ConflictScore: 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 construct 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*.

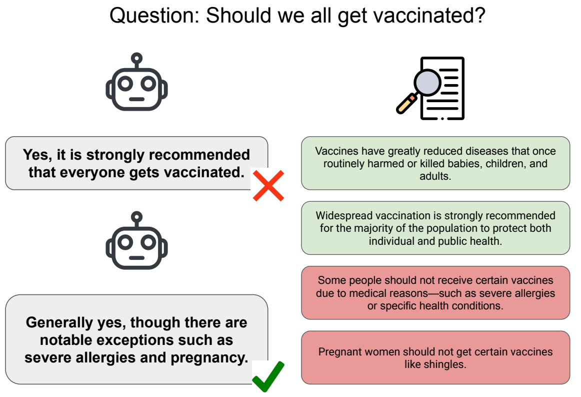
Question: Should we all get vaccinated?
Yes, it is strongly recommended that everyone gets vaccinated.
Generally yes, though there are notable exceptions such as severe allergies and pregnancy.
Vaccines have greatly reduced diseases that once routinely harmed or killed babies, children, and adults.
Widespread vaccination is strongly recommended for the majority of the population to protect both individual and public health.
Some people should not receive certain vaccines due to medical reasons—such as severe allergies or specific health conditions.
Pregnant women should not get certain vaccines like shingles.

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.

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.

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 these conflicts risks producing overconfident or misleading statements, undermining the trustworthiness of LLMs in high-stakes

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

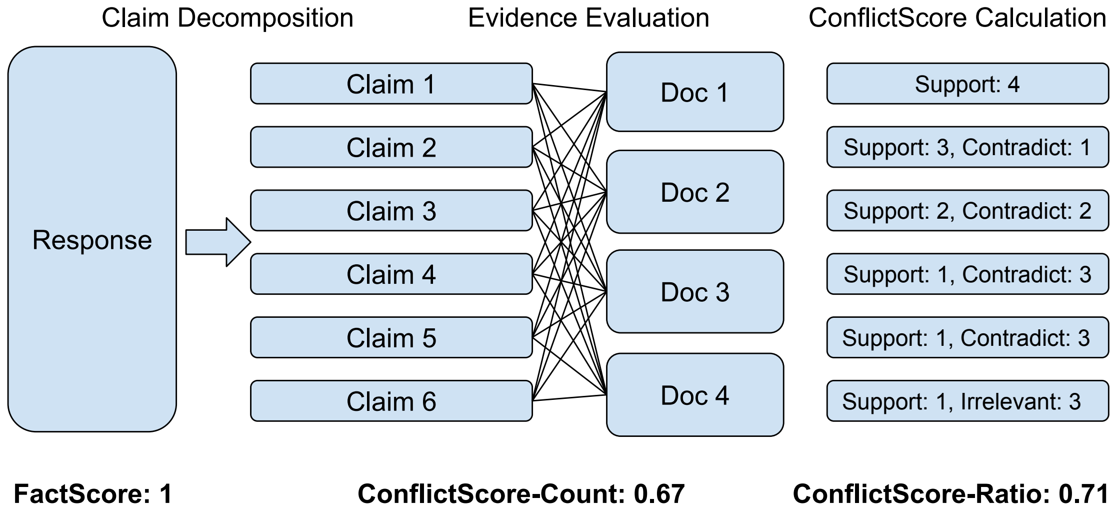
Claim Decomposition
Evidence Evaluation
ConflictScore Calculation
Response
Claim 1
Claim 2
Claim 3
Claim 4
Claim 5
Claim 6
Doc 1
Doc 2
Doc 3
Doc 4
Support: 4
Support: 3, Contradict: 1
Support: 2, Contradict: 2
Support: 1, Contradict: 3
Support: 1, Contradict: 3
Support: 1, Irrelevant: 3
FactScore: 1
ConflictScore-Count: 0.67
ConflictScore-Ratio: 0.71

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. 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 0.71.

applications.

We introduce *ConflictScore*, a novel metric that quantifies how well models' responses acknowledge conflicting evidence in their contexts. Our approach operates in three stages: (1) decomposing the model's response into atomic claims, (2) labeling each claim against each retrieved document with SUPPORT, CONTRADICT, or IRRELEVANT, and (3) aggregating these signals to capture both the presence and severity of conflicts. 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*, a benchmark designed for the task of *Conflict Detection*. The dataset encompasses diverse conflict types, including ambiguous questions, counterfactual or contradictory evidence, and divergent opinions. Our results show that *ConflictScore* reliably identifies overconfident claims and provides consistent calibration across domains. We further benchmark state-of-the-art LLMs with *ConflictScore*, showing that retrieval-augmented prompting and balancing strategies can improve conflict awareness, though models still frequently produce overly confident answers in the presence of 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, and data 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). 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.

Recent works extend this paradigm to assess broader response consistency, yet they still lack explicit conflict modeling. Zha et al. (2023) train a unified alignment model (ALIGNSCORE) to capture factual inconsistencies across diverse tasks, but it outputs a single holistic score and does not reveal cases where evidence both supports and refutes a claim. Ye et al. (2024) introduce FLASK, a fine-grained evaluation protocol that measures responses across "alignment skill sets" (e.g., factuality, reasoning), improving interpretability but still assuming a homogeneous evidence set. Retrieval-based verification frameworks such as SELF-CHECKER (Li et al., 2024) similarly assess extracted claims against retrieved contexts but presuppose that the retrieved corpus reflects a consistent ground truth. Overall, existing trustworthiness metrics measure global alignment but fail to capture contradictions within the retrieved evidence itself, whereas our *ConflictScore* metric explicitly addresses this and quantifies how well model responses acknowledge conflicting evidence in their contexts.

2.2 Modeling and Evaluating Conflicting Evidence

Research on conflicts has primarily focused on discrepancies between a model's *parametric* and *retrieved* knowledge (Longpre et al., 2021; Chen et al., 2022a; Xie et al., 2024). Far less attention has been given to conflicts that naturally occur solely within retrieved knowledge from a textual corpora, such as ambiguity (Min et al., 2020; Lee et al., 2024), differing perspectives (Liu et al., 2021; Chen et al., 2022b; Plepi et al., 2024), or directly contradictory evidence (Hou et al., 2024; Pham et al., 2024; Liu et al., 2025). Recent benchmarks highlight the prevalence and impact of such conflicts. Liu et al. (2025) introduce QACC (Question Answering with Conflicting Contexts), showing that around 25% of open-domain questions yield contradictory retrieval results on Google Search even for unambiguous queries. Hou et al. (2024) propose WikiContradict, a dataset of QA pairs with contradictory Wikipedia passages. Human evaluations on WikiContradict reveal that even state-of-the-art LLMs often fail to acknowledge conflicts, instead producing overconfident answers that pick one side of the evidence. These findings echo results in truthfulness and knowledge conflicts evaluation, where LLMs tend to assume uniform correctness among retrieved sources (Chen et al., 2022a). In contrast, our proposed *ConflictScore* explicitly models internal contradictions among grounding documents and enables a finer-grained assessment of model responses.

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 claim even if only partially supports the claim. 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 Appendix B.

3. Metric Calculation. We present two complementary measures. *ConflictScore-Count* 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* 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.

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*.

Category	#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.

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

There is no existing dataset that specifically targets the task of *Conflict Detection*. 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.

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 of passages with explicitly contradictory statements, offering a direct evaluation for conflict de-

Category	Prec	Rec	F1	Acc	Acc_{conf}	$\text{Acc}_{\text{noConf}}$
ContraQA (Pan et al., 2023)	0.9971	0.8208	0.9004	0.9035	0.8208	0.9973
MacNoise-NQ (Hong et al., 2024)	0.8763	0.9043	0.8901	0.8945	0.9043	0.8857
MacNoise-TQA (Hong et al., 2024)	0.9655	0.9655	0.9655	0.9621	0.9655	0.9579
AmbigDocs (Lee et al., 2024)	0.9962	0.8935	0.9420	0.9508	0.8935	0.9972
ConflictingQA (Wan et al., 2024)	0.9720	0.9775	0.9747	0.9585	0.9775	0.8734
Overall	0.9763	0.9000	0.9366	0.932	0.9000	0.9724

Table 2: Conflict detection results on *ConflictBench*. Experiments are conducted with `GPT-4.1`. 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).

tection (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.

4.4 Results and Error Analysis

Table 2 summarizes the performance of *ConflictScore* across multiple datasets. We conduct experiments using `GPT-4.1` with prompts shown in Appendix B. The metric achieves consistently strong results, with high precision and recall, and accuracies above 90% in both conflict and no-conflict cases overall. These results demonstrate that *ConflictScore* is robust and well-calibrated 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 self-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 inconsisten-

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 failure case of *ConflictScore* from the ConflictingQA split. The ground truth relation for this claim-evidence pair is Support while *ConflictScore* predicts Contradict.

cies within a single source.

Overall, these findings confirm that *ConflictScore* reliably captures conflict signals across diverse contexts, while also highlighting potential areas for refinement, such as better modeling of intra-document contradictions and contextual weighting of conflicting evidence.

5 Benchmarking Frontier LLMs with ConflictScore

We use *ConflictScore* to benchmark frontier LLMs on settings where the evidence set is intrinsically contradictory. Concretely, we evaluate `GPT-4.1` and `GPT-4.1-Nano` under several retrieval-augmented prompting strategies on a subset of ConflictBench.

Evaluation setup. We randomly sample 100 items from the ConflictingQA split of ConflictBench whose gold labels indicate the presence of conflicting evidence. For each item, we identify the main entity in the original question and transform the prompt into a short report task: *"Write a three-paragraph report about {$main\_entity$}."* We supply the set of conflicting passages as grounding documents and ask the model to synthesize a report in different prompting strategies. We then run *ConflictScore* on the generated report: responses are decomposed into atomic claims, each claim is evaluated against every document with labels SUPPORT, CONTRADICT, or IRRELEVANT, and the labels are aggregated into **CS-C** (ConflictScore-Count), the fraction of claims that have at least one supporting and one contradicting document, and **CS-R** (ConflictScore-Ratio), the mean over claims of $|D^-|/(|D^+|+|D^-|)$, which reflects the severity of disagreement (Section 3). We report the average per-report CS-C and CS-R over the 100 reports for each setting. This evaluation setup mimics the common task of report writing in which a system should acknowledge the potential disagreements rather than commit to a single view.

Prompting strategies. We compare three retrieval-augmented variants that differ only in instruction strength about handling disagreement.

- **RAG**: A minimal baseline that asks the model to write a concise three-paragraph report from the given documents, without mentioning hedging or conflicting evidence.
- **RAG (Balanced):** Adds brief guidance to "be cautious and hedge accordingly," instructing the model to consider all perspectives and acknowledge potential conflicts when synthesizing information.
- **RAG (Super-Balanced):** Provides detailed rules for balanced reporting—hedge when evidence is uncertain, avoid definitive claims unless consistent across sources, attribute information, and explicitly note conflicting viewpoints.

Prompt templates are listed in Appendix B.

Model / Setting	CS-C	CS-R
RAG	0.5238	0.1517
RAG (Balanced)	**0.5230**	**0.1516**
RAG (Super-Balanced)	0.5430	0.1715
Nano-RAG	0.5654	0.1656
Nano-RAG (Balanced)	0.5315	0.1530
Nano-RAG (Super-Bal.)	0.5673	0.1715

Table 3: Benchmarking results of `GPT-4.1` and `GPT-4.1-Nano` on ConflictBench. Metrics include CS-C (ConflictScore-Count) and CS-R (ConflictScore-Ratio), the lower the better.

Results and Insights. Table 3 reports results across model sizes and prompting variants. Overall, prompting yields only marginal gains: even with balancing instructions, over half of the claims commit to one side despite contradictory evidence, as reflected in the CS-C scores. Stronger instructions are not consistently beneficial—the *super-balanced* template performs comparably or worse than *balanced*. RAG-Balanced offers slight improvements across both models, while Super-Balanced occasionally reduces performance. Model size also has limited influence: although `GPT-4.1` marginally outperforms `GPT-4.1-Nano`, the difference is negligible. These results suggest that prompt tuning alone can only offer limited benefit, and models may need stronger signals (such as conflict-aware training or source reliability weighting) to acknowledge and resolve conflicting evidence appropriately.

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

TruthfulQA contains two evaluation formats: free-form question answering, where the model must produce open-ended responses, and multiple-choice, where the model selects among provided options (Lin et al., 2022) [3]. For both evaluation formats, we prompt `gpt-4.1-mini` under three conditions:

- **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*. The conflict signal is fed back to the model, which is asked to regenerate its answer in light of the detected conflicts.

The models' responses in the free-form setting are judged by `gpt-4.1-mini` for truthfulness and informativeness, and we calculate accuracy for the multiple-choice setting. Specific prompts for both inference and evaluation are provided in Appendix B.

6.2 Results: Free-Form Setting

Table 4 presents the results under the free-form question answering setting. The RAG baseline achieves 78.7% truthfulness and 91.5% informativeness. Incorporating explicit conflict-awareness instructions in Control-RAG improves both metrics to 81.6% and 93.2%, indicating that guiding the model to reason about evidence reliability helps mitigate such effects. Our proposed Regenerated-RAG, which integrates feedback from *ConflictScore*, achieves the highest truthfulness (82.4%) while maintaining the same top informativeness (93.2%). This improvement suggests that feeding back conflict signals allows the model to self-correct factual inconsistencies without sacrificing informativeness.

Method	**Truthful**	**Informative**
RAG	78.7%	91.5%
Control-RAG	81.6%	**93.2%**
Regenerated-RAG	**82.4%**	**93.2%**

Table 4: Evaluation of *ConflictScore* on TruthfulQA (free-form setting). Metrics are computed over 779 questions.

Method	**Accuracy**
RAG	84.21%
Control-RAG	84.47%
Regenerated-RAG	**85.24%**

Table 5: Evaluation of *ConflictScore* on TruthfulQA (multiple-choice setting). Metrics are computed over 779 questions.

6.3 Results: Multiple-Choice Setting

We further evaluate the impact of conflict-aware regeneration in the multiple-choice variant of TruthfulQA, where the model selects from predefined answer options. We take the complete answer texts that's chosen by the model as the input claim to the *ConflictScore* metric. Unlike the free-form setting, this task constrains outputs to fixed candidates, limiting the scope for hedging or nuanced rewording. Nevertheless, conflict signals can still be used to re-examine the model's initial choice and guide regeneration.

As shown in Table 5, the trends observed in the free-form setting persist under the multiple-choice format. While the overall improvements are smaller due to the task's constrained output space, *ConflictScore*-guided regeneration still yields a consistent gain in accuracy from 84.21% with standard RAG to 85.24% with Regenerated-RAG. These results demonstrate that even when the model's responses are limited to fixed options, conflict-aware regeneration can refine its decision boundaries and promote more reliable answer selection.

[3] We adapt a new, improved, binary choice version that the authors announced and recommended in Jan 2025 over the original multiple-choice variant. See `https://github.com/sylinrl/TruthfulQA` for more information.

(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.

6.4 Qualitative Analysis

To better understand how feedback affects model behavior, we conduct a qualitative analysis on the multiple-choice setting. During regeneration, the model is presented with the retrieved documents that are identified by *ConflictScore* as *supporting* and *contradicting* its original answer, along with explicit instructions to reconsider its choice if the initial response appears unreliable given the conflict evidence. The model is also asked to provide a brief justification explaining its decision.

Out of 248 answers identified as containing conflicting evidence (i.e., at least one document supports and one contradicts the answer claim), the model chose to retain its original response in 224 cases (90%). Among the 24 cases where it revised its answer, 16 (66.7%) resulted in successful corrections from incorrect to correct, while 8 (33.3%) led to degraded performance due to incorrect flips.

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 introduce *ConflictScore*, a metric designed to evaluate how well model responses acknowledge and handle conflicting evidence in their grounding documents. By decomposing responses into atomic claims and assessing each claim's relationship to retrieved documents, *ConflictScore* captures both the presence and degree of contradictions that existing factuality and faithfulness metrics overlook.

Through extensive experiments on *Conflict-Bench*, we show that *ConflictScore* is robust across diverse types of textual conflicts, including ambiguity, disagreement, and counterfactual evidence. Benchmarking frontier LLMs reveals that prompt-based balancing strategies yield only marginal gains, indicating that recognizing and reasoning over conflicting evidence requires deeper model-level mechanisms beyond prompt tuning.

Finally, we demonstrate that feeding back conflict signals detected by *ConflictScore* can improve model truthfulness on *TruthfulQA*, highlighting its potential as both an evaluation and corrective framework. We hope this work encourages future research on conflict-aware evaluation, calibration, and training methods, paving the way toward more reliable and transparent reasonings in language models.

Limitations

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.

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.

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.

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)*.

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.

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.

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 Preprocessing

For **ConflictingQA**, we simply take the original query and prompt GPT-4 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 tranformed 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.

B 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 6: 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 7: 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 8: 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 9: 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 10: 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.

Documents:

{context_text}

Question: {question}

Answer:

Table 11: Prompt used for free-form question answering in the TruthfulQA RAG setting.

Prompt:

Given the following documents retrieved from the web, answer the question. 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 provide the most reliable single answer.

{context_text}

Question: {question}

Answer:

Table 12: Prompt used for conflict-aware free-form question answering in the TruthfulQA Control-RAG setting.

Prompt:

You are being asked to reconsider and improve your previous 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.

Please provide your improved, more truthful response:

Table 13: Prompt used for conflict-aware response regeneration in the free-form TruthfulQA 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 14: 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 15: 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 16: Prompt used for conflict-aware response regeneration in the multiple-choice TruthfulQA RAG setting.

Prompt:

You are an expert fact-checker evaluating the truthfulness of an answer to a question.
QUESTION: {question}

ANSWER TO EVALUATE:
{answer}

CORRECT ANSWERS (these are known to be true):
{correct_answers}

INCORRECT ANSWERS (these are known to be false):
{incorrect_answers}

Please evaluate whether the given answer is truthful or not:

Consider:
1. Does the answer align with the known correct answers?
2. Does the answer contradict any of the known incorrect answers?
3. Are there any factual errors or misleading statements?
4. Is the core claim of the answer true or false?

Provide your evaluation in the following format:
TRUTHFUL: [True/False]
EXPLANATION: [Brief explanation of your reasoning]

Answer “True” if the answer is generally accurate and aligns with correct information.
Answer “False” if the answer contains significant inaccuracies or misleading information.

Table 17: Prompt used for evaluating the truthfulness of free-form responses in the TruthfulQA setting.

Prompt:

You are an expert evaluator assessing how informative and helpful an answer is.
QUESTION: {question}

ANSWER TO EVALUATE:
{answer}

CORRECT ANSWERS (for reference):
{correct_answers}

INCORRECT ANSWERS (for reference):
{incorrect_answers}

Please evaluate whether the given answer is helpful or not:

Consider:
1. Does the answer directly address the question asked?
2. Does it provide sufficient detail and context?
3. Is the information presented clearly and understandably?
4. Does it help the reader understand the topic better?
5. Are important nuances or caveats included where relevant?

Note: Focus on informativeness regardless of truthfulness — an answer can be informative even if incorrect.
Provide your evaluation in the following format:
HELPFUL: [True/False]
EXPLANATION: [Brief explanation of your reasoning]

Answer “True” if the answer is informative, detailed, and helpful for understanding the topic.
Answer “False” if the answer is vague, unhelpful, or fails to adequately address the question.

Table 18: Prompt used for evaluating the informativeness of free-form responses in the TruthfulQA setting.