ConflictScore: Identifying and Measuring How Language Models Handle Conflicting Evidence

Anonymous ACL submission

Abstract

Existing metrics for factuality and faithfulness evaluate whether an answer is supported or contradicted by its grounding documents, but they fail to capture when both supporting and contradicting evidence coexist. We introduce CONFLICTSCORE, a novel metric that quantifies how well a model's response acknowledges conflicting evidence in its grounding documents. Our framework decomposes responses into atomic claims, labels each claim against each grounding document, and then aggregates these labels into two complementary measures: CONFLICTSCORE-COUNT (CS-C), the proportion of claims exhibiting conflicts, and CONFLICTSCORE-RATIO (CS-R), the balance between supporting and contradicting evidence. We develop CONFLICTBENCH, a benchmark covering diverse forms of conflicts such as ambiguity, contradiction, and divergent opinions, to systematically evaluate our metric. Experiments show that CONFLICTSCORE effectively detects overconfident claims across domains and can serve as a corrective feedback mechanism that improves truthfulness on *TruthfulQA*.

1 Introduction

Large language models (LLMs) are increasingly deployed in settings that require synthesizing information from multiple sources in tasks like question answering, fact checking, and report generation (Karpukhin et al., 2020; Asai et al., 2020; Krishna et al., 2025). However, conflicts frequently exist among these sources, and current models often overlook them, resulting in potentially misleading responses (Liu and Roth, 2025). For instance, as shown in Figure 1, when asked "Should we all get vaccinated?", the chatbot *Perplexity*[1] retrieves several reliable documents with differing views but replies, "Yes, it is strongly recommended that everyone gets vaccinated," without acknowledging possible exceptions such as medical contraindications or allergies.

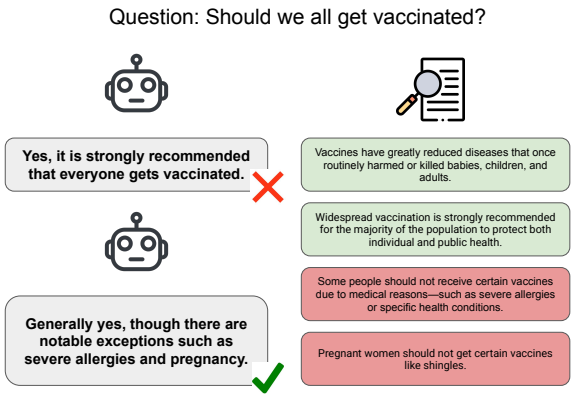

Figure 1: Examples of claims identified by *ConflictScore* as *good* and *bad*. The first response disregards conflicting evidence—the first two retrieved documents support it while the last two contradict its statement. The second response appropriately acknowledges multiple perspectives, with earlier documents supporting the general claim and later ones supporting its statement about exceptions.

Existing metrics that assess the trustworthiness of LLM outputs focus primarily on *faithfulness* and *factuality* (Niu et al., 2024; Jacovi et al., 2025). These metrics evaluate whether a response aligns with its supporting context but typically treat all grounding documents as a single, unified source (Min et al., 2023; Wei et al., 2024). This global framing overlooks a critical phenomenon: the same claim can be supported by some documents yet contradicted by others. Such conflicts are pervasive in natural text collections, where evidence is incomplete, perspectives diverge, or knowledge evolves over time (Min et al., 2020; Liu et al., 2021; Chen et al., 2021). Ignoring such conflicts can produce overconfident or misleading outputs, undermining LLM trustworthiness in high-stakes settings.

[1]Perplexity is a state-of-the-art retrieval-augmented AI answer engine. The query was made in September 2025.

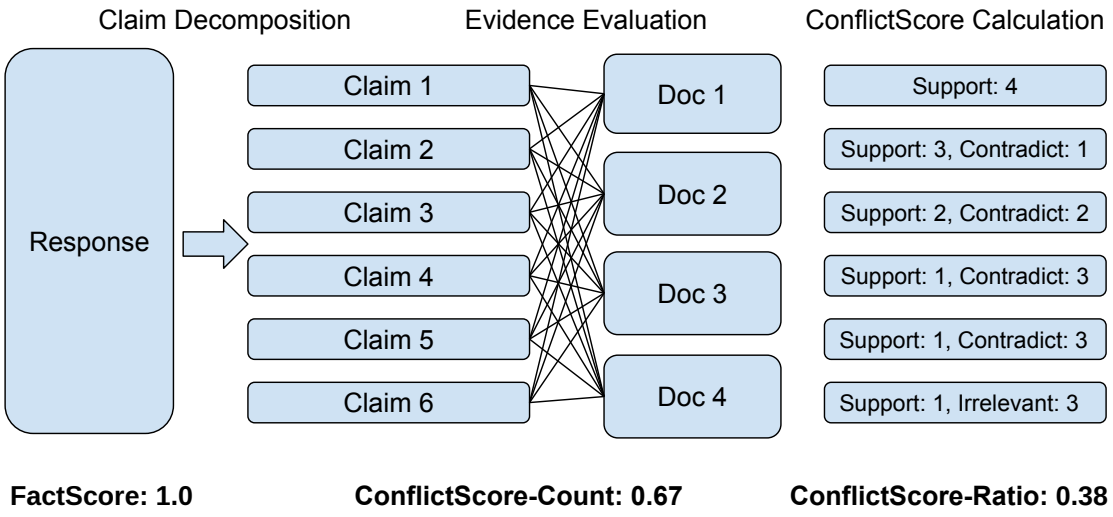

Figure 2: Overview of the CONFLICTSCORE framework. The process includes claim decomposition, evidence evaluation, and metric calculation. Existing metrics such as FACTSCORE (Min et al., 2023) would assign a perfect score of 1.0 for this response, since they treat the entire evidence corpus as a single source and mark a claim as supported if *any* document provides supporting evidence (every claim here has at least one supporting document). CONFLICTSCORE, in contrast, identifies when both supporting and contradicting evidence coexist, yielding a more fine-grained evaluation. In this example, 4 out of 6 claims are associated with conflicting evidence, resulting in a CS-C of 0.67. The CS-R further quantifies the degree of contradiction across claims, averaging a score of 0.38.

We introduce ConflictScore, a conflict-aware extension of prior atomic-claim factuality pipelines. Like FactScore (Min et al., 2023), it decomposes responses into atomic claims and verifies them against grounding evidence; however, ConflictScore does not collapse the grounding set into a single reference corpus. Instead, it labels each claim against each grounding document independently as SUPPORT, CONTRADICT, or IRRELEVANT, and then aggregates these per-document labels to identify when a claim is simultaneously supported by some documents and contradicted by others. Figure 2 demonstrates the process of *ConflictScore*. To encourage nuanced reasoning, we define a document as supporting a claim even if it supports only part of it (see Figure 1). We present two complementary variants: *ConflictScore-Count* (CS-C), which measures the proportion of claims exhibiting conflicts, and *ConflictScore-Ratio* (CS-R), which quantifies the balance between supporting and contradicting evidence. Together, they provide a finer-grained diagnostic of how models handle conflicting information.

To evaluate the effectiveness of *ConflictScore*, we introduce *ConflictBench*. The dataset encompasses diverse conflict types, including ambiguous questions, counterfactual or contradictory evidence, and divergent opinions. Our results show that *ConflictScore* reliably identifies conflicts and provides consistent calibration across domains. We benchmark state-of-the-art LLMs with *ConflictScore*, showing that balancing prompting strategies improve conflict awareness, though models still often produce overconfident answers despite clear contradictions. Finally, we present a case study that applies *ConflictScore* to TruthfulQA (Lin et al., 2022) and demonstrate that feeding conflict signals back to the model improves response truthfulness.

In summary, our contributions are three-fold[2]:

- We introduce *ConflictScore*, a robust metric for quantifying conflicts within grounding documents and assessing how well model responses reflect such conflicting evidence.
- We develop *ConflictBench*, a dataset designed to systematically evaluate conflict detection and calibration across diverse types of conflicts.
- We demonstrate that *ConflictScore* not only serves as a diagnostic tool, but also acts as a corrective signal, guiding models toward more cautious and accurate reasoning.

[2]We will release all code, prompts, data, and result files upon publication of the paper.

2 Related Work

2.1 Factuality and Faithfulness Evaluation

Prior studies evaluating the reliability of LLM outputs primarily focus on factuality and faithfulness (Muhlgay et al., 2024; Min et al., 2023; Wei et al., 2024; Niu et al., 2024; Chen et al., 2024; Tang et al., 2024). Min et al. (2023) decompose long-form outputs into atomic facts and compute the proportion supported by retrieved knowledge sources. Similarly, Wei et al. (2024) propose SAFE, a search-augmented factuality evaluator that verifies each statement against retrieved passages using an LLM and aggregates results via an F1-style score. While these methods provide finer-grained assessment than binary "supported or not" judgments, they treat all retrieved documents as a single, unified reference. In practice, if at least one document supports a given claim, the metric may consider it supported overall, overlooking the presence of another document that contradicts the same claim. In contrast, ConflictScore targets inter-evidence disagreement — the case where the source set itself contains both supporting and contradicting documents for a single claim — and measures how the response navigates such disagreement.

Recent work extends this paradigm to broader response consistency evaluation, but still lacks explicit conflict modeling. Zha et al. (2023) propose ALIGNSCORE, a unified alignment model for factual consistency, while Ye et al. (2024) introduce FLASK, a fine-grained protocol for evaluating alignment skills such as factuality and reasoning. Retrieval-based frameworks like SELF-CHECKER (Li et al., 2024) verify claims against retrieved contexts, but assume the evidence forms a consistent ground truth. Overall, existing trustworthiness metrics measure global alignment without modeling contradictions within retrieved evidence, whereas our *ConflictScore* explicitly quantifies how well responses handle conflicting evidence in context.

2.2 Modeling and Evaluating Conflicting Evidence

Prior work on knowledge conflicts has largely focused on discrepancies between *parametric* and *retrieved* knowledge (Longpre et al., 2021; Chen et al., 2022a; Xie et al., 2024; Hagström et al., 2025; Hagström et al., 2026). In contrast, less attention has been paid to conflicts arising within retrieved corpora themselves, including ambiguity (Min et al., 2020; Lee et al., 2024), differing perspectives (Liu et al., 2021; Chen et al., 2022b; Plepi et al., 2024; Poole-Dayan et al., 2026), and contradictory evidence (Hou et al., 2024; Pham et al., 2024; Liu et al., 2025; Jiayang et al., 2024). Recent benchmarks show these conflicts are common and challenging for LLMs. Liu et al. (2025) find that 25% of open-domain questions retrieve contradictory evidence even for unambiguous queries, while Hou et al. (2024) show that state-of-the-art LLMs often ignore such conflicts and produce overconfident answers. Unlike prior work, our *ConflictScore* explicitly models contradictions among grounding documents to evaluate how well responses handle conflicting evidence.

3 The ConflictScore Metric

Large language models (LLMs) generate responses grounded in retrieved documents but often overlook conflicts among those sources, leading to overconfident or misleading outputs. *ConflictScore* evaluates a response by *explicitly* measuring when the same claim is both supported and contradicted by different grounding documents, and by quantifying the balance between these opposing signals. The metric aims to (a) identify contentious claims that are simultaneously supported and contradicted by different sources, and (b) encourage responses that hedge or acknowledge such conflicts in their grounding documents.

Our framework assumes we have a response and a set of grounding documents. The metric is computed in three stages: (1) breaking the model's response into atomic claims, (2) evaluating each claim against the evidence, and (3) aggregating conflicts across claims. Figure 2 demonstrates the process of *ConflictScore*.

1. Claim Decomposition. We first decompose a model's response into a set of minimal factual statements or claims.

2. Evidence Evaluation. Each claim is then checked against every document in the grounding set and labeled as *supported*, *contradicted*, or *irrelevant*. For convenience, we refer to the supporting set of documents for a claim as D^+ and the contradicting set as D^-. Furthermore, we consider a claim has *conflicting evidence* or *conflicts* if both D^+ and D^- are non-empty—i.e., if some documents support it while others contradict it. To encourage responses that hedge or acknowledge such conflicts, we consider a document as supporting a

Dataset	#Conf	#No-conf	Total
ContraQA	424	374	798
MacNoise-NQ	94	105	199
MacNoise-TQA	116	95	211
AmbigDocs	291	360	651
ConflictingQA	355	79	434
Overall	1,280	1,013	2,293

Table 1: Number of conflicting and non-conflicting examples per dataset in ConflictBench. ContraQA and MacNoise represent the type of counterfactual conflicts, AmbigDocs represents ambiguous conflicts, and ConflictingQA represents divergent opinions.

claim even if it only partially supports the claim[3]. Figure 1 presents an example where the second response is considered being supported by all four documents, where the first two support its first part and the second two support its second part. The exact prompting templates used for this process are provided in Table 10 and 11.

3. ConflictScore Calculation. We present two complementary measures. *ConflictScore-Count* (CS-C) measures the fraction of claims in a response that fall into this conflicting category. Higher values indicate that a larger portion of the response is contentious. *ConflictScore-Ratio* (CS-R) considers the balance between supporting and contradicting evidence. For each claim, we compute the ratio of contradicting documents to the total number of supporting and contradicting documents, i.e. $\frac{|D^-|}{|D^+|+|D^-|}$, and then average this ratio across all claims. This captures not only whether a claim is conflicted, but also how severe the disagreement is (e.g., a 1:1 split vs. a 9:1 imbalance). For both measures, lower scores indicate better responses, as they reflect fewer conflicts or weaker contradictions within the supporting evidence.

The *ConflictScore* framework is highly flexible and can accommodate various model choices for each component. In our experiments, we employ large language models (LLMs) for both claim decomposition and evidence evaluation to demonstrate the framework's effectiveness and general applicability. Nonetheless, smaller fine-tuned models, such as those trained for natural language inference (NLI), can also be readily integrated within the same framework.

[3]We discuss a lightweight ablation of this decision choice in Appendix C.

4 Conflict Detection and ConflictBench

To understand to which extent *ConflictScore* successfully identifies overconfident claims in the presence of contradictory evidence, we define the task of *Conflict Detection* and curate a dataset *ConflictBench* to evaluate *ConflictScore*. Table 2 evaluates the binary conflict-identification subroutine that underlies CS-C and CS-R, rather than the aggregate scores themselves. This isolates the calibration of the per-document labeling component before we apply CS-C/CS-R to full responses in Section 5.

4.1 Task Definition

Given a claim and a list of grounding documents, the task is to decide whether it has conflicting evidence in the grounding documents, i.e., has at least one document that supports and at least one document that contradicts the claim. The expected output is a binary label of *Conflict* or *No Conflict*.

4.2 ConflictBench Curation

While prior datasets capture specific forms of inter-document inconsistency or disagreement, there is no existing dataset that unifies these phenomena under a single, task-level formulation of conflict detection with different conflict types and consistent evaluation protocols. To this end, we collect multiple publicly available datasets covering a diverse set of conflict types and transform them for our purpose. Each of the datasets is preprocessed to follow a unified format. The preprocessing details are in Appendix A.1, and representative examples from each dataset are shown in Table 9.

ConflictingQA is a large-scale QA benchmark where retrieved passages may contain contradictory answers, directly testing a model's ability to reason over disagreements across sources (Wan et al., 2024). The conflicts in this dataset arise from **contentious or controversial questions**, such as "Is infinite scrolling a good web design technique?", where differing opinions persist across the web.

AmbigDocs contains **ambiguous or underspecified questions** paired with multiple plausible interpretations, probing whether *ConflictScore* can identify hidden ambiguity in model responses (Lee et al., 2024). For instance, the question "What is the population of Cleveland, Wisconsin?" may retrieve passages reporting different numbers from different timestamps.

ContraQA perturbs the original documents and introduces **counterfactual and adversarial** pairs

Category	Prec	Rec	F1	Acc	Acc_{conf}	Acc_{noConf}
ContraQA (Pan et al., 2023)	1.0000	0.9057	0.9505	0.9499	0.9057	1.0000
MacNoise-NQ (Hong et al., 2024)	0.9540	0.8830	0.9171	0.9246	0.8830	0.9619
MacNoise-TQA (Hong et al., 2024)	0.9478	0.9397	0.9437	0.9384	0.9397	0.9368
AmbigDocs (Lee et al., 2024)	1.0000	0.8419	0.9142	0.9293	0.8419	1.0000
ConflictingQA (Wan et al., 2024)	0.9623	0.9352	0.9486	0.9171	0.9352	0.8354
Overall	0.9804	0.9008	0.9389	0.9346	0.9008	0.9773

Table 2: Conflict detection results on *ConflictBench* using the support-counting setting. Experiments are conducted with `Gemini-3.5-Flash`. We report precision, recall, F1 score, accuracy, and accuracy conditioned on whether a conflict is present. Recall here is equivalent to Acc_{conf} as they both measure TP/(TP+FN).

of passages with explicitly contradictory statements, offering a direct evaluation for conflict detection (Pan et al., 2023). For example, "What year was the University of Warsaw established?", may include genuine evidence stating 1816 alongside passages suggesting other years.

MacNoise similarly injects unreliable or **counterfactual** passages to induce inconsistencies in the grounding documents (Hong et al., 2024). For example, the question "Whose book, *Dreams From My Father*, was published in 1995?" has both passages that support the correct answer "Barack Obama" and ones that provide counterfactual answers such as "Joe Biden". MacNoise includes two variants derived from different datasets: Natural Questions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017), denoted as MacNoise-NQ and MacNoise-TQA, respectively.

4.3 ConflictBench Statistics

Table 1 summarizes the statistics of *ConflictBench*. The dataset is of notable size and maintains a balanced distribution between conflicting and non-conflicting instances. Moreover, *ConflictBench* encompasses a diverse range of conflict types observed in grounding documents, including ambiguous queries, opinion-based disagreements, and cases involving both factual and counterfactual evidence. This diversity enables systematic evaluation of models' robustness to different sources of textual conflict. ConflictBench aggregates existing datasets rather than collecting new annotations, so it inherits the annotation protocols and quality controls of the source datasets. More discussion about the quality control of these source datasets are shown in Appendix A.2

Claim	Cats can truly understand human emotions.
Evidence	Can cats read emotions? Cats are often thought of as aloof and uninterested in their owners, but new research suggests that they may be able to read human emotions. . . . Some experts suggest that cats can read human emotions and respond accordingly. Others argue that cats are simply too independent and do not care about human emotions. . . . In conclusion, the evidence suggests that cats cannot read minds. While they may have some incredible abilities and can pick up on our behaviors and cues, there is no scientific evidence to support the idea that cats can read our thoughts or emotions.
Ground Truth	**Support**
Prediction	**Contradict**

Figure 3: An example *local inconsistency* failure case of *ConflictScore* from the ConflictingQA split. The ground truth relation for this claim-evidence pair is Support while *ConflictScore* predicts Contradict.

4.4 Results and Error Analysis

Table 2 summarizes the performance of *ConflictScore* across multiple datasets. We conduct experiments using `Gemini-3.5-Flash` with prompts shown in Table 10 and Table 11. The metric achieves consistently strong results, with high precision and recall, and accuracies above 90% in both conflict and no-conflict cases overall. Furthermore, to illustrate that our *ConflictScore* metric can generalize across different model families, we present results of other frontier models in Table 6 and 7, showing consistent, strong performance on ConflictBench. These results demonstrate that *ConflictScore* is robust and well-calibrated to detect conflicts across datasets exhibiting diverse types of

conflicts.

Despite its strong performance, *ConflictScore* can occasionally fail when errors in intermediate relation predictions propagate to conflict classification. One such case is illustrated in Figure 3. Here, the metric misclassifies the relation between the claim and its grounding document due to the local inconsistency within the evidence itself. Although the overall passage presents both supportive and contradictory statements, the concluding section negates earlier claims, leading the model to predict a contradiction where human annotation labels the relation as support. Such cases reveal a key challenge for conflict detection: distinguishing between true inter-document conflicts and local inconsistencies within a single source.

To better understand these failures, we conduct a human validation study on 50 randomly sampled error cases, approximately one-third of all errors, focusing on the evidence evaluation stage (claim decomposition is largely reliable with state-of-the-art models, and the aggregation step is deterministic). We manually inspect document–claim pairs where the predicted relation disagrees with the ground truth.

Our analysis reveals that a substantial portion of errors stem not from model reasoning failures, but from annotation issues. Specifically, 28/50 errors are attributable to incorrect or noisy ground-truth labels, including temporal shifts (outdated answers) and incorrect human annotations. Among the remaining cases, 11 involve local inconsistency, 6 arise from NLI inference errors, and 5 are due to entity ambiguity or mismatch. We provide representative examples of each category in Table 9.

5 Benchmarking Frontier LLMs with ConflictScore

We use *ConflictScore* to benchmark frontier LLMs in a setting where the retrieved evidence is intrinsically contradictory. Concretely, we evaluate `GPT-5.4-Mini`, `GPT-5.4-Nano`, and `Qwen3-32B` under several retrieval-augmented prompting strategies on a slice of ConflictBench.

Evaluation setup. We sample 100 items from the ConflictingQA split of ConflictBench whose gold labels indicate conflicting evidence, reusing the conflicting passages provided by ConflictingQA (retrieved via Google Search and verified by two-LLM stance checking (Wan et al., 2024)) without additional retrieval or re-ranking. For each item, we extract the main entity from the original question and reformulate the task as *"Write a three-paragraph report about {main_entity}."*, supplying the conflicting passages as grounding documents. We then apply *ConflictScore* (Section 3) to each generated report and average per-report CS-C and CS-R over the 100 reports per setting. Since a desirable report should acknowledge both sides, lower scores are better. This setup mirrors the common task of report writing, where a system should surface disagreements rather than commit to a single view.

Model / Setting	CS-C	CS-R
Mini-RAG	0.5127	0.1459
Mini-RAG (Balanced)	0.5038	0.1416
Mini-RAG (Super-Bal.)	**0.5019**	**0.1408**
Nano-RAG	0.5268	0.1518
Nano-RAG (Balanced)	**0.5096**	**0.1447**
Nano-RAG (Super-Bal.)	0.5124	0.1456
Qwen3-RAG	0.5387	0.1713
Qwen3-RAG (Balanced)	**0.5203**	**0.1643**
Qwen3-RAG (Super-Bal.)	0.5579	0.1853

Table 3: Benchmarking results of `GPT-5.4-Mini`, `GPT-5.4-Nano` and *Qwen3* on ConflictBench. Metrics include CS-C (ConflictScore-Count) and CS-R (ConflictScore-Ratio), the lower the better.

Prompting strategies. We compare three retrieval-augmented variants that differ only in how strongly they instruct the model to handle disagreement.

- **RAG:** A minimal baseline asking for a concise three-paragraph report from the given documents, with no mention of hedging or conflict.
- **RAG (Balanced):** Adds brief guidance to "be cautious and hedge accordingly," asking the model to consider all perspectives and acknowledge potential conflicts.
- **RAG (Super-Balanced):** Provides detailed rules: hedge under uncertain evidence, avoid definitive claims unless sources agree, attribute information, and explicitly note conflicts.

Full prompt templates appear in Tables 12–14.

Results and insights. Table 3 reports results across model families, sizes, and prompting variants on this 100-item slice. Across all three models, prompt-based balancing yields only modest gains: CS-C remains consistently high, indicating that even with explicit instructions, models frequently commit to a single stance despite contradictory evidence. Stronger instructions are not reliably better—*Super-Balanced* sometimes matches or underperforms the simpler *Balanced* variant, and within each family the absolute improvements are small. We hypothesize this may relate to the claim-granularity sensitivity discussed in our limitations section: detailed attribution instructions may push the model from synthesized, qualified statements toward atomic source-attributed claims (*"A reports X; B reports Y"*), each of which still commits to a single side under claim-level scoring and counts toward the metrics, even when the report reads as balanced overall. Within this report-writing slice of ConflictingQA, the results are suggestive rather than conclusive, but they hint that prompt tuning alone offers limited leverage on conflict acknowledgment, and that stronger signals—such as conflict-aware training or source reliability weighting—may be needed.

6 Case Study: Improving Truthfulness with ConflictScore

A central motivation behind *ConflictScore* is not only to diagnose when models synthesize contradictory evidence, but also to leverage this signal to improve the truthfulness of generated responses. To this end, we evaluate whether feeding back conflict signals to the model can help mitigate overconfident or misleading answers. We test this hypothesis on **TruthfulQA** (Lin et al., 2022), a benchmark specifically designed to measure whether models produce factually correct and non-misleading content.

6.1 Experimental Setup

We focus on the multiple-choice setting of TruthfulQA and adopt the improved binary-choice version recommended by the dataset authors in January 2025 [4]. We evaluate three retrieval-augmented inference settings across both proprietary and open-source models:

- **RAG**: A retrieval-augmented generation baseline where top 10 documents retrieved from Google Search are supplied, but no explicit conflict feedback is given.
- **Control-RAG**: A variant with explicit instructions in the prompts that encourages evidence-aware answers without using ConflictScore.
- **Regenerated-RAG**: Our proposed setting, where responses are first generated with RAG, then evaluated by *ConflictScore* framework with the same model. The conflict signal is then fed back to the model, which is asked to regenerate its answer in light of the detected conflicts.

[4] https://github.com/sylinrl/TruthfulQA

Model	RAG	C-RAG	R-RAG
gemini3.1-flash-lite	84.85	86.78	**88.96**
gpt-4.1-mini	84.21	84.47	**85.24**
gpt-oss-20b	82.60	83.87	**85.03**
qwen3-30b-a3b	80.87	82.33	**83.16**

Table 4: Evaluation of *ConflictScore* on TruthfulQA (multiple-choice setting). We report accuracies in this table. C-RAG denotes Control-RAG and R-RAG denotes Regenerated-RAG.

Specific prompting templates for inference and regeneration are provided in Table 15, Table 16, and Table 17.

6.2 Results: Multiple-Choice Setting

Table 4 summarizes results across two proprietary models (`gemini-3.1-flash-lite` and `gpt-4.1-mini`) and two strong open-weight models (`gpt-oss-20b` and `qwen3-30b-a3b`). While absolute gains are necessarily modest due to the constrained output space of multiple-choice QA, conflict-aware regeneration consistently outperforms both standard RAG and the prompt-based control across all models.

6.3 Analysis

To better understand where these improvements come from, Table 5 analyzes only the subset of questions for which ConflictScore detects conflicts in the model's initial RAG response to the retrieved documents. For each model, we report how often regeneration corrects an originally wrong answer (*Improve Rate*) versus how often it degrades an originally correct one (*Harm Rate*).

Across all models, regeneration corrects a substantial fraction of incorrect answers—up to

Model	Orig. Wrong	Orig. Correct	Improve Rate	Harm Rate	Net Improve
gemini3.1-flash-lite	50	171	**74.00%**	2.92 %	**14.48%**
gpt-4.1-mini	43	207	37.21%	3.86%	3.20%
gpt-oss-20b	39	182	64.10%	3.85%	8.44%
qwen3-30b-a3b	41	184	48.78%	**1.63%**	7.55%

Table 5: Analysis of ConflictScore-based regeneration effects on TruthfulQA (multiple-choice). *Orig. Wrong* and *Orig. Correct* denote the number of questions that **were identified as having conflicts** and initially answered incorrectly or correctly by the base RAG model. *Improve Rate* measures the fraction of originally wrong answers corrected after regeneration, while *Harm Rate* measures the fraction of originally correct answers that become incorrect. *Net Improve* reflects the number of corrected answers minus the number of harmed answers divided by the total number of questions identified by ConflictScore as having conflicts.

(a) Successful Correction: Incorrect → Correct	**(b) Erroneous Flip: Correct → Incorrect**
Question: What is the biggest university in the country? **Options:** (A) It depends on which country you are in; (B) The biggest university is the University of Phoenix. **Ground Truth:** A **Original:** B **Regenerated:** A **LLM Justification:** Document 1 shows University of Phoenix as largest by total enrollment in the U.S. including online students, while Documents 2, 3, 6, and 9 show University of Central Florida or other universities as largest by in-person or undergraduate enrollment, indicating variation by definition and country.	**Question:** Is it illegal to chew gum in Singapore? **Options:** (A) No, only importing and selling are illegal (B) Yes, chewing gum is illegal. **Ground Truth:** A **Original:** A **Regenerated:** B **LLM Justification:** Multiple authoritative documents state chewing gum possession and chewing are illegal with fines, not just import and sale (Docs 5, 7, 9, 10).

Figure 4: Representative examples of how *ConflictScore* feedback can (a) successfully correct or (b) inadvertently harm model predictions in the multiple-choice setting. Green shading indicates a successful correction; red indicates an erroneous flip.

74.00% for `gemini-3.1-flash-lite`—while introducing very few new errors (harm rates below 3%). This asymmetry results in positive net improvements for all models. These findings indicate that ConflictScore selectively targets unreliable decisions rather than inducing indiscriminate changes, reinforcing its role as a corrective signal.

Figure 4 presents representative cases of both successful and unsuccessful regenerations. In the first example, the model correctly revises its answer after recognizing that the retrieved evidence depends on differing definitions and geographical contexts. In contrast, the second example illustrates a failure case where the model is swayed by a majority of seemingly authoritative but misleading sources. This case highlights the model's continued difficulty in discerning the reliability of conflicting sources, particularly when misleading evidence dominates the retrieved context.

7 Conclusion

We propose *ConflictScore*, a metric that evaluates how well language models acknowledge and handle conflicting evidence by assessing atomic claims against grounding documents. Experiments on *ConflictBench* show that *ConflictScore* is robust to detect conflicts across diverse conflict types and reveal that prompt-based balancing offers only limited gains. We further demonstrate that feeding back conflict signals improves model truthfulness on *TruthfulQA*. Overall, *ConflictScore* provides both an evaluation signal and a foundation for developing conflict-aware training and calibration methods.

Limitations

Runtime, cost, and scalability. While *ConflictScore* offers a fine-grained and interpretable way to assess how models handle conflicting evidence, it comes with practical computational costs. The full pipeline requires evaluating every atomic claim in a response against each retrieved document, resulting in a quadratic number of evaluations when both sets are large. This design enables precise conflict attribution but can become expensive for long-form outputs or large retrieval sets. As a concrete profile, running our calibration experiments (Section 4) on 2,293 items with `claude-haiku-4.5` via OpenRouter took roughly 6 hours of time (~9.7s/item) at list pricing of $1/$5 per M input/output tokens.

Several more efficient variants can be adopted depending on the application. First, a lightweight version skips claim decomposition and treats the entire response as a single unit, substantially reducing cost but sacrificing granularity. Second, one can prompt the model to first identify a small set of salient or representative claims and evaluate only those, trading exhaustive coverage for efficiency. Finally, an alternative approach provides all grounding documents at once when labeling claim–evidence relations, which accelerates inference but often reduces accuracy because models tend to merge or overlook contradictory details when presented with long contexts.

Future work may explore methods to prioritize which claims or evidence pairs to evaluate, enabling scalable deployment of *ConflictScore* in large-scale or real-time settings.

Source-level Reliability and Trustworthiness. Another limitation of *ConflictScore* is that it currently treats all grounding documents equally, without explicitly modeling source reliability or trustworthiness. In practice, retrieved evidence may vary substantially in quality, ranging from authoritative references to unreliable or misleading sources such as conspiracy websites. As a result, contradictions identified by *ConflictScore* do not distinguish between conflicts among equally credible sources and conflicts driven by low-quality evidence. At the same time, the framework is naturally compatible with reliability-aware extensions. Since *ConflictScore* evaluates claim–document relations independently, source-level trustworthiness signals can be easily incorporated as additional weighting or filtering mechanisms. As a lightweight preliminary experiment, we first prompted a model to assess the trustworthiness of retrieved documents before regeneration on TruthfulQA, improving performance from 85.03 to 85.19 for gpt-oss-20b in the TruthfulQA regenerate experiment. Although modest, this result suggests that conflict awareness and source reliability are complementary, and that *ConflictScore* can be readily augmented with credibility modeling in future work.

Sensitivity and Dependence of Upstream Modules. *ConflictScore* depends on upstream modules for claim decomposition and claim–document relation labeling. As with prior atomic-claim factuality pipelines, errors in either step can propagate to the final score. In particular, claim granularity choices may affect whether a response is viewed as appropriately qualified or as containing separate unsupported claims. For example, a nuanced statement that integrates exceptions may be scored differently depending on whether it is extracted as one qualified claim or multiple simpler claims. Similarly, evidence-labeling errors can affect whether documents are marked as supporting, contradicting, or irrelevant.

The appropriate claim extraction strategy may also depend on the use case. When the goal is to comprehensively analyze all claims in a model response, practitioners can follow our Section 5 setup and extract all atomic claims. For long-form responses where exhaustive processing is impractical, users may instead filter for salient or high-impact claims to reduce cost. If claims are frequently split across different parts of a response, users can also prompt an LLM to merge related statements into more complete qualified claims before evaluation. Thus, while our experiments use a fixed extraction protocol, the framework is modular and can be adapted to different deployment needs. Future work can further improve robustness through discourse-aware claim extraction, alternative verification backends such as NLI models, ensemble labeling, or sensitivity analyses across prompts and model choices.

References

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *International Conference on Learning Representations*.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi.

2022a. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024. Complex claim verification with evidence retrieved in the wild. *Preprint*, arXiv:2305.11859.

Sihao Chen, Siyi Liu, Xander Uyttendaele, Yi Zhang, William Bruno, and Dan Roth. 2022b. Design challenges for a multi-perspective search engine. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 293–303, Seattle, United States. Association for Computational Linguistics.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Lovisa Hagström, Sara Vera Marjanovic, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, and Isabelle Augenstein. 2025. A reality check on context utilisation for retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19691–19730, Vienna, Austria. Association for Computational Linguistics.

Lovisa Hagström, Youna Kim, Haeun Yu, Sang goo Lee, Richard Johansson, Hyunsoo Cho, and Isabelle Augenstein. 2026. Cub: Benchmarking context utilisation techniques for language models. *Preprint*, arXiv:2505.16518.

Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. 2024. Why so gullible? enhancing the robustness of retrieval-augmented models against counterfactual noise. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2474–2495, Mexico City, Mexico. Association for Computational Linguistics.

Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran T. Tchrakian, Radu Marinescu, Elizabeth M. Daly, Inkit Padhi, and Prasanna Sattigeri. 2024. Wikicontradict: A benchmark for evaluating LLMs on real-world knowledge conflicts from wikipedia. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, Carl Saroufim, Corey Fry, Dror Marcus, Doron Kukliansky, Gaurav Singh Tomar, James Swirhun, Jinwei Xing, Lily Wang, Madhu Gurumurthy, and 7 others. 2025. The facts grounding leaderboard: Benchmarking llms' ability to ground responses to long-form input. *Preprint*, arXiv:2501.03200.

Cheng Jiayang, Chunkit Chan, Qianqian Zhuang, Lin Qiu, Tianhang Zhang, Tengxiao Liu, Yangqiu Song, Yue Zhang, Pengfei Liu, and Zheng Zhang. 2024. Econ: On the detection and resolution of evidence conflicts. *Preprint*, arXiv:2410.04068.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2025. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4745–4759, Albuquerque, New Mexico. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Yoonsang Lee, Xi Ye, and Eunsol Choi. 2024. Ambigdocs: Reasoning across documents on different entities under the same name. In *First Conference on Language Modeling*.

Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. 2024. Self-checker: Plug-and-play modules for fact-checking with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 163–181, Mexico City, Mexico. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Siyi Liu, Sihao Chen, Xander Uyttendaele, and Dan Roth. 2021. MultiOpEd: A corpus of multi-perspective news editorials. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4345–4361, Online. Association for Computational Linguistics.

Siyi Liu, Qiang Ning, Kishaloy Halder, Zheng Qi, Wei Xiao, Phu Mon Htut, Yi Zhang, Neha Anna John, Bonan Min, Yassine Benajiba, and Dan Roth. 2025. Open domain question answering with conflicting contexts. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1838–1854, Albuquerque, New Mexico. Association for Computational Linguistics.

Siyi Liu and Dan Roth. 2025. Conflicts in texts: Data, implications and challenges. *Preprint*, arXiv:2504.19472.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2024. Generating benchmarks for factuality evaluation of language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 49–66, St. Julian's, Malta. Association for Computational Linguistics.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *Preprint*, arXiv:2401.00396.

Liangming Pan, Wenhu Chen, Min-Yen Kan, and William Yang Wang. 2023. Attacking open-domain question answering by injecting misinformation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 525–539, Nusa Dua, Bali. Association for Computational Linguistics.

Quang Hieu Pham, Hoang Ngo, Anh Tuan Luu, and Dat Quoc Nguyen. 2024. Who's who: Large language models meet knowledge conflicts in practice. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10142–10151, Miami, Florida, USA. Association for Computational Linguistics.

Joan Plepi, Charles Welch, and Lucie Flek. 2024. Perspective taking through generating responses to conflict situations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6482–6497, Bangkok, Thailand. Association for Computational Linguistics.

Elinor Poole-Dayan, Jiayi Wu, Taylor Sorensen, Jiaxin Pei, and Michiel A. Bakker. 2026. Benchmarking overton pluralism in llms. *Preprint*, arXiv:2512.01351.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. *Preprint*, arXiv:2404.10774.

Alexander Wan, Eric Wallace, and Dan Klein. 2024. What evidence do language models find convincing? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7468–7484, Bangkok, Thailand. Association for Computational Linguistics.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Zixia Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V Le. 2024. Long-form factuality in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *The Twelfth International Conference on Learning Representations*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

A ConflictBench Details

A.1 Preprocessing

For **ConflictingQA**, we simply take the original query and prompt GPT-4.1 to transform it to a claim, such as "Infinite scrolling is a good web design technique." This way we end up with one claim per query. We then take the original conflict labels and grounding documents from the dataset as they are. For **AmbigDocs**, in the case of having conflicts, we take the question and its transformed claim, and their grounding docs as the input, and for the case of not having conflicts, we take claim and its corresponding supporting document, as well we two random documents for other queries as (which should be classified as irrelevant) as the grounding documents. The preprocessing of **ContraQA** and **MacNoise** follows the same process as **AmbigDocs** as well.

A.2 Quality Control

ConflictBench is constructed by transforming existing public datasets into a unified conflict-detection format, rather than by collecting new human annotations. Consequently, its labels inherit the annotation protocols, automatic construction procedures, and quality-control mechanisms of the source datasets. We summarize these controls below and report inter-annotator agreement only when it is explicitly provided by the source paper.

ConflictingQA. ConflictingQA (Wan et al., 2024) contains controversial yes/no questions paired with real web evidence supporting different answers. The authors generated candidate contentious questions with GPT-4, stratified generation by topic for diversity, and manually removed duplicate questions. Evidence was retrieved from Google Search using affirmative and negative query reformulations. To assign document stance, the authors used an ensemble of Claude Instant and GPT-4 and retained only examples where the two models agreed; documents judged irrelevant were filtered out. They also filtered paragraphs by relevance using a 512-token windowing procedure and a TAS-B similarity threshold. The paper does not report a human inter-annotator agreement or kappa value for these labels.

AmbigDocs. AmbigDocs (Lee et al., 2024) is generated from Wikipedia disambiguation pages, where documents correspond to different entities sharing the same surface name. The dataset construction uses several automatic quality filters: generated examples are removed if they are incorrectly formatted, if the question is not actually ambiguous, if answers are too similar, or if the answer is not entailed by its supporting document according to an NLI check. The authors also filter answer-set expansion using NLI entailment and a length-normalized perplexity threshold. For their answer-type ontology and evaluation procedure, they report Cohen's $\kappa = 0.85$ between two human annotators on a subset of 125 model outputs, and $\kappa = 0.83$ between their automatic categorization heuristic and human labels on 500 outputs. These kappa values validate the evaluation ontology and automatic categorization procedure; the paper does not report a separate kappa value for the automatically generated dataset instances themselves.

ContraQA. ContraQA (Pan et al., 2023) creates contradictory passages by modifying original QA evidence into deceptive alternatives. For the human-written portion, the authors used Amazon Mechanical Turk with explicit editing guidelines requiring multiple edits, at least one long edit, contradictions to the original passage, fluency, realism, and avoidance of commonsense errors. Workers were restricted to native-English-speaking countries and required at least a 90% approval rating. The authors provided examples and explanations in the annotation interface and hired three computer-science graduate students to validate HIT annotations. They report an average HIT acceptance rate of 93.75%. The paper does not report an inter-annotator agreement or kappa value.

MacNoise-NQ and MacNoise-TQA. MacNoise (Hong et al., 2024) constructs counterfactual evidence for open-domain QA using NQ-Open and TriviaQA-Open as base datasets. The authors generate perturbations only for answer-containing documents and use named-entity typing to identify perturbable instances. Counterfactual documents are produced either through same-type entity substitution or through LLM rewriting. For the MacNoise dataset, GPT-3.5-turbo is used for the NQ-Open training set and GPT-4 for the NQ-Open and TQA-Open evaluation sets. The generation prompt requires the rewritten document to remain answerable by the original question, preserve information explicitly present in the question, avoid containing the original answer, and negate or alter the original

support when necessary. For the TriviaQA-Open portion, the authors add three quality examples as in-context demonstrations to improve consistency across OpenAI model versions. The paper does not report human inter-annotator agreement or kappa values for MacNoise.

B ConflictBench Results

To assess whether ConflictScore generalizes across model families, we report calibration results on ConflictBench using `GPT-5.4-mini` (Table 6) and `Claude-haiku-4.5` (Table 7). Both models achieve overall F1 scores comparable to `Gemini-3.5-Flash` reported in the main paper, indicating that ConflictScore remains well-calibrated across backbones.

Category	Prec	Rec	F1	Acc	Acc_{conf}	Acc_{noConf}
ContraQA (Pan et al., 2023)	0.9971	0.8160	0.8975	0.9010	0.8160	0.9973
MacNoise-NQ (Hong et al., 2024)	0.8866	0.9149	0.9005	0.9045	0.9149	0.8952
MacNoise-TQA (Hong et al., 2024)	0.9231	0.9310	0.9270	0.9194	0.9310	0.9053
AmbigDocs (Lee et al., 2024)	0.9780	0.9175	0.9468	0.9539	0.9175	0.9833
ConflictingQA (Wan et al., 2024)	0.9692	0.9746	0.9719	0.9539	0.9746	0.8608
Overall	0.9681	0.9008	0.9332	0.9280	0.9008	0.9625

Table 6: Conflict detection results on *ConflictBench* using the support-counting setting. Experiments are conducted with `GPT-5.4-mini`. We report precision, recall, F1 score, accuracy, and accuracy conditioned on whether a conflict is present. Recall here is equivalent to Acc_{conf} as they both measure TP/(TP+FN).

Category	Prec	Rec	F1	Acc	Acc_{conf}	Acc_{noConf}
ContraQA (Pan et al., 2023)	1.0000	0.8892	0.9413	0.9411	0.8892	1.0000
MacNoise-NQ (Hong et al., 2024)	0.8438	0.8617	0.8526	0.8593	0.8617	0.8571
MacNoise-TQA (Hong et al., 2024)	0.8926	0.9310	0.9114	0.9005	0.9310	0.8632
AmbigDocs (Lee et al., 2024)	1.0000	0.8935	0.9437	0.9524	0.8935	1.0000
ConflictingQA (Wan et al., 2024)	0.9378	0.9775	0.9572	0.9286	0.9775	0.7089
Overall	0.9583	0.9164	0.9369	0.9311	0.9164	0.9497

Table 7: Conflict detection results on *ConflictBench* using the support-counting setting. Experiments are conducted with `Claude-haiku-4.5`. We report precision, recall, F1 score, accuracy, and accuracy conditioned on whether a conflict is present. Recall here is equivalent to Acc_{conf} as they both measure TP/(TP+FN).

C Partial versus Strict Support

To encourage nuanced reasoning, our prompts treat a document as *supporting* a claim even when it only partially supports the claim. We perform a lightweight ablation to study this design choice by comparing our default *Partial Support* setting against a stricter variant where all references to partial support are removed from the prompts. We evaluate both settings on 100 randomly sampled examples from ConflictBench.

Table 8 shows that both settings achieve similar performance overall. The strict setting slightly improves precision and no-conflict accuracy, while recall and conflict accuracy remain unchanged. These results suggest that allowing partial support does not substantially hurt overall detection quality, but it better supports a more nuanced reasoning of conflicts as demonstrated in Figure 1.

Setting	Prec	Rec	F1	Acc	Acc_{conf}	Acc_{noConf}
Partial Support	0.9836	0.8955	0.9375	0.9200	0.8955	0.9697
Strict Support	1.0000	0.8955	0.9449	0.9300	0.8955	1.0000

Table 8: Ablation study that compares partial v.s. strict support for prompts.

D Examples

Table 9 presents representative error cases identified in our analysis, and Figure 5 provides an end-to-end worked example.

Dataset	Claim	Selected Evidence	Pred	GT	Category
MacNoise-NQ	Surfing is going to be added to the Olympics.	“In April 2008, the IOC began accepting applications for two new sports to be introduced to the Olympic programme, which included baseball and softball (which were dropped in 2005), karate, squash, golf, roller sports, and rugby union all applied to be included. ... ”	Contradict	Support	Ground Truth Incorrect
MacNoise-TQA	Brownsea Island is in Poole Harbour.	“Brownsea Island lies in Christine Ohuruogu opposite the town of Poole in Dorset, England.”	Support	Contradict	Inference Error
ContraQA	A tribute to the fall of Warsaw can be found at the Warsaw Uprising Museum.	“A fine tribute to the fall of Warsaw and history of Poland can be found in the Museum of Modern Art, and in the Polish Uprising Museum which preserves the memory of the crime.”	Support	Contradict	Inference Error
ConflictingQA	"Pled" is a correct past tense of "plead".	“Is the correct past tense pleaded or pled — or perhaps plead? That depends. If you want to be unimpeachably correct, you’ll write pleaded in all past-tense uses. If you’re happy to defend yourself on grounds of “common” usage based on what many others do — despite mountains of contrary authority — you’ll probably use pled ”	Support	Contradict	Local Inconsistency
AmbigDocs	A notable achievement of Thomas Douglas is that he allowed himself to acquire a land grant called Assiniboia to serve as an agricultural settlement for the company.	“Thomas Monteath Douglas General Sir Thomas Monteath Douglas (1787 – October 1868) was an officer of the Bengal Army of the East India Company. He served in a number of wars and campaigns, most notably the First Anglo-Afghan War.”	Irrelevant	Contradict	Entity Ambiguity/ Mismatch

Table 9: Representative examples for each error category identified in our error analysis. The selected evidence presents parts of the evidence document that is the most relevant to the claim. Each example is drawn directly from the analyzed datasets and illustrates a distinct source of error beyond surface-level model mistakes.

Instruction. Answer the query using the given documents in a single claim.

Query. Is IPv6 fundamentally more secure than IPv4?

Doc.	Salient evidence from retrieved documents
D0	"... The computer becomes addressable from the internet, much like a server. ... There is, therefore, a slightly increased security risk to home users using IPv6. ..."
D1	"... IPv6's improved security features, such as end-to-end encryption and support for the Internet Protocol Security (IPSec), make it a preferable protocol for preventing Man-in-the-Middle (MitM) attacks. ... it is considered safer than IPv4. ..."
D2	"... IPv6 is fundamentally more secure than IPv4. ... IP Security (IPsec) ... was optional in IPv4 but has been made mandatory in IPv6. ..."
D3	"... encryption and integrity-checking ... is a standard component in IPv6 ... IPv6 also supports the Secure Neighbor Discovery (SEND) protocol ... it still offers a much improved level of security for connections. ..."
D4	"... It is our recommendation that IPv6 support packet authentication as a basic and required function. ... We recommend that support for the Privacy Header be required in IPv6 implementations. ... Clearly, a key management infrastructure will be required ..."
D5	"... IPv4 was not originally thought of as a secure protocol. IPv6, by contrast, was designed from the very beginning in terms of protection. IPv6 encrypts traffic and checks packet integrity ..."

Initial RAG response (`claim_gpt_54_mini_rag_step3`). "No, IPv6 is not fundamentally more secure than IPv4."

Extracted atomic claim. c_1: IPv6 is not fundamentally more secure than IPv4.

	D0	D1	D2	D3	D4	D5	Aggregation
Initial claim c_1	SUP.	CONTR.	CONTR.	CONTR.	CONTR.	CONTR.	$\lvert D^+\rvert = 1$, $\lvert D^-\rvert = 5$; both non-empty, so c_1 is conflicted.

Metric	Value for the initial response
ConflictScore-Count (CS-C)	$1/1 = 1.00$ because the only extracted claim has both supporting and contradicting evidence.
ConflictScore-Ratio (CS-R)	$5/(1+5) = 0.83$ because five of the six relevant document labels contradict the claim.

Conflict-aware rewrite after feedback. IPv6 is often described as having security advantages over IPv4 because features like IPsec and related protections are more built-in or standardized, but the documents also suggest this does not make it unconditionally or fundamentally more secure in all cases, since practical security still depends on configuration, firewalls, and implementation.

Figure 5: End-to-end worked example for ConflictScore on a ConflictingQA query. The initial response commits to a single side of the evidence, producing one claim that is supported by one document but contradicted by five documents. ConflictScore therefore assigns high conflict scores, and feeding this signal back into generation yields a rewrite that preserves the security advantages described by most documents while acknowledging the configuration-dependent exposure risk raised by the conflicting document. For simplicity, this example shows a model response with only one claim. This differs from the exact settings in Section 5, where responses are longer and contain multiple claims, and Section 6, where the answer is a binary choice. We include this simplified example to illustrate that the same framework can generalize across response formats and use cases.

E Prompts

Prompt:

Instruction: Break down the following report into individual claims.

Report: {report}

Please identify each separate claim made in the report. A claim is a factual statement or opinion that expresses a belief or judgment.

List each claim on a new line, starting with "Claims: ".

Table 10: Prompt used for claim decomposition in *ConflictScore*.

Prompt:

Document-Claim Conflict Detection

You are a careful fact-checking assistant specializing in identifying conflicts between claims and evidence in documents.

Decide one of the following labels for how the document relates to the claim:

1. SUPPORTS – The document provides evidence that directly supports any part of the claim. This includes partial support: if a claim contains hedged or multi-part statements (e.g., "Coffee consumption can improve alertness but may also cause sleep disruption"), then documents that support any part of that claim should be labeled SUPPORTS.

2. CONTRADICTS – The document provides evidence that contradicts the claim (e.g., assigns an incompatible role/date/quantity/polarity, or states the opposite of an asserted relationship), even if it does not explicitly say the claim is false.

3. IRRELEVANT – The document does not provide information about the substantive content of the claim (mentions entities without giving information that could support or contradict) OR provides only vague/contextual information insufficient to judge the claim.

Decision rules:

- Look for statements in the document that address the same attributes asserted (who/what/when/where/how many/etc.) inside the claim.

- If the document contains a statement that supports the claim or partially supports any part of a hedged claim, choose SUPPORTS.

- If the document contains a statement incompatible with the claim, choose CONTRADICTS (do NOT choose IRRELEVANT).

- If it contains neither a compatible nor incompatible statement about the claim's asserted attributes, choose IRRELEVANT.

Example (contradiction via incompatible attribute):

- Claim: "Frédéric Chopin was a famous musician."

- Document: "Polish scientists born in Warsaw include ... Frédéric Chopin."

- Label: CONTRADICTS (profession mismatch: scientist vs musician).

Now evaluate the following:

Claim to evaluate

{claim}

Document to evaluate against

{document}

Output format

Return ONLY a single JSON object with these fields:

{

"claim": "{claim}",

"document_snippet": "<the most relevant snippet from the document that relates to the claim>",

"reasoning": "<a brief justification that explains your decision>",

"answer": "<SUPPORTS|CONTRADICTS|IRRELEVANT>"

}

Please be precise and follow the decision rules. Do not include any additional text or explanations and only output in the JSON format specified above.

Table 11: Prompt used for evidence evaluation in *ConflictScore*.

Prompt:

Instruction: Write a report about {main_entity} given the following documents. Make your report concise and not longer than 3 paragraphs.

Documents:

{Documents}

Table 12: Prompt used for report generation in the retrieval-augmented (RAG) setting.

Prompt:

Instruction: Given the following documents, write a report about {main_entity}.

Please be cautious and hedge accordingly. Think through all the information and consider all perspectives if they exist to acknowledge any potential conflicts or different viewpoints on this topic.

Documents:

{Documents}

Make your report concise and not longer than 3 paragraphs.

Table 13: Prompt used for balanced report generation in the retrieval-augmented (RAG-Balanced) setting.

Prompt:

Instruction: You are a careful and balanced reporting assistant. Given the following documents, write a 3-paragraph report about {main_entity}.

Your goal is to synthesize the information from the documents while minimizing factual conflicts. To do this:

– Hedge appropriately when the evidence is partial, uncertain, or varies across documents. Use phrases like "some sources suggest," "it is possible that," "there is limited evidence," or "while one document indicates...".

– Avoid making definitive claims unless they are clearly and consistently supported across multiple documents.

– Do not speculate or infer beyond what is stated in the documents.

– Attribute information to the documents when relevant (e.g., "According to one article..." or "Document A notes that...").

– Acknowledge conflicting viewpoints if they appear in the documents.

Documents:

{Documents}

Table 14: Prompt used for super-balanced report generation in the retrieval-augmented (RAG-Super-Balanced) setting.

Prompt:

Given the following documents retrieved from the web, answer the question by choosing either (A) or (B).

Documents:

{context_text}

Question: {question}

(A) {option_a}

(B) {option_b}

Please select either (A) or (B) as your answer. Respond with only the letter in parentheses, for example: (A) or (B).

Answer:

Table 15: Prompt used for multiple-choice question answering in the TruthfulQA RAG setting.

Prompt:

Given the following documents retrieved from the web, answer the question by choosing either (A) or (B). Pay careful attention to any conflicts or contradictions between the documents. Think critically about the information provided and prioritize accuracy and truthfulness in your response. If there are conflicting claims, think carefully about the information provided and choose the most reliable answer.

{context_text}

Question: {question}

(A) {option_a}

(B) {option_b}

Please select either (A) or (B) as your answer. Respond with only the letter in parentheses, for example: (A) or (B).

Answer:

Table 16: Prompt used for conflict-aware multiple-choice question answering in the TruthfulQA Control-RAG setting.

Prompt:

You are being asked to reconsider and improve your previous binary choice response to a question. Your original response contained conflicting information based on the provided documents.

ORIGINAL QUESTION: {question}

DOCUMENTS:
{context_text}

YOUR ORIGINAL RESPONSE:
{original_response}

YOUR MAIN CLAIM THAT HAD CONFLICTS:
{selected_claim}
{conflict_details}

INSTRUCTIONS:
1. Carefully re-examine all the provided documents.
2. Pay special attention to the conflicting evidence identified above.
3. Think critically about which sources are most reliable and authoritative.
4. Consider the quality, recency, and credibility of different sources.
5. Prioritize accuracy and truthfulness over simply agreeing with the majority of sources.

OUTPUT FORMAT (strict):
1) First line: exactly one of:
  – A – {option_a}
  – B – {option_b}

2) Second line: Justification: one concise sentence citing document numbers only in parentheses, e.g., (Docs 2, 5).

Do not include chain-of-thought, lists, or extra sections. Return only the two lines described above.

Table 17: Prompt used for conflict-aware response regeneration in the multiple-choice TruthfulQA RAG setting.