Check for updates

Received: 27 March 2026 | Revised: 24 July 2026 | Accepted: 11 August 2026

DOI: 10.4218/etrij.2026-0242

ETRI Journal WILEY

ORIGINAL ARTICLE

Special Issue

Diagnosing judgment flip risk of large language models in fact verification under conflicting evidence

Jihyeon Roh | **Kyoungman Bae**

Vision Intelligence Research Section, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea

Correspondence
Jihyeon Roh, Vision Intelligence Research Section, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea.
Email: jihyeon.roh@etri.re.kr

Funding information
Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT)
(No. 2022-0-00369, (Part 4) Development of AI Technology to support Expert Decision-making that can Explain the Reasons/Grounds for Judgment Results based on Expert Knowledge).

Abstract

Large language model (LLM)-based fact verifiers can encounter conflicting evidence in retrieval-augmented pipelines, but their robustness to evidence composition remains underexamined. We introduce a flip-rate diagnostic protocol that measures how often initially correct binary judgments become incorrect when the supporting and refuting evidence is perturbed. Using six models and two datasets (FEVER and SciFact), we vary the evidence ratio, quantity, and ordering and evaluate chain-of-thought (CoT) prompting and three lightweight mitigations. On FEVER, one refuting passage alongside three supporting passages flips correct judgments in 3.3%–22.2% of cases; under full conflict, flip rates exceed 60% for every model on FEVER and 90% on SciFact. For most models, placing refuting evidence later also increases flip rates. CoT reduces flips under full conflict but is unreliable under mixed conflict. Conflict-aware stance-grouped prompting is the most consistent mitigation across datasets. These findings indicate that baseline accuracy alone is insufficient for assessing LLM fact-verifier reliability and that evidence composition should be treated as a separate axis of robustness evaluation.

KEYWORDS
conflicting evidence, fact verification, judgment flip, large language models, robustness evaluation

1 | INTRODUCTION

Large language model (LLM)-based fact verification is being increasingly adopted, often as part of retrieval-augmented generation (RAG) pipelines that condition a model's verdict on the retrieved evidence [1]. However, in realistic deployment settings, retrieved evidence is rarely clean because search noise, document staleness, and editorial inconsistencies across sources can naturally yield *conflicting evidence* for the same claim. Moreover, natural language processing systems have repeatedly been shown to be vulnerable to subtle input perturbations [2], and the possibility of *evidence poisoning*, the injection of manipulated documents into retrieval corpora, has been discussed as a practical threat [3, 4]. As LLM-based systems are deployed in high-stakes decision-

This is an Open Access article distributed under the term of Korea Open Government License (KOGL) Type 4: Source Indication + Commercial Use Prohibition + Change Prohibition (http://www.kogl.or.kr/info/licenseTypeEn.do).
1225-6463/$ © 2026 ETRI

making domains, such as content moderation and misinformation detection, it becomes critical to assess their trustworthiness beyond average accuracy.

Fact verification is particularly sensitive to such reliability concerns. Unlike open-ended question answering, where conflicting contexts may lead to partially correct or hedged responses (primarily a quality degradation issue), fact verification enforces a binary verdict (TRUE/FALSE) that often maps directly to downstream actions. For example, in content moderation, a flip between TRUE and FALSE can translate into a difference between removing lawful speech and allowing misinformation to spread. Therefore, if a system's judgment changes depending on how the evidence is presented, even when the underlying evidence content is held fixed, the system may be unreliable for deployment.

Most research on fact verification has evaluated models using standard metrics such as accuracy or F1. However, high average performance does not guarantee stable behavior under conflict, and verdicts in mixed-evidence settings may become highly sensitive to small perturbations in evidence composition, including the number of passages, their ordering, and the support–refute ratio. Such *judgment instability* poses a critical concern for deployment yet remains invisible under standard metrics.

Prior work in question answering and retrieval-augmented settings has examined how models use retrieved evidence for reasoning [5, 6] and how their behavior changes when such evidence conflicts with other information [7–9]. Although these studies provided important insights, they did not systematically evaluate the risk of *judgment flips* in binary decision tasks. More recent work has begun to consider conflicting evidence in fact verification [10]; however, there remains a lack of integrated analysis that (i) diagnoses which models are vulnerable to which conflict patterns and (ii) evaluates lightweight, model- and condition-appropriate mitigations.

To bridge this gap, we present a diagnostic study that quantifies when and how severely LLM fact verifiers break down under conflicting evidence. Rather than proposing a conflict-resolution method, we introduce a controlled stress test that compares a no-conflict baseline (gold-consistent evidence only) with controlled evidence perturbations that mix the supporting and refuting passages (Figure 1) and measures the risk that *baseline-correct* judgments become incorrect using the correct→incorrect flip rate.

Our stress test isolates three orthogonal drivers of judgment instability: (1) **Evidence ratio**, varied across eight support–refute compositions to identify where instability emerges; (2) **evidence quantity**, varied by changing the total number of passages while holding the ratio fixed to separate quantity effects; and (3) **evidence ordering**, varied by permuting the presentation order of an identical evidence set to quantify position bias. We further conduct a **chain-of-thought (CoT) ablation** to

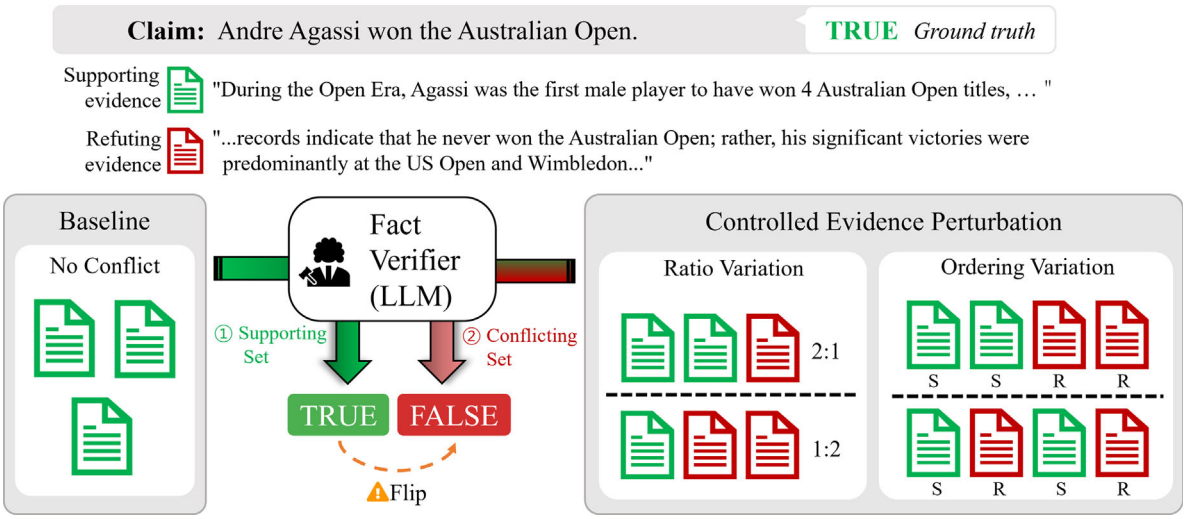

FIGURE 1 Overview of the experimental framework. Given a claim and a set of evidence passages, an LLM-based fact verifier outputs a binary TRUE/FALSE verdict. In the baseline condition (①), only supporting evidence is provided, and the model produces a correct verdict. In the controlled evidence perturbation condition (②), we construct a conflicting evidence set by mixing supporting and refuting passages, with controlled perturbations along two dimensions: *ratio variation*, which changes the support-to-refute composition, and *ordering variation*, which rearranges the same passages in different sequences. A *judgment flip* occurs when a judgment that was correct in the baseline condition becomes incorrect under perturbation. Green and red icons denote supporting and refuting evidence, respectively.

22337326, 2026, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.4218/etrij.2026-0242 by Electronics and Telecommunications, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

examine whether explicit reasoning elicitation provides a consistent safeguard under conflict. Finally, to assess practical mitigations, we evaluate three training-free, **lightweight defenses**—conflict-aware prompting (CA), evidence-level independent judgment (EIJ), and conflict-aware stance-grouped prompting (CA+SGP)—and synthesize the results to characterize model-specific vulnerability patterns and inform defense selection.

Our contributions can be summarized as follows:

- We introduce a flip-rate diagnostic protocol for LLM-based fact verification under conflicting evidence. The protocol evaluates whether initially correct TRUE/FALSE judgments become incorrect as evidence ratio, quantity, and ordering are systematically varied.
- We provide an empirical evaluation across six models from four families and two datasets, FEVER and SciFact, revealing ratio-driven vulnerability, full-conflict fragility, and late-evidence weighting.
- We evaluate CoT prompting and three lightweight mitigation strategies (CA, EIJ, CA+SGP). We find that CA+SGP provides the most consistent flip-rate reductions across models and datasets, whereas CoT and EIJ are effective only under specific model, conflict, or domain conditions.

2 | RELATED WORK

2.1 | Robustness of LLM verifiers

With the growing adoption of LLM-based fact verification [1, 11] in retrieval-augmented pipelines [12], there has been increasing interest in improving robustness to variation in the quality of retrieved evidence [13, 14]. Self-RAG [15] aims to reduce the impact of noisy evidence by selectively deciding when retrieval is needed and by critiquing generations via self-reflection. FIRE [6] proposes an iterative pipeline that interleaves retrieval and verification, adaptively deciding whether additional retrieval is warranted. Pan et al. [16] highlighted the risk that LLM-generated misinformation can pollute downstream knowledge sources and proposed defenses such as misinformation detection and vigilant prompting.

A related line of work examines how LLMs are influenced by input framing and surface-level properties beyond factual correctness, including sycophancy toward user-stated beliefs [17] and style-over-substance evaluation bias [18]. These studies share the concern that LLM judgments can be swayed by such input characteristics; our setting differs in that it does not involve user-stated beliefs and uses quality-screened evidence pools to better focus on evidence-composition effects.

Unlike prior robustness studies that primarily target end-to-end accuracy or specific threat models, we diagnose how easily baseline-correct binary verdicts flip under controlled evidence composition changes.

2.2 | Conflicting evidence in LLMs

Recent studies have actively investigated knowledge conflicts in LLMs. Xu et al. [19] organized such conflicts into three types: context–memory, inter-context, and intra-memory conflict. Our setting primarily corresponds to inter-context conflict, whereas it may also involve context–memory conflict when the retrieved evidence contradicts a model's prior beliefs. Longpre et al. [7] formalized entity-level knowledge conflicts and demonstrated that models may overrely on parametric knowledge rather than contextual evidence. Xie et al. [9] examined how models resolve conflicts between parametric memory and external evidence and found that evidence coherence and persuasiveness influence model behavior. Our work builds on this line of inquiry but differs in what it measures. We focus on binary fact-verification verdicts and measure paired flips from baseline-correct judgments under a controlled evidence-composition grid that jointly varies ratio, quantity, and ordering.

A related line of work studies contrastive evidence for robust fact verification. VitaminC [20] constructs contrastive claim–evidence pairs from Wikipedia revisions, in which near-identical passages differ only in whether they support or refute a claim, and shows that training on them improves robustness to subtle factual changes. Our study complements this line of work but shifts the focus from robustness training to diagnostic measurement. Rather than training models on contrastive evidence, we use fixed supporting and refuting evidence pools to measure how often the initially correct verdicts flip as the support–refute ratio, quantity, and ordering are varied.

Recent approaches have also attempted to handle conflicting evidence through context-faithful prompting [21], multiagent debate [8], conflict-driven summarization [22], soft-prompting-based conflict detection [23], and source-credibility annotations [10]. However, much of this literature either proposes methods for handling conflicts or studies robustness through training rather than quantifying how easily existing verifiers break under controlled conflict perturbations. Moreover, many studies have been conducted in question answering or free-form generation settings and have not treated judgment flips in binary (TRUE/FALSE) decisions as a first-class deployment risk.

22337326, 2026, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.4218/etrij.2026-0242 by Electronics and Telecommunications, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2.3 | CoT and position bias

Reasoning-elicitation interventions form another relevant line of research. CoT prompting [24] is widely used to improve performance on tasks that require multistep reasoning by encouraging models to generate intermediate reasoning steps. In parallel, the impact of information position in long contexts was carefully analyzed by Liu et al. [25], who showed that performance can vary substantially even under a simple reordering of documents. Position bias mitigation approaches have also been explored, including permutation self-consistency [26], which extends self-consistency [27] by averaging over multiple input orderings to reduce order sensitivity. Our CA+SGP defense shares the motivation to reduce order sensitivity but operates at the stance level by grouping evidence according to stance classification rather than averaging over arbitrary permutations.

However, these studies have largely focused on long-context information use and general reasoning benchmarks. In fact-verification settings where conflicting evidence coexists, it remains unclear whether such interventions consistently improve judgment robustness. It is also unclear whether they can degrade stability under certain conditions, and how such condition dependence varies across models.

Taken together, prior work has studied knowledge conflicts, input-driven and persuasion-related bias, contrastive evidence, position bias, and reasoning-based interventions, often as separate problems. Our contribution is a unified flip-rate diagnostic protocol for binary fact verification. Starting from claims that are initially judged correctly under support-only evidence, we measure how often those verdicts become incorrect under controlled changes in the evidence ratio, quantity, and ordering and evaluate CoT and lightweight defenses under the same paired setup.

3 | EXPERIMENTAL SETUP

In this study, we addressed an evidence-based fact-verification task [28]. Given an input claim and a set of evidence passages, the model outputs a binary verdict indicating whether the claim is TRUE or FALSE. To evaluate judgment robustness in conflicting-evidence settings, we systematically varied the composition of the evidence set–specifically, its ratio, ordering, and quantity–and measured how often a prediction that was correct under a reference condition flipped to an incorrect prediction under these controlled perturbations.

Throughout this section, we define *supporting* and *refuting* evidence with respect to the gold label rather than the surface polarity of the claim text. Supporting evidence is evidence that aligns with the gold label (i.e., it supports the correct decision), whereas refuting evidence is evidence that contradicts the gold label (i.e., it tends to overturn the correct decision). Consequently, for REFUTES claims, *supporting evidence* can be evidence that refutes the claim under our gold-label-based definition.

3.1 | Datasets and evidence construction

Datasets. We randomly sampled 500 SUPPORTS and 500 REFUTES claims from the FEVER dataset [28] and constructed a pool of three supporting and three refuting evidence items per claim using the pipeline described below. Each generated passage was scored on a 1–5 quality-screening scale by Qwen3-32B [29], a model from a family different from that of the generator, and passages scoring below 3 were excluded. Claims for which this quality gate left fewer than three items in either pool were discarded, removing 226 claims. The final set of 774 claims (426 SUPPORTS and 348 REFUTES) was held fixed across all conditions, enabling paired comparisons.

To assess whether the diagnosed patterns generalize beyond Wikipedia-style claims, we used SciFact [30] as a cross-domain validation dataset, constructing evidence sets from SUPPORT and CONTRADICT claims in the scientific domain. We mapped these labels to the SUPPORTS/REFUTES notation used in the experiments. We applied the same evidence construction and quality-screening pipeline. After quality screening, the final SciFact set contained 610 of the 773 candidate claims (339 SUPPORTS, 271 REFUTES), as described in Appendix D. SciFact results are summarized in Section 4.5 with full tables in Appendix D.

Supporting evidence pool. For each claim, we formed a pool of three supporting evidence passages consisting of one dataset-provided gold evidence passage and two paraphrases generated by GPT-4o [31]. For FEVER, the gold passage was obtained from the FEVER evidence annotation, and for SciFact, it was obtained from the corresponding gold evidence annotation. We used a temperature of 0.8 for generation and included previously generated paraphrases in the prompt to encourage lexical and syntactic diversity. Paraphrases scoring below 3 on the quality-screening scale were excluded, as with all passages generated in this study.

Refuting evidence pool. If the quality of the refuting evidence differs substantially from that of the supporting evidence in either direction, then the observed flip rates may reflect a quality mismatch rather than evidence-composition effects. Systematically weaker counter-

22337326, 2026, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.4218/etrij.2026-0242 by Electronics and Telecommunications, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

evidence would underestimate the flip risk, whereas systematically stronger counter-evidence would overestimate it. Therefore, we constructed counter-evidence whose quality-screening score was close to that of the gold evidence, rather than maximizing counter-evidence-quality. For each claim, GPT-4o generated multiple counter-evidence candidates that argued the opposite stance from the gold label and were constrained to one to two sentences to match the typical length of the gold evidence. We then scored the gold evidence and all counter-evidence candidates on the same 1–5 Qwen3-32B quality-screening scale, removed candidates scoring below 3, selected the candidate whose score was closest to the gold evidence, and applied rewriting or regeneration to reduce the remaining gap. Two paraphrases were added to form the final three-item refuting pool. In the FEVER construction, 93.3% of the generated counter-evidence candidates passed the quality threshold.

On the Qwen3-32B automatic quality-screening scale, the final supporting three-item pool (gold evidence plus two paraphrases) had a mean score of 4.67, whereas the final refuting three-item pool (a generated counter-evidence passage plus two refuting paraphrases) had a mean score of 4.13, yielding a residual support–refute gap in automatic scores of 0.54. All means were computed per claim and then averaged across claims. Because this residual gap favors supporting evidence, it would, if anything, tend to make the reported flip rates conservative rather than inflate them because of stronger refuting evidence. Stage-wise gap reduction, final pool decomposition, before/after rewriting examples, and human validation of the evidence pool are reported in Appendix B. For SciFact, the corresponding final pool-level gap was 0.04.

Evidence selection rule. We denote each experimental condition by $s:r$, where s and r indicate the numbers of evidence passages supporting the gold label and refuting the gold label, respectively. For each condition, we selected evidence using a deterministic prefix rule. From the supporting pool [GOLD, PARA1, and PARA2] and the refuting pool [COUNTER-ORIG, COUNTER-PARA1, and COUNTER-PARA2], we took the required number of passages from the front. For example, the 2:1 condition used [GOLD, PARA1] and [COUNTER-ORIG]. This design induces superset relations across conditions (e.g., 3:2 contains 3:1), which helps isolate the marginal effect of added evidence under a fixed prefix-selection rule. For the ordering experiment, we used the fixed orders defined in Section 3.2.

3.2 | Experimental design

We use the following terminology: 3:0 denotes a *support-only* baseline, 0:3 a *refute-only* setting, and conditions that mix both types (e.g., 2:1 and 1:2) are referred to as *mixed-conflict* settings.

Ratio conditions (Section 4.1). We constructed eight ratio conditions: 3:0, 3:1, 3:2, 2:1, 2:3, 1:2, 1:3, and 0:3. By increasing the refuting proportion from 0% to 100%, we examined how judgment instability changed with escalating conflict.

Quantity conditions (Section 4.1). To isolate the effect of evidence *quantity* while maintaining a balanced ratio, we varied only the total number of passages under a fixed 1:1 composition: 1:1 ($n=2$), 2:2 ($n=4$), and 3:3 ($n=6$). The 3:3 setting used the default randomized (fixed-seed) ordering shared across the ratio and quantity experiments, whereas the ordering conditions rearranged the same evidence set according to predefined rules. Performance differences between the default 3:3 condition and the predefined ordering conditions are therefore attributable to the ordering effects.

Ordering conditions (Section 4.2). Under a balanced 3:3 setting, we varied only the *presentation order* of an identical evidence set, where S and R denote individual supporting and refuting passages, respectively. We defined four presentation orders as follows: `support_first` (SF; [S,S,S,R,R,R]), `refute_first` (RF; [R,R,R,S,S,S]), `interleaved` (IL; [S,R,S,R,S,R]), and `interleaved_reverse` (IR; [R,S,R,S,R,S]). Because content and composition are fixed, differences across these conditions isolate the effect of ordering under conflicting evidence.

CoT ablation (Section 4.3). We ran all conditions with both a BASIC prompt and a CoT prompt to examine whether explicit reasoning elicitation improves judgment robustness under conflict [24].

Mitigation strategies (Section 4.4). To examine whether the vulnerabilities identified in Sections 4.1–4.3 could be mitigated, we evaluated three prompt- and pipeline-level defense strategies. Their design was motivated by the experimental findings and is described in detail in Section 4.4.

3.3 | Models and prompts

Models. We evaluated six models spanning four families: GPT-4o [31], GPT-5.4 [32], Gemini-3.1-Pro [33], Llama-3.1-70B-Instruct, Llama-3.1-8B-Instruct [34], and Qwen3-32B-Instruct [29] (hereafter Llama-70B, Llama-8B, and Qwen-32B, respectively). This selection covers three closed-source models (GPT-4o, GPT-5.4, and Gemini-3.1-Pro), two large-scale open-source models from different families (Llama-70B and Qwen-32B), and a smaller model (Llama-8B) to examine the effects of model scale and generation. The open-source models

22337326, 2026, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.4218/etrij.2026-0242 by Electronics and Telecommunications, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

were executed using a vLLM inference server [35]. Unless otherwise specified, we used a temperature of 0 to reduce stochastic decoding variability and improve reproducibility. The model-specific inference settings, including reasoning-mode configurations, are detailed in Appendix A.

Prompting. Across all experimental conditions in Sections 4.1–4.3, we used a unified input format that lists a claim and an enumerated set of evidence passages. Under the BASIC prompt, the model must output a single-word verdict (TRUE or FALSE) based on the evidence provided. Under the CoT variant, we elicited step-by-step reasoning but constrained the final line to a single-word label (TRUE/FALSE). Except for the CoT prompt, all final verdict verification calls used a direct single-word TRUE/FALSE output format without eliciting intermediate reasoning, thereby isolating the effect of evidence composition on verdict formation. We treated CoT prompting as an ablation study in Section 4.3 to assess whether explicit reasoning elicitation improves judgment robustness under conflicting-evidence conditions. The mitigation strategies in Section 4.4 modify the prompting format as described therein. For output parsing, we extracted TRUE or FALSE from the BASIC response and the final line of the CoT response; if a valid label could not be extracted, we treated the instance as incorrect. All prompt templates used in this study, including evidence generation, quality screening, and defense prompts, are provided in Appendix A.

3.4 | Evaluation metrics

Flip rate (primary metric). For each model m, we define the *reference set* as the claims answered correctly under the baseline condition of 3:0 (support only) *before* applying CoT prompting or any defense strategy.

$$R_m = \{i | \hat{y}^{\text{before}}_{m,3:0}(i) = y(i)\}, \quad N_{\text{ref}} = |R_m|. \tag{1}$$

Note that R_m differs across models because baseline accuracy under the 3:0 condition varies by model. We keep R_m fixed when computing the flip rates both *before* and *after* an intervention (CoT or defense), enabling a paired comparison on the same reference set. For any condition c, the flip rate in setting $s \in \{\text{before}, \text{after}\}$ is

$$\text{Flip Rate}^s_m(c) = \frac{\left|\left\{i \in R_m \mid \hat{y}^s_{m,c}(i) \neq y(i)\right\}\right|}{|R_m|} = \frac{N^s_{\text{flip}}(c)}{N_{\text{ref}}}. \tag{2}$$

This one-sided (correct→ incorrect) definition quantifies the risk of correct baseline judgments becoming incorrect under evidence perturbations; incorrect→ correct transitions are not counted. To measure the effect of CoT prompting or a defense strategy, we define

$$\Delta\text{Flip}_m(c) = \text{FlipRate}^{\text{after}}_m(c) - \text{FlipRate}^{\text{before}}_m(c). \tag{3}$$

A positive $\Delta\text{Flip}_m(c)$ indicates degraded robustness (more correct→ incorrect flips), whereas a negative value indicates an improvement.

Accuracy (secondary). To contextualize overall performance, we also report standard accuracy for each condition.

For each model, the same baseline-correct claims were evaluated across evidence-composition conditions within a dataset; therefore, condition-level comparisons were paired at the claim level. We use McNemar's test on binary per-claim flip indicators to assess whether paired conditions differ significantly [36] and report significance markers for these paired contrasts where applicable. McNemar's test is used only for paired comparisons on the same claim set, whereas cross-dataset comparisons and comparisons between disjoint subsets are treated as descriptive. In addition, for scalar estimates in which sampling uncertainty is of interest rather than a paired condition contrast, we report 95% bootstrap confidence intervals computed by resampling claims with replacement [37]. These estimates include the minority-refuting (3:1) and full-conflict (0:3) flip rates, and the EIJ confidence–correctness correlations.

4 | ROBUSTNESS ANALYSIS

In this section, we evaluate judgment robustness using the flip rate defined in (2), with the all-supporting 3:0 condition as the baseline. Sections 4.1–4.4 present the FEVER analysis, covering evidence ratio and quantity (Section 4.1), evidence ordering (Section 4.2), CoT prompting (Section 4.3), and lightweight mitigation strategies (Section 4.4). Section 4.5 examines whether these patterns persist in SciFact, and the complete results are presented in Appendix D.

4.1 | Evidence ratio and quantity effects

We used the ratio and quantity conditions defined in Section 3.2 with the 3:0 support-only setting as the no-conflict baseline.

22337326, 2026, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.4218/etrij.2026-0242 by Electronics and Telecommunications, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Minority refuting evidence. The key result in Table 1 is that even a small proportion of refuting evidence can induce many correct→ incorrect flips.

Under the 3:1 condition, in which refuting evidence accounted for only 25% of the evidence, the flip rates ranged from 3.3% (GPT-5.4) to 22.2% (Llama-8B), and paired McNemar's tests confirmed that the increase from the 3:0 baseline to 3:1 was significant for all six models (all $p < 0.001$). The 2:1 condition showed a similar pattern, with flip rates of up to 28.9%. Gemini-3.1-Pro (4.8%) and GPT-5.4 (3.3%) showed substantially lower 3:1 flip rates than the other four models (12.7%–22.2%), suggesting greater robustness to minority conflicts. However, this advantage diminishes under higher conflict. At 0:3, all six models exhibited high flip rates ranging from 60.8% to 79.9%, with Gemini-3.1-Pro (65.5%) and GPT-5.4 (68.8%) no longer showing a clear advantage over GPT-4o (60.8%). The full paired significance tests for adjacent ratio and quantity contrasts are presented in Appendix C.

Figure 2 shows the flip rate as a function of the proportion of refuting evidence. Across all six models, evidence ratio is a major driver of judgment instability, with flip rates increasing substantially as the refuting proportion increases.

Model comparison. Baseline accuracy alone does not explain robustness to conflicting evidence. Gemini-3.1-Pro and Llama-70B had identical 3:0 baseline accuracies (97.7%), yet their flip rates diverged substantially under conflict: at 1:2, Gemini-3.1-Pro had a flip rate of 11.2%, whereas Llama-70B had a flip rate of 26.2%. Thus, high accuracy under no-conflict evidence does not imply stable judgments under conflicting-evidence conditions.

Flip rate disparity across claim labels. When stratified by gold labels, SUPPORTS-labeled claims generally exhibited higher flip rates than REFUTES-labeled claims under high-conflict conditions, although the pattern is more nuanced under mixed conflict. The full label-stratified results are reported in Appendix C and discussed in Section 5.

Evidence quantity effects. In contrast to the strong ratio effect, varying only the total number of passages under a fixed 1:1 ratio yielded more varied behavior (Table 1, rightmost columns). For GPT-4o and the three open-source models, the flip rate generally increased from $n = 2$ to $n = 6$, with GPT-4o showing the largest increase (12.9 percentage points [pp]). In contrast, both Gemini-3.1-Pro and GPT-5.4 exhibited nonmonotonic patterns (Gemini-3.1-Pro: 11.1%→ 8.6%→ 9.3%; GPT-5.4:10.0%→ 13.5%→ 11.4%), indicating that the quantity effects were model-dependent rather than uniformly destabilizing. These results caution against assuming that simply retrieving more evidence (e.g., increasing the top-k) is always safe for RAG-style fact-verification systems.

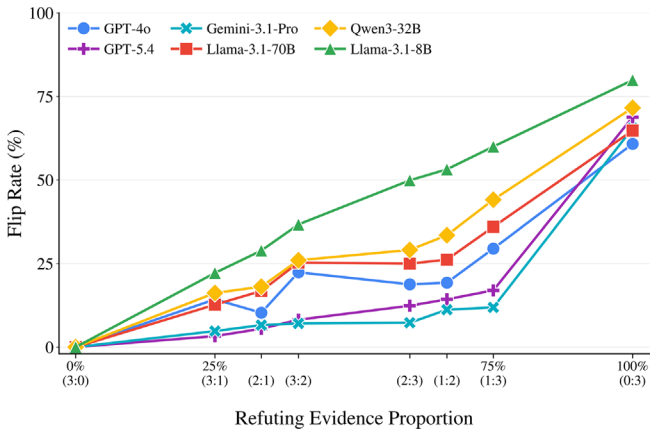

FIGURE 2 Flip rate as a function of the refuting evidence proportion. Each line shows the flip rates for one model across eight ratio conditions, from 3:0 (support-only) to 0:3 (refute-only).

TABLE 1 Flip rates across evidence-ratio and quantity conditions on FEVER. Flip rate is the percentage of claims correctly judged under the no-conflict condition (3:0) for which the judgment becomes incorrect under each condition. Each condition is denoted by $s:r$, where s and r represent the numbers of supporting and refuting evidence passages, respectively. Base Acc. denotes accuracy (%) under the all-supporting 3:0 condition. n denotes the total number of evidence passages under the fixed 1:1 ratio. Paired McNemar's tests are reported in Appendix C.

	Base	Evidence ratio				Quantity conditions (1:1)		
Model	Acc. (%)	3:1	2:1	1:2	0:3	n=2	n=4	n=6
GPT-4o	97.5	14.3	10.3	19.3	60.8	15.0	20.4	27.9
GPT-5.4	97.8	3.3	5.5	14.3	68.8	10.0	13.5	11.4
Gemini-3.1-Pro	97.7	4.8	6.6	11.2	65.5	11.1	8.6	9.3
Llama-3.1-70B	97.7	12.7	16.8	26.2	64.8	21.4	25.3	28.2
Qwen3-32B	97.8	16.2	18.1	33.6	71.6	23.6	27.2	30.5
Llama-3.1-8B	83.7	22.2	28.9	53.2	79.9	39.4	36.4	42.8

22337326, 2026, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.4218/etrij.2026-0242 by Electronics and Telecommunications, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

SciFact validation revealed the same qualitative ratio-driven pattern. Across all six models, flip rates were substantially higher under refute-majority and full-conflict conditions, with the 0:3 condition producing flip rates above 90% (Section 4.5, Table 4).

4.2 | Evidence ordering effects

Results for the four ordering conditions under the balanced 3:3 setting are summarized in Table 2.

Late-evidence weighting. Evidence ordering affected flip rates, indicating that model judgments were not invariant to the ordering of the same set of conflicting passages. Under the BASIC prompt, RF yielded lower flip rates than SF for all six models, which is consistent with a late-evidence weighting pattern: When supporting evidence appears later, judgments are more stable, whereas when refuting evidence appears later, flip rates tend to increase. This effect was most pronounced for GPT-4o, for which RF yielded 14.3% compared with 25.8% for SF, corresponding to an 11.5 pp gap attributable to ordering alone. Llama-70B exhibited a similar 8.6 pp SF–RF gap. Paired McNemar's tests confirmed that the SF–RF contrast was significant for GPT-4o ($p < 0.001$), Llama-70B ($p < 0.001$), and Gemini-3.1-Pro ($p = 0.015$). The same directional pattern appeared for GPT-5.4 ($p = 0.13$) and Qwen-32B ($p = 0.07$), although these differences were not statistically significant.

Model-specific sensitivity to ordering. The magnitude of ordering sensitivity was model-dependent. GPT-4o and Llama-70B exhibited the largest ordering ranges, whereas GPT-5.4, Gemini-3.1-Pro, and Qwen-32B exhibited the smallest ranges. Llama-8B deviated from the main late-evidence pattern: IL yielded the lowest flip rate, and its overall flip rates remained high across all orderings, suggesting that ordering optimization alone is unlikely to yield large absolute gains for smaller models. SciFact validation showed a similar qualitative pattern, with a positive SF–RF gap in five of the six models and larger ordering gaps for most models compared with FEVER (Section 4.5, Table 4).

4.3 | CoT ablation

The effects of CoT under ordering perturbations are reported in Table 2, and those under the ratio conditions are shown in Figure 3. To enable a direct comparison, both BASIC and CoT flip rates were computed against the same reference set: Instances correctly judged under the BASIC 3:0 condition, as defined in (2).

CoT effects under ordering perturbations. Across the four ordering conditions listed in Table 2, CoT does not consistently improve robustness. It increases the mean flip rate for four of the six models—significantly for Llama-70B (+2.0 pp, $p = 0.002$), Qwen-32B (+2.9 pp, $p < 0.001$), and Llama-8B (+3.3 pp, $p = 0.002$), and non-significantly for GPT-4o (+1.1 pp, $p = 0.06$)—whereas CoT improves robustness for Gemini-3.1-Pro (−1.3 pp, $p < 0.001$) and produces no significant change for GPT-5.4. Thus, CoT does not reliably mitigate sensitivity to ordering.

TABLE 2 Flip rates (%) across evidence ordering conditions under the balanced 3:3 ratio on FEVER for BASIC and CoT prompting. All conditions use an identical evidence set with only the presentation order varied. SF = `support_first`, RF = `refute_first`, IL = `interleaved`, IR = `interleaved_reverse`. Avg. denotes the mean across the four ordering conditions. Bold indicates the lowest flip rate per model per prompt type.

Model	Prompt	SF	RF	IL	IR	Avg.
GPT-4o	BASIC	25.8	**14.3**	21.9	17.2	19.8
	CoT	23.2	**16.7**	22.6	21.2	20.9
GPT-5.4	BASIC	9.6	**8.1**	10.3	10.4	9.6
	CoT	8.5	**8.2**	9.5	11.5	9.4
Gemini-3.1-Pro	BASIC	8.3	**6.3**	9.0	7.5	7.8
	CoT	7.3	**5.0**	7.7	6.0	6.5
Llama-3.1-70B	BASIC	28.7	**20.1**	28.2	26.3	25.8
	CoT	29.2	**23.4**	29.2	29.2	27.8
Qwen3-32B	BASIC	30.8	28.0	29.6	**27.6**	29.0
	CoT	**31.3**	**31.3**	31.8	33.0	31.9
Llama-3.1-8B	BASIC	44.0	43.1	**42.1**	48.3	44.4
	CoT	**46.6**	47.4	49.4	47.4	47.7

22337326, 2026, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.4218/etrij.2026-0242 by Electronics and Telecommunications, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

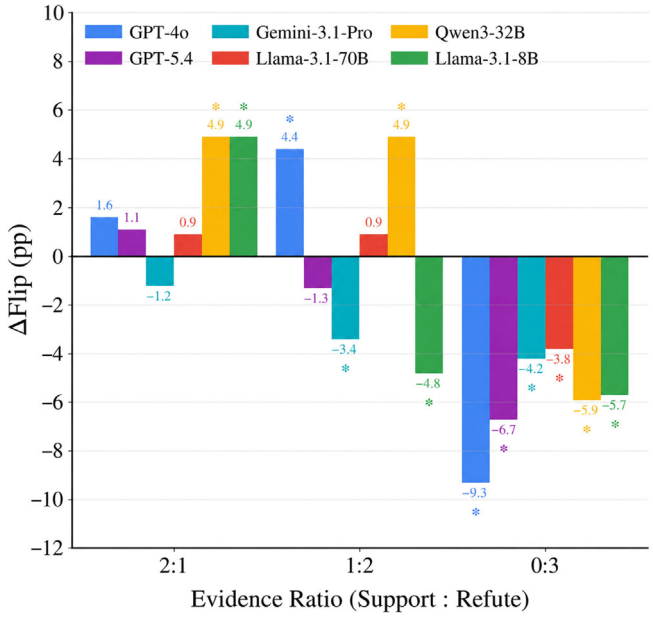

FIGURE 3 Effect of CoT on flip rate (Δ Flip = $\text{Flip}_{\text{CoT}} - \text{Flip}_{\text{Basic}}$, in percentage points [pp]). Both flip rates are computed against the same reference set (instances correctly judged under the Basic 3:0 condition). Bars above zero indicate that CoT increases the flip rate (worse), whereas bars below zero indicate that CoT decreases the flip rate (better). Three representative ratio conditions are shown (2:1, 1:2, 0:3). CoT consistently reduces the flip rate at 0:3 (full conflict) across all models, while increasing it at intermediate conflict levels for most models. An asterisk (∗) marks $p < 0.05$ (paired McNemar's test, CoT vs. Basic).

Consistent improvement under full conflict. The condition in which CoT most consistently reduced the flip rate was 0:3 (refute-only), with decreases across all six models: GPT-4o −9.3 pp, GPT-5.4 −6.7 pp, Qwen-32B −5.9 pp, Llama-8B −5.7 pp, Gemini-3.1-Pro −4.2 pp, and Llama-70B −3.8 pp (all $p < 0.01$). This suggests that when all the retrieved evidence contradicts the baseline-correct decision, step-by-step reasoning can encourage the model to re-examine whether to accept the provided evidence.

Degradation under mixed conflict. In the mixed-conflict settings, CoT increased the flip rate for most models (Figure 3). Under the 2:1 (support-majority) condition, the flip rate increased for five of the six models, significantly so for Qwen-32B ($p = 0.003$) and Llama-8B ($p = 0.02$), indicating that CoT does not improve stability even when supporting evidence outnumbers refuting evidence by a factor of two. Gemini-3.1-Pro was an exception, showing consistent improvement across most conditions (2:1: −1.2 pp; 1:2: −3.4 pp, $p < 0.001$). GPT-5.4 showed mixed effects (+1.1 pp at 2:1, −1.3 pp at 1:2; neither significant).

Summary. CoT consistently reduced the flip rate under the 0:3 condition across all six models but increased it under mixed-conflict ratio conditions for most models and under ordering perturbations for four of the six models. In FEVER, Gemini-3.1-Pro was the only model that benefited from CoT across most conditions, whereas the other models showed more mixed effects, including degradation, under some mixed-conflict settings.

In SciFact, the benefit of CoT under full conflict was only partially preserved; we discuss this domain dependence in Section 4.5.

4.4 | Mitigation experiments

The ordering study (Section 4.2) shows that even with an identical evidence set, the flip rate can differ by up to 11.5 pp purely because of the presentation order, suggesting that *presentation structure*, in addition to evidence content, plays a key role in judgment stability. Motivated by this finding, we designed and evaluated (i) two structural defenses that explicitly modify evidence presentation (EIJ and CA+SGP) and (ii) a lightweight prompt-only baseline (CA) as a point of comparison.

4.4.1 | Mitigation methods

(1) **Conflict-aware prompting (CA).** We augmented the Basic prompt with a brief warning that the evidence may contain conflicting information and instructed the model to assess the quality and credibility of each passage before answering. We included CA as a prompt-only baseline for comparison.

(2) **Evidence-level independent judgment (EIJ).** The instability observed in Sections 4.1 and 4.2 arises when multiple evidence passages are presented jointly. EIJ addresses this setting by decomposing verification into two stages: *independent judgment* and aggregation. In Stage 1, the model receives each evidence passage as a separate input and outputs: (i) a TRUE/FALSE judgment and (ii) a confidence score on a 1–5 scale. In Stage 2, we compare the mean confidence of the TRUE and FALSE sides and output the side with the higher mean as the final verdict. We use the mean rather than a sum or simple majority vote because the latter two structurally favor whichever side has more passages, meaning that under ratio-imbalanced conditions, they measure evidence-count effects rather than the defense's own contribution. In the case of a tie, we retain the baseline judgment.

(3) **Conflict-aware stance-grouped prompting (CA +SGP).** CA+SGP combines the ordering sensitivity

22337326, 2026, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.4218/etrij.2026-0242 by Electronics and Telecommunications, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

observed in Section 4.2 (in particular, the relative stability of the RF layout) with a CA warning. The process proceeds in two stages. In Stage 1, we classify the stance of each evidence passage as SUPPORT or REFUTE with respect to the claim. In Stage 2, we perform a final verification using a structured prompt that places refuting passages first and supporting passages last; we include the CA warning only when both stances are present.

All defense experiments were conducted using `temperature = 0`. We evaluated them on the full set of 774 claims under six test conditions: two ratio conditions (2:1 and 1:2) and four ordering conditions. Details of the call structure and measured token costs are presented in Appendix A.

4.4.2 | Results

Table 3 summarizes the mean baseline flip rates and defense-induced changes averaged across the six test conditions, including two ratio settings (2:1 and 1:2) and four ordering settings.

Overall effectiveness. At the model level, CA+SGP yielded the most consistent flip-rate reductions, reducing the mean flip rate for all six models, with Δ ranging from −1.0 pp for GPT-4o to −7.4 pp for Gemini-3.1-Pro. It achieved the best results for four of the six models (Gemini-3.1-Pro, GPT-5.4, Llama-70B, Qwen-32B). EIJ was effective for specific models (GPT-4o −3.2 pp, Qwen-32B −5.4 pp, and Llama-8B −4.0 pp) but worsened performance for Gemini-3.1-Pro (+1.8 pp) and GPT-5.4 (+1.1 pp), indicating that per-passage independent judgment does not universally improve robustness. This reversal suggests that EIJ's decomposition step may remove useful cross-evidence interactions in models that can otherwise exploit the jointly presented evidence. CA remains a low-cost option but has inconsistent effects. Condition-wise McNemar's tests showed that CA+SGP provided significant reductions in the majority of model-condition pairs, whereas EIJ provided significant reductions only for selected models and conditions (full condition-wise results are provided in Appendix C).

Condition-wise patterns. The condition-level results align with each defense's design. Because EIJ reduces multi-evidence verification to per-passage judgments, it removes ordering information and yields large gains in ordering tests (e.g., mean Δ of −3.9 pp for GPT-4o and −6.1 pp for Qwen-32B). CA+SGP classifies stance and enforces a fixed RF-style layout, so it is most effective where baseline ordering vulnerability is high. For Llama-70B, it yields Δ of SF, IL, and IR by −8.5, −8.6, and −7.1 pp, respectively but yields only −1.1 pp under the already stable RF. For GPT-4o, it worsens RF by +3.3 pp, suggesting added noise from stance classification. For Gemini-3.1-Pro, CA+SGP reduced the mean flip rate to 0.8%, indicating near-complete mitigation under the tested FEVER conditions.

Finally, CA+SGP sharply reduced variation across ordering conditions, shrinking GPT-4o's range from 14.3%–25.8% (gap 11.5 pp) to 17.6%–18.4% (gap 0.8 pp), with similar reductions for Llama-70B (8.6→ 1.2 pp), Qwen-32B (3.2→ 0.9 pp), Gemini-3.1-Pro (2.7→ 0.3 pp), and GPT-5.4 (2.3→ 0.5 pp), indicating that stance-based regrouping largely neutralizes sensitivity to input ordering.

Limitations and practical considerations. For the smaller model (Llama-8B), the gains from EIJ and CA+SGP were relatively limited. EIJ (−4.0 pp) was comparable to CA (−3.5 pp), and CA+SGP produced an overall change of only −1.5 pp, with a nonsignificant increase of +2.2 pp ($p = 0.15$) under the 2:1 ratio condition. This suggests that stance-classification-based defenses can be sensitive to classification errors in smaller models. By contrast, CA reduced the flip rate in all six conditions for Llama-8B (−2.8 to −4.5 pp) at no additional cost, making it the most practical option when deploying smaller models.

TABLE 3 Defense effectiveness on FEVER, averaged across six conditions (2:1, 1:2, and four ordering conditions). The BL column reports the baseline mean flip rate (%); the CA, EIJ, and CA+SGP columns report Δ, the change from baseline in percentage points (negative = improvement). Bold indicates the best defense per model. Per-condition paired McNemar's tests against the BASIC baseline are reported in Appendix C.

		Defense (Δ from baseline, pp)		
Model	**BL (%)**	**CA**	**EIJ**	**CA+SGP**
GPT-4o	18.1	+0.5	**−3.2**	−1.0
GPT-5.4	9.7	−1.5	+1.1	**−5.0**
Gemini-3.1-Pro	8.2	−6.1	+1.8	**−7.4**
Llama-3.1-70B	24.4	−0.7	−1.7	**−5.5**
Qwen3-32B	27.9	−1.2	−5.4	**−6.5**
Llama-3.1-8B	43.3	−3.5	**−4.0**	−1.5

4.5 | Cross-domain validation

To assess whether the diagnostic findings hold beyond FEVER, we applied the same experimental protocol to 610 SciFact claims, using the same six models. Table 4 summarizes the key SciFact validation results, and the

22337326, 2026, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.4218/etrij.2026-0242 by Electronics and Telecommunications, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 Summary of the SciFact cross-domain validation. The 3:1, 1:2, and 0:3 columns report Basic flip rates (%). SF–RF denotes the SF minus RF flip-rate gap (pp) under Basic; CoT Δ (0:3) is the CoT minus Basic flip rate at 0:3; CA+SGP Δ(avg) is the mean change from baseline across the six defense conditions (pp; negative = improvement). Asterisks on SF–RF and CoT indicate paired McNemar's significance ($^{*}p < 0.05$). Baseline accuracies and condition-wise significance tests are reported in Appendix D.

	Flip rate (%)					
Model	3:1	1:2	0:3	SF–RF gap	CoT Δ(0:3)	CA+SGP Δ(avg)
GPT-4o	31.6	67.3	92.1	24.3*	−20.4*	−8.2
GPT-5.4	10.4	25.5	93.4	6.3*	−5.9*	−11.1
Gemini-3.1-Pro	8.9	23.6	90.5	5.0*	−1.4	−16.6
Llama-3.1-70B	27.5	54.0	94.9	32.5*	−2.3*	−15.4
Qwen3-32B	25.6	55.5	93.9	14.5*	−2.5*	−19.4
Llama-3.1-8B	44.4	77.5	92.2	−1.6	+0.8	−9.8

complete per-condition results are presented in Appendix D.

The main diagnostic patterns largely hold for SciFact. All six models show clear ratio-driven vulnerability, with full-conflict flip rates exceeding 90%—higher than the corresponding FEVER range (60.8%–79.9%). Because SciFact differs from FEVER in domain, claim style, and baseline difficulty, we focus on whether the qualitative patterns persist rather than comparing absolute magnitudes. The late-evidence weighting pattern also mostly holds: support_first yields higher flip rates than refute_first in five of the six models, with larger ordering gaps for most models than in FEVER (e.g., Llama-70B: 32.5 pp vs. 8.6 pp).

The effects of CoT and EIJ are more domain-dependent. First, CoT significantly reduces the 0:3 flip rate for four of the six SciFact models, compared with all six on FEVER; Llama-8B and Gemini-3.1-Pro show no significant reduction. Second, the effects of EIJ are strongly model- and domain-dependent: Appendix D shows that EIJ improves robustness for some models on SciFact, including Llama-70B and Llama-8B, but increases the mean flip rates for GPT-5.4 (+12.4 pp) and Gemini-3.1-Pro (+6.1 pp). In contrast, CA+SGP provides the most consistent mitigation across the two datasets, reducing the mean flip rate for all six models on SciFact. These results support the conclusion that evidence-composition vulnerability is not specific to FEVER, while the effectiveness of reasoning- or decomposition-based interventions depends on the model and domain.

5 | DISCUSSION

Flip rate and deployment implications. The flip rate metric captures a form of instability that standard accuracy does not reflect; an initially correct judgment can become incorrect when the evidence composition changes. In binary fact-verification settings, such reversals directly change the final TRUE/FALSE verdict and result in deployment-relevant failures. Gemini-3.1-Pro and GPT-5.4 show greater robustness at low conflict levels but remain vulnerable under full conflict, with 0:3 flip rates exceeding 65%. This suggests that high baseline accuracy alone provides limited assurance of stability when the retrieved evidence contains conflicting signals.

Flip risk across claim labels. The label-stratified results show that the flip risk is not symmetric across gold labels. Under full conflict (0:3), SUPPORTS claims flipped at higher rates than REFUTES claims for all six models, with gaps ranging from 5.0 pp (Llama-8B) to 51.3 pp (GPT-4o). Under mixed conflict, the pattern is more model-dependent: at 1:2, the SUPPORTS–REFUTES gap is positive for five of the six models, whereas Gemini-3.1-Pro shows the opposite pattern. These results suggest that for most models in our evaluation, correct SUPPORTS judgments were more vulnerable to opposing evidence than were correct REFUTES judgments. Possible contributors include differences in how supporting and refuting evidence are expressed as well as model-level sensitivity to negation or contradiction. Isolating these causes requires more controlled experiments. Therefore, we consider label asymmetry an important direction for future research.

Reasoning-based versus structure-based mitigation. The CoT and mitigation results highlight the practical distinction between reasoning- and structure-based interventions. CoT improves robustness under full conflict for all models on FEVER but can degrade it under mixed conflict for most models, suggesting that explicit reasoning does not simply add robustness and may change how models react to detected conflict. EIJ, which decomposes the input into single-passage judgments, is also model-dependent; it reduces flip rates for four models but increases them for Gemini-3.1-Pro and GPT-

22337326, 2026, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.4218/etrij.2026-0242 by Electronics and Telecommunications, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

5.4, suggesting that per-passage decomposition may remove useful cross-evidence interactions in some models. In contrast, CA+SGP provides the most consistent reductions in mean flip rate across models and datasets by reorganizing evidence by stance before prompting. These findings suggest that restructuring evidence presentation was more consistently effective in our experiments than simply eliciting additional reasoning, although no single intervention fully resolved the vulnerability across all models and conditions.

Limitations and future directions. This study had several limitations. First, our evidence construction used GPT-4o for counter-evidence generation and Qwen3-32B for automatic quality screening, both of which were also included among the evaluated models. This may have introduced self-preference or scorer-specific effects. We mitigated this concern by separating generation and screening across model families, providing human validation of evidence quality, and showing that models not involved in evidence construction exhibit the same core ratio-driven vulnerability; however, the automatic score remains a proxy rather than a calibrated human-equivalent measure. The evidence-quality gaps and quality-gate selection effects are further analyzed in Appendices B and C. Second, although SciFact extends the evaluation beyond FEVER, two datasets still cover a limited range of domains and claim styles. Our experiments also used at most six evidence passages, a controlled small-k setting designed to isolate the ratio and ordering effects, rather than simulating a large-k retrieval deployment. Therefore, we do not establish how flip risk scales to the larger top-k values used in deployed RAG systems. Third, the main experiments used a temperature of 0 for reproducibility; a check at a temperature of 0.7 in Appendix C suggests that composition effects exceed sampling variability, but broader decoding settings remain untested. Addressing these limitations requires evidence construction with generator and scorer models independent of the evaluated systems, larger top-k values over a wider range of domains and models, and robustness checks under diverse decoding settings.

6 | CONCLUSION

This study examined LLM-based fact verification under conflicting evidence using a diagnostic stress test. Instead of evaluating only whether a model was accurate under clean evidence, we measured whether judgments that were initially correct remained stable when the composition, quantity, and order of the supporting and refuting evidence were systematically varied. Across the six models and two datasets, this flip-rate protocol exposed reliability risks that were not visible from standard accuracy alone.

The results showed that judgment instability is not limited to extreme or artificially balanced conflict settings. Even a small proportion of refuting evidence can overturn correct decisions, full-conflict conditions produce severe fragility, and the same evidence set can lead to different outcomes depending on the presentation order. The relative severity of these vulnerabilities varies across models, indicating that conflict robustness should be treated as a separate evaluation axis rather than inferred from baseline performance. We also find that flip risk is label-asymmetric, with SUPPORTS claims generally more vulnerable than REFUTES claims.

The intervention results further suggest that robustness cannot be guaranteed by eliciting additional reasoning. CoT prompting helps in some full-conflict cases but can be unreliable under mixed conflict, whereas EIJ remains model-dependent, and CA+SGP provides the most consistent reductions in mean flip rate across the tested settings. However, no intervention fully removes the vulnerability, and the effects of reasoning- and decomposition-based methods remain model- and domain-dependent.

Overall, our findings suggest that reliable fact-verification systems should be evaluated not only for accuracy under clean evidence but also for stability under realistic conflicting-evidence conditions. Future work should extend this diagnostic framework to realistic retrieval settings with larger evidence sets, additional domains and model families, and adaptive interventions such as conflict-triggered CoT or stance-aware evidence organization.

ORCID

Jihyeon Roh https://orcid.org/0000-0002-1708-7259
Kyoungman Bae https://orcid.org/0000-0001-9007-4027

REFERENCES

1. A. Dmonte, R. R. Oruche, M. Zampieri, P. Calyam, and I. Augenstein, *Claim verification in the age of large language models: A survey*, (Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics, Volume 4: Student Research Workshop, San Diego, CA, USA), 2026, pp. 15–29.
2. D. Zeng, Z. Zuo, L. Yang, X. Xiao, and Z. Tang, *Text adversarial attacks using policy gradients against deep learning classifiers*, ETRI J. **47** (2025), no. 6, 1085–1103.
3. H. He, Y. Li, B. B. Zhu, D. Wen, R. Cheng, and F. C. M. Lau, *Fact2Fiction: Targeted poisoning attack to agentic fact-checking system*, (Proceedings of the AAAI Conference on Artificial Intelligence, Singapore), 2026, pp. 30943–30950.
4. Y. Wu, X. Liu, Y. Li, Y. Gao, Y. Ding, J. Ding, X. Zheng, and X. Ma, *ADMIT: Few-shot knowledge poisoning attacks on RAG-*

22337326, 2026, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.4218/etrij.2026-0242 by Electronics and Telecommunications, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

based fact checking, arXiv preprint (2025), DOI 10.48550/arXiv:2510.13842.

5. J. Roh, M. Kim, and K. Bae, *Towards a small language model powered chain-of-reasoning for open-domain question answering*, ETRI J. **46** (2024), no. 1, 11–21.
6. Z. Xie, R. Xing, Y. Wang, J. Geng, H. Iqbal, D. Sahnan, I. Gurevych, and P. Nakov, *FIRE: fact-checking with iterative retrieval and verification*, (Findings of the Association for Computational Linguistics: Naacl 2025, Albuquerque, NM, USA), 2025, pp. 2901–2914.
7. S. Longpre, K. Perisetla, A. Chen, N. Ramesh, C. DuBois, and S. Singh, *Entity-based knowledge conflicts in question answering*, (Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic), 2021, pp. 7052–7063.
8. H. Wang, A. Prasad, E. Stengel-Eskin, and M. Bansal, *Retrieval-augmented generation with conflicting evidence*, (Second Conference on Language Modeling (COLM), Montreal, Canada), 2025.
9. J. Xie, K. Zhang, J. Chen, R. Lou, and Y. Su, *Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts*, (The Twelfth International Conference on Learning Representations, Vienna, Austria), 2024.
10. Z. Ge, Y. Wu, D. W. K. Chin, R. K.-W. Lee, and R. Cao, *Resolving conflicting evidence in automated fact-checking: a study on retrieval-augmented LLMs*, (Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, Montreal, Canada), 2025, pp. 9656–9664.
11. Z. Guo, M. Schlichtkrull, and A. Vlachos, *A survey on automated fact-checking*, Trans. Assoc. Comput. Linguist. **10** (2022), 178–206.
12. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, *Retrieval-augmented generation for knowledge-intensive NLP tasks*, Adv. Neural Inf. Process. Syst. **33** (2020), 9459–9474.
13. F. Cuconasu, G. Trappolini, F. Siciliano, S. Filice, C. Campagnano, Y. Maarek, N. Tonellotto, and F. Silvestri, *The power of noise: Redefining retrieval for RAG systems*, (Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, Washington DC, USA), 2024, pp. 719–729.
14. F. Shi, X. Chen, K. Misra, N. Scales, D. Dohan, E. H. Chi, N. Schärli, and D. Zhou, *Large language models can be easily distracted by irrelevant context*, (Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA), 2023, pp. 31210–31227.
15. A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, *Self-RAG: Learning to retrieve, generate, and critique through self-reflection*, (The twelfth International Conference on Learning Representations, Vienna, Austria), 2024.
16. Y. Pan, L. Pan, W. Chen, P. Nakov, M.-Y. Kan, and W. Wang, *On the risk of misinformation pollution with large language models*, (Findings of the Association for Computational Linguistics: Emnlp 2023, Singapore), 2023, pp. 1389–1403.
17. M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. Bowman, E. Durmus, Z. Hatfield-Dodds, S. Johnston, S. Kravec, and T. Maxwell, *Towards understanding sycophancy in language models*, (The Twelfth International Conference on Learning Representations, Vienna, Austria), 2024.
18. M. Wu and A. F. Aji, *Style over substance: Evaluation biases for large language models*, (Proceedings of the 31st International Conference on Computational Linguistics, Abu Dhabi, United Arab Emirates), 2025, pp. 297–312.
19. R. Xu, Z. Qi, Z. Guo, C. Wang, H. Wang, Y. Zhang, and W. Xu, *Knowledge conflicts for LLMs: A survey*, (Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA), 2024, pp. 8541–8565.
20. T. Schuster, A. Fisch, and R. Barzilay, *Get your vitamin C! robust fact verification with contrastive evidence*, (Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies), 2021, pp. 624–643.
21. W. Zhou, S. Zhang, H. Poon, and M. Chen, *Context-faithful prompting for large language models*, (Findings of the Association for Computational Linguistics: Emnlp 2023, Singapore), 2023, pp. 14544–14556.
22. J. Chen, B. Bi, W. Zhang, J. Sui, X. Zhu, Y. Wang, L. Mei, and S. Liu, *Rethinking all evidence: Enhancing trustworthy retrieval-augmented generation via conflict-driven summarization*, arXiv preprint (2025), DOI 10.48550/arXiv:2507.01281.
23. E. Choi, J. Park, H. Lee, and J. Lee, *Conflict-aware soft prompting for retrieval-augmented generation*, (Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou, China), 2025, pp. 26981–26995.
24. J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou, *Chain-of-thought prompting elicits reasoning in large language models*, Advances in Neural Information Processing Systems **35** (2022), 24824–24837.
25. N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, *Lost in the middle: How language models use long contexts*, Trans. Assoc. Comput. Linguist. **12** (2024), 157–173.
26. R. Tang, C. Zhang, X. Ma, J. Lin, and F. Türe, *Found in the middle: Permutation self-consistency improves listwise ranking in large language models*, (Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Mexico City, Mexico), 2024, pp. 2327–2340.
27. X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou, *Self-consistency improves chain of thought reasoning in language models*, (The Eleventh International Conference on Learning Representations, Kigali, Rwanda), 2023.
28. J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, *FEVER: a large-scale dataset for fact extraction and verification*, (Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, LA, USA), 2018, pp. 809–819.
29. A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, and C. Zheng, *Qwen3 technical report*, arXiv preprint (2025), DOI 10.48550/arXiv:2505.09388.
30. D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, and H. Hajishirzi, *Fact or fiction: Verifying scientific claims*, (Proceedings of the 2020 Conference on Empirical

22337326, 2026, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.4218/etrij.2026-0242 by Electronics and Telecommunications, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Methods in Natural Language Processing (EMNLP)), 2020, pp. 7534–7550.
31. A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. J. Ostrow, A. Welihinda, A. Hayes, A. Radford, and A. Mądry, *GPT-4o system card*, arXiv preprint (2024), DOI 10.48550/arXiv:2410.21276.
32. OpenAI, *Introducing GPT-5.4*, 2026. [Accessed: 2026. 6. 4]
33. Google DeepMind, *Gemini 3.1 Pro model card*, 2026. [Accessed: 2026. 6. 4]
34. A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, and A. Yang, *The Llama 3 herd of models*, arXiv preprint (2024), DOI 10.48550/arXiv:2407.21783.
35. W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, and I. Stoica, *Efficient memory management for large language model serving with pagedattention*, (Proceedings of the 29th Symposium on Operating Systems Principles, Koblenz, Germany), 2023, pp. 611–626.
36. Q. McNemar, *Note on the sampling error of the difference between correlated proportions or percentages*, Psychometrika **12** (1947), no. 2, 153–157.
37. B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*, Chapman and Hall/CRC, 1994.

AUTHOR BIOGRAPHIES



Jihyeon Roh received her Ph.D. in electrical engineering from the Korea Advanced Institute of Science and Technology in 2022. Since 2022, she has worked at the Electronics and Telecommunications Research Institute, where she is currently a member of the Vision Intelligence Research Section. Her research interests include large language models, natural language processing, trustworthy AI, and the reliability and robustness of language models.



Kyoungman Bae received his B.S., M.S., and Ph.D. degrees in computer engineering from Dong-A University, Busan, Republic of Korea, in 2004, 2006, and 2016, respectively. Since 2016, he has worked at the Electronics and Telecommunications Research Institute, where he is currently a member of the Vision Intelligence Research Section. His primary research interests include large language models, natural language processing, explainable AI, and generative AI.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: J. Roh and K. Bae, *Diagnosing judgment flip risk of large language models in fact verification under conflicting evidence*, ETRI Journal **48** (2026), 838–851, DOI 10.4218/etrij.2026-0242

22337326, 2026, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.4218/etrij.2026-0242 by Electronics and Telecommunications, Wiley Online Library on [29/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License