three supporting facts, two weakening facts, and two context facts. The supporting category shares a single cap across full and partial support. Each fact has a model-assigned importance score from 1 (minor relevance) to 5 (decisive); supporting facts also have a support-type label. The table below presents the contents of `Extracted_Evidence`, retaining all facts and their metadata. Here, I denotes importance; a dash indicates that support type does not apply to that category.

Support type	I	Fact
`supporting_facts`		
Full	5	The cited report lists Texas at 1,063 incarcerated people per 100,000, compared with 475 in Russia and 284 in Iran.
Partial	4	The Council of Europe reports Russia's rate as 467 per 100,000, still below Texas's cited rate.
Partial	4	An Iran human-rights expert estimated Iran's rate at 272 per 100,000, close to the cited 284.
`weakening_facts`		
–	4	A criminal-justice expert said incarceration comparisons with undemocratic countries are unreliable because some prisoners may not be counted.
–	3	A Russia researcher said Russia's reported rate could understate incarceration by excluding people held in local police stations.
`missing_context`		
–	5	Texas's 1,063 rate includes broad correctional populations, while another measure counts only sentenced prisoners and reports 600 per 100,000 residents.
–	4	The comparison uses different sources and definitions for incarceration, including Texas correctional populations and international prison counts.

Formatted evidence. The paragraphs below reproduce the corresponding `Formatted_Evidence` value, with its headings typeset in bold. The formatter groups supporting facts by support type, maps `weakening_facts` to the heading *Contradiction*, and maps `missing_context` to *Context*. In this example, all seven facts are retained in their within-category order, while importance scores and JSON syntax are omitted. The resulting text contains the same factual content, with support types expressed as headings. The heading *Contradiction* is the formatter's label for the weakening category; these particular items qualify the reliability of the comparison rather than directly refute its numerical ordering.

> **Full support:** The cited report lists Texas at 1,063 incarcerated people per 100,000, compared with 475 in Russia and 284 in Iran.
>
> **Partial support:** The Council of Europe reports Russia's rate as 467 per 100,000, still below Texas's cited rate. An Iran human-rights expert estimated Iran's rate at 272 per 100,000, close to the cited 284.
>
> **Contradiction:** A criminal-justice expert said incarceration comparisons with undemocratic countries are unreliable because some prisoners may not be counted. A Russia researcher said Russia's reported rate could understate incarceration by excluding people held in local police stations.
>
> **Context:** Texas's 1,063 rate includes broad correctional populations, while another measure counts only sentenced prisoners and reports 600 per 100,000 residents. The comparison uses different sources and definitions for incarceration, including Texas correctional populations and international prison counts.

For both verifiers, the best-performing configuration uses $k = 2$. Qwen achieves its highest accuracy and Macro-F1 with the original claim, whereas GPT achieves its highest scores with the augmented claim. Increasing k provides no consistent improvement, with only marginal gains for Qwen when moving from $k = 2$ to $k = 3$ under claim augmentation. These results suggest that increasing the extraction budget does not necessarily improve downstream verdict prediction. Further analysis of evidence counts on the same validation set shows that increasing (k) from 2 to 4 raises the average number of evidence items per article from 3.43 to 5.47. Evidence counts are unequal across categories at all three extraction limits, with weakening evidence consistently having the highest

Table 5: PolitiFact six-class validation results on 2,004 examples. Each cell reports Accuracy / Macro-F1 (%). Here, k is the maximum number of evidence items per extraction category: supporting facts, weakening facts, and missing context. Supporting facts are formatted into full and partial support groups, which share a total budget of k. Bold indicates the best results for each verifier. The original-claim configuration with $k = 2$, which performs best for Qwen 4B, is selected for subsequent training experiments.

Verifier	Claim	$k = 2$	$k = 3$	$k = 4$
QWEN 3-4B	Original Claim + Formatted Evidence	**35.48 / 14.17**	35.08 / 13.27	35.13 / 13.67
	Augmented Claim + Formatted Evidence	35.08 / 13.57	35.18 / 14.02	34.53 / 12.18
GPT 5.6-LUNA	Original Claim + Formatted Evidence	40.30 / 38.55	40.10 / 38.31	38.90 / 37.35
	Augmented Claim + Formatted Evidence	**41.50 / 39.57**	40.15 / 38.07	40.25 / 38.44

average count. At (k=4), weakening evidence reaches the extraction limit in 36.7% of articles, compared with 4.0% for supporting evidence and 0.3% for missing context. These differences raise the possibility that verifiers are sensitive to the relative number of items across categories. However, the counts may reflect the available evidence, and further analysis is required to determine whether they contribute to the observed performance differences. The effect of claim augmentation also varies across verifiers. For GPT, augmented claims improve both accuracy and Macro-F1 at every evaluated value of k. At $k = 2$, augmentation increases accuracy from 46.95% to 48.10% and Macro-F1 from 39.11% to 40.64%. For Qwen, augmentation slightly reduces both metrics at $k = 2$ but slightly improves both at $k = 3$. Thus, additional claim context consistently benefits GPT in these experiments, while its effects on Qwen are small and mixed. For subsequent training experiments on Qwen 4B and 9B models, we select the original claim paired with formatted evidence a $k = 2$. This configuration achieves the highest validation accuracy and macrio-f1 for Qwen 4B.

The following prompt was used for both the unfiltered RL model and the unfiltered RL model with a local reward judge. A shorter prompt was used for downstream verdict prediction task used in Table 5

```
A conversation between User and Assistant. The user asks a question,
and the Assistant solves it. The assistant first thinks about the
reasoning process in the mind and then provides the user with the
answer.

User: You are given a PolitiFact claim and the article's analysis text.
Use the article analysis to examine the claim the way a fact-checker
would: identify what the claim says, compare it against the evidence
in the analysis, and decide the PolitiFact verdict.

Allowed verdict labels:
- True
- Mostly True
- Half True
- Mostly False
- False
- Pants on Fire

Claim: {claim}
Speaker: {speaker}
Statement context: {statement_context}
Statement date: {statement_date}

Article analysis and evidence:
{evidence}

Write your response in exactly this structure:
<analysis>
A brief fact-checking analysis that uses the article
analysis/evidence to examine the claim.
</analysis>
<answer>One allowed verdict label</answer>

Assistant: Let me examine the article analysis and the claim.
```

D POST TRAINING

D.1 FILTERED TRAINING DATASET

Table 6: 4B and 9B filtered RL results across four input settings. Each cell reports Accuracy / Balanced Accuracy (%). Formatted Ev. + Article includes the original claim, article fields, and formatted evidence. Bold marks the best scores in each row.

Model	Claim only	Claim + Article	Aug. Claim	Formatted Ev. + Article
4B Filtered RL	31.32 / 20.23	48.66 / 39.66	31.81 / 22.45	**57.12 / 49.36**
9B Filtered RL	30.40 / 24.81	45.41 / 37.90	23.85 / 23.83	**54.09 / 47.35**

For 4B, filtered training gives lower accuracy in all three settings, by 3.53–4.64 percentage points. The article-plus-formatted-evidence condition remains the strongest for both models. In this condition, the filtered model's recall drops from 63.54% to 52.08% on Half True and from 42.94% to 31.07% on Mostly True, while False recall increases from 91.14% to 94.25%. Thus, the decrease is accompanied by weaker discrimination of partially true claims and a stronger tendency to predict False.

For the 9B filtered model, formatted/extracted evidence gives the highest accuracy and balanced accuracy, 54.09% and 47.35%, improving over claim-only input by 23.69 and 22.54 percentage points, respectively. Augmented claims alone score below the original claims, indicating that claim augmentation does not replace the benefit of supplying evidence in this setting. In both cases, removing data leakage from the training dataset leads to a decrease in model performance. Interestingly, the 4B model outperforms the 9B model in the evidence-rich setting.

D.2 FULL CONTEXT TRAINING DATASET

We additionally trained Qwen3-4B-Thinking-2507 with article analysis and evidence together with the full article context. The run used 23,474 training examples. Table 7 reports the final checkpoint's independent evaluation on 2,612 held-out examples, with the 4B unfiltered RL results from Table 2 as a reference.

Table 7: 4B full-article-context training results and the unfiltered RL reference. Each cell reports Accuracy / Balanced Accuracy (%). Claim + Article uses sanitized article evidence in the independent evaluation.

Model	Claim only	Claim + Article	Aug. Claim
4B Unfiltered RL	32.47 / 25.72	51.72 / 45.24	30.63 / 27.37
4B Full-context RL	30.59 / 20.97	46.78 / 39.80	32.66 / 24.01

Across all evaluated input settings, the full-context model achieves lower balanced accuracy than the unfiltered model. One possible explanation is that longer article contexts make it harder for a small model to identify relevant evidence and learn the intended reasoning patterns. Irrelevant or redundant information may distract the model from important qualifications and contextual details, weakening its ability to distinguish between closely related verdict categories.

D.3 FORMATTED ONLY TRAINING DATASET

The original article evidence block is replaced by extracted facts organized into full support, partial support, contradiction, and context. The claim and its speaker, date, and statement-context metadata are retained. This variant uses 23,464 training examples. Table 8 reports the results on 2,607 evaluation examples, with the unfiltered RL model from Table 2 as a reference.

Formatted-only training achieves 57.65% accuracy and 51.62% balanced accuracy with formatted evidence and article content, below the unfiltered reference's 60.64% and 54.36%. With Claim + Article, accuracy also decreases from 51.72% to 47.30%. Although claim-only accuracy improves slightly, balanced accuracy is lower across all four settings.

Table 8: 4B formatted-only training results and the unfiltered RL reference. Each cell reports Accuracy / Balanced Accuracy (%). Column headings describe the evaluation inputs; Formatted Ev. + Article includes the original claim, article fields, and formatted evidence.

Model	Claim only	Claim + Article	Aug. Claim	Formatted Ev. + Article
4B Unfiltered RL	32.47 / 25.72	51.72 / 45.24	30.63 / 27.37	60.64 / 54.36
4B Formatted-only RL	33.56 / 23.54	47.30 / 39.62	30.46 / 22.25	57.65 / 51.62

One possible explanation is that extracting a compact set of facts can omit qualifications and contextual details needed to distinguish closely related verdict categories. In addition, article-based evaluation supplies longer, less structured inputs than those used during formatted-only training. This change in input representation may make it harder for the model to use the additional article information effectively. These observations suggest that organized evidence alone may not preserve all the information needed for fine-grained verdict prediction; they do not establish a causal effect of the evidence representation in isolation. However, combining article content with formatted evidence helps the model learn more effectively and yields the best performance among the evaluated input settings.

D.4 Fix Reward Function

The fixed reward combines verdict correctness with deterministic checks of response format and analysis content. Let $\hat{y}$ and y denote the predicted and gold verdicts. We extract the analysis and verdict from a final `<analysis>` block followed by an `<answer>` block containing one of the six allowed labels, with no subsequent non-whitespace text. Let f_a and f_v indicate successfully parsed analysis and answer blocks, respectively, and let k indicate an allowed predicted label. Responses that fail this joint parsing rule receive zero reward.

For a parsed analysis, the auxiliary quality score is

$$q = 0.30f_a + 0.25\ell + 0.25o + 0.20m, \tag{1}$$

where ℓ measures analysis length, o measures keyword overlap with the reference, and m measures the presence of predefined reasoning markers. For an analysis of n words, the length score is

$$\ell(n) = \begin{cases} n/20, & n < 20, \\ 1, & 20 \le n \le 180, \\ \max\left(0.5,\, 1 - \dfrac{n-180}{240}\right), & n > 180. \end{cases} \tag{2}$$

Let K_a and K_r be keyword sets extracted from the analysis and reference, respectively. The reference concatenates the claim, article analysis, short summary, and article description. Keyword extraction lowercases the text, removes a fixed stopword list, retains words of at least five characters, and keeps the first 120 unique keywords for the analysis and 80 for the reference. Let h count distinct predefined markers appearing in the lowercased analysis. The remaining scores are

$$o = \min\left(\frac{|K_a \cap K_r|}{8}, 1\right), \qquad m = \min\left(\frac{h}{4}, 1\right). \tag{3}$$

The marker list is: claim, evidence, analysis, according, shows, indicates, because, however, therefore, supports, contradicts, misleading, exaggerates, and partly. Each marker is counted at most once using substring matching.

The final reward is

$$R_{\text{fixed}} = \begin{cases} 0.62 + 0.13f_v + 0.25q, & \hat{y} = y, \\ \min(0.20,\, 0.03k + 0.03f_v + 0.14q), & \text{otherwise.} \end{cases} \tag{4}$$

For a valid parsed response, $f_a = f_v = k = 1$. This design ensures that every correct verdict receives a higher reward than any incorrect verdict, whose reward is capped at 0.20. The auxiliary terms encourage the requested structure and evidence-related analysis, but they are heuristic signals and do not directly establish factual grounding or reasoning validity. All coefficients are manually fixed; learning their relative weights during training remains an open research direction.

D.5 An Exploration of Another Local Judge Reward Function

We also explored a local model-based judge to evaluate the analyses generated by the Qwen3-4B policy during reinforcement learning. The judge is a separate, frozen language model. It receives the claim, available speaker and statement context, article evidence, gold verdict, and the policy's analysis and predicted verdict. It returns five scores in $[0, 1]$: evidence grounding (e), alignment with the gold verdict (v), claim understanding (c), reasoning quality (r), and focus and clarity (f). The judge is instructed to assess whether the analysis uses the evidence to justify the gold verdict, rather than rewarding fluency or verbosity alone. The analysis-quality score is

$$q = 0.30e + 0.40v + 0.10c + 0.15r + 0.05f. \tag{5}$$

The final reward combines this quality score with the predicted verdict's proximity to the gold label:

$$R = q\, w(\hat{y}, y), \qquad w(\hat{y}, y) = \begin{cases} 1.0, & \hat{y} = y, \\ 0.6, & \hat{y} \text{ and } y \text{ are adjacent labels}, \\ 0, & \text{otherwise}. \end{cases} \tag{6}$$

Adjacency follows the ordered scale True, Mostly True, Half True, Mostly False, False, and Pants on Fire. Responses without valid final `<analysis>` and `<answer>` blocks receive zero reward. The judge supplies the training reward; independent test predictions are produced by the trained policy alone.

Table 9: 4B results with Article + Formatted Evidence on the same 2,607 evaluation examples, using greedy decoding and a 4,096-token output limit. Both metrics are percentages.

Model	Accuracy	Balanced accuracy
4B Base	53.59	45.37
4B Unfiltered RL (rule-based reward)	**60.64**	**54.36**
4B Local-judge RL	57.23	50.38

Local-judge RL improves both metrics over the base model, but remains below the rule-based reward variant. This suggests that adding a learned assessment of analysis quality does not necessarily improve verdict prediction in these training configurations. One possible explanation is that uncertainty in the judge's quality estimates, together with partial credit for adjacent incorrect labels, makes the reward less directly aligned with exact-label accuracy. Furthermore, the reward coefficients are manually specified and may not be optimal. Learning these weights during training remains an open research direction.

E Datasets

E.1 Dataset Details

- **Snopes** Snopes: Our balanced four-class subset contains 1,000 examples collected from Snopes. With filtered article evidence, it tests cross-source transfer under shared truthfulness labels; in our experiments, RL gains are larger with evidence.
- **ClaimReview 2025Q4** He et al. (2026): This imbalanced dataset contains 901 claims with three verdict labels. Our claim-only evaluation probes closed-book transfer; differing accuracy and balanced-accuracy trends caution against interpreting aggregate gains as uniform improvements.
- **SciFact** Wadden et al. (2020): Our evaluation uses 339 claim–document pairs from 300 scientific claims to distinguish support, refutation, and insufficient evidence. The results expose sensitivity to whether the supplied evidence covers every material claim detail.
- **ClimateCheck** (Abu Ahmad et al., 2025): The test set contains 1,904 claim–abstract pairs from 176 climate-related claims, extending evidence-sufficiency evaluation to a specialized domain. Our error analysis identifies NEI errors as a substantial contributor to the observed RL regressions when using original prompt.

- **RealFactBench** (Yang et al., 2025): Our filtered text-only binary subset contains 700 examples and probes closed-book verification. Unfiltered RL improves both metrics for both backbones, although this setting does not test evidence retrieval or NEI recognition.
- **SciClaimHunt** (Kumar et al., 2025): Our evaluation uses a 1,000-example sample of scientific claim–evidence pairs. The dataset uses binary labels and synthetic negatives generated through methods including negation and claim–document mismatching; improvements therefore do not directly establish better separation of REFUTED from NEI.
- **QuanTemp++** (Venktesh V et al., 2025): Our subset contains 1,475 numerical claims after excluding PolitiFact examples. Supplied corpus snippets probe numerical and scope sensitivity; for the 4B model, declining accuracy alongside improved balanced accuracy reveals class-level trade-offs.
- **AVeriTeC** (Schlichtkrull et al., 2023): Our 438-example development subset uses supplied human-annotated question–answer evidence and four verdict categories. RL gains show that introducing insufficient-evidence and conflicting-evidence categories does not inevitably cause regression.
- **TSVer** (Strong & Vlachos, 2025): Our 123-example subset uses oracle time-series evidence to test numerical and temporal verification. The 9B gains suggest transfer beyond prose evidence, although the small evaluation subset limits precision.
- **FACTors** (Altuncu et al., 2025): Our five-class subset contains 800 claims published from 2024 onward and probes fine-grained closed-book transfer. Limited gains show that familiar verdict labels alone do not ensure generalization.

Table 10: Dataset configurations for external evaluation.

Dataset	Input	N	K	Prompt
Snopes (C)	C	1,000	4	O
Snopes (C+E)	C+E	1,000	4	O
ClaimReview 2025Q4	C	901	3	O
SciFact	C+E	339	3	M
ClimateCheck	C+E	1,904	3	M
RealFactBench	C	700	2	O
SciClaimHunt	C+E	1,000	2	O
QuanTemp++ (non-PolitiFact)	C+E	1,475	3	O
AVeriTeC dev (oracle QA)	C+E	438	4	O
TSVer (oracle series)	C+E	123	4	O
FACTors 2024+ (5-class)	C	800	5	O

Notes. C: claim only; C+E: claim with supplied evidence. N: number of evaluated examples. K: number of target categories. O: original prompt. M: modified detail-checking prompt.

Checkpoints and decoding settings. Base denotes the released checkpoint before our PolitiFact RL. The 4B and 9B backbones are Qwen3-4B-Thinking-2507 and Qwen3.5-9B, respectively. 4B uses native thinking, $T = 0.6$, top-$p = 0.95$, top-$k = 20$, and a 32,768-token output limit. All 9B and Qwen 27B results use non-thinking mode, as our experiments suggest that it yields better performance, with $T = 0.7$, top-$p = 0.8$, top-$k = 20$, a presence penalty of 1.5, and a 4,096-token output limit. The sampling seed is $42+$ row index.

Reference-model settings and answer extraction. Jev uses `typesafe/jev-1.13` and returns a native typed choice. Both Luna versions use medium reasoning, a 4,096-token output limit, and the same seed schedule. Their API model IDs are `openai/gpt-6-luna` and `openai/gpt-5.6-luna`. The historical 4B Snopes results marked with † use greedy decoding, a 4,096-token output limit, and last-label extraction. Apart from Jev and this historical exception, all results require final-answer tags. Invalid model answers are retained without output repair.

Prompt variants and evaluation scope. O results use the original prompts and the latest complete inference controls, except for the historical 4B Snopes decoding exception described above.

M applies the same development-selected detail-checking and output-format instructions to all ten models; Jev uses its native decision interface. Model weights are unchanged. O and M evaluations of the same dataset use identical examples, evidence, and labels. M results are exploratory follow-ups after test-error analysis, not an independent blind benchmark.

E.2 MODIFIED PROMPT

Below is the modified prompt used for SciFact and ClimateCheck

```
Verification procedure: compare the exact claim with the supplied evidence before deciding.
Check each material assertion, including entity names and roles, quantities and units,
↪  dates, locations, negation, comparison direction, causal statements, and scope words
↪  such as all, any, only, or always.
Do not silently rewrite, weaken, or omit a claim detail to match the evidence.
List any specifically contradicted detail and any detail whose truth is not established.
Use REFUTED if the evidence explicitly contradicts a material assertion in the same scope.
Use SUPPORTED only if the evidence supports all material assertions.
Otherwise use NEI: missing information is not support and is not by itself a contradiction.
Do not substitute general world knowledge, topic similarity, or plausibility for evidence.
A statement being partly mentioned is not sufficient to support its extra details.

Output format: after reasoning, write a brief <analysis>...</analysis> block, then exactly
↪  <answer>LABEL</answer>, replacing LABEL with one allowed label.
The tag name is answer, never the label itself.
Nothing follows the closing answer tag.
```