Large Language Models in Resolving Contextual Knowledge Conflicts

Xinye Yang[1] **Zhenyang Liu**[1] **Ruisi Li**[2] **Yuanyuan Lei**[3]

[1]Northwestern University, Evanston, IL, [2]New York University, New York, NY
[3]Computer & Information Science and Engineering, University of Florida, Gainesville, FL
XinyeYang2027@u.northwestern.edu, yuanyuan.lei@ufl.edu

Abstract

Most prior works focused on conflicts between an LLM's internal parametric knowledge and externally provided context. In contrast, we investigate how LLMs handle conflicts that arise within contextual knowledge itself. We introduce a taxonomy of six types of contextual conflicts (*misinformation, inferential, temporal, granularity, perspective, and ambiguity*) and contribute a comprehensive dataset CONTEXTCONFLICT for this setting. The dataset contains 5,781 samples, covers both reasoning and summarization tasks, and includes both explicit contradictions and implicit conflicts that require multi-step reasoning. Experiments on seven LLMs show that current models still fall short in resolving contextual knowledge conflicts. We further provide mechanistic interpretability insights into how LLMs process such conflicts, revealing their latent awareness of conflicts and the representational geometry underlying conflict processing. In addition, our analysis uncovers a consistent model bias towards earlier evidence, and this positional preference serves as a key obstacle to effective conflict resolution. Motivated by these findings, we further propose a simple training-free, label-free steering method that steers activations to encourage a more comprehensive incorporation of evidences for better conflict resolution. On our dataset, the method consistently improves accuracy on reasoning tasks and generates higher-quality, more balanced summaries for summarization tasks. [1][2]

1 Introduction

Large language models (LLMs) deployed in retrieval-augmented or multi-document settings routinely encounter conflicting information from multiple sources. Figure 1 illustrates a representative case: when asked about social media identity-verification policies, an LLM must synthesize opposing viewpoints that prioritize different values (safety versus privacy). The reliability of downstream applications therefore depends on how well models handle such conflicts. When conflict resolution fails, outputs may become hallucinated, factually incorrect, or one-sided (Shi et al., 2024; Chen et al., 2022), with the failure mode extending to news aggregation, medical decision support, misinformation detection, and legal reasoning in high-stakes settings. Characterizing how LLMs process multi-source conflicts is therefore a prerequisite for diagnosing and correcting these failures (Longpre et al., 2021; Chen et al., 2022).

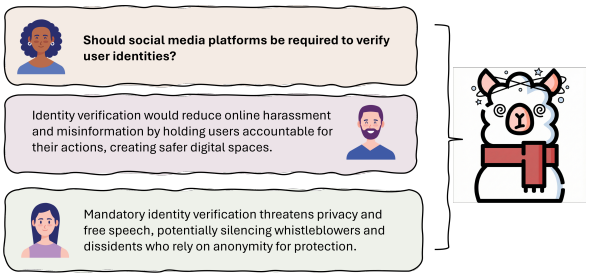

Figure 1: Example of contextual knowledge conflict: two contexts provide divergent perspectives.

Prior datasets on contextual knowledge conflicts exhibit four interrelated limitations. First, they rely on template-based synthetic construction, most commonly entity replacement, which fails to capture real-world complexity (Su et al., 2024; Longpre et al., 2021). Second, they focus on explicit factual contradictions and overlook implicit conflicts that require multi-step reasoning (Su et al., 2024; Lazaridou et al., 2021; Du et al., 2022). Third, they offer limited domain breadth and conflict-type coverage. Fourth, they can suffer from class imbalance, which makes category-level analysis statistically unreliable (Xu et al., 2024; Xie et al., 2024).

[1]The link for dataset and code is: https://github.com/lei-nlp-lab/context_conflict_emnlp_2026.

[2]CONTEXTCONFLICT dataset is also released on Hugging Face: https://huggingface.co/datasets/AsherYang/ContextConflict

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 21142–21169
October 24-29, 2026 ©2026 Association for Computational Linguistics

To address the above gaps, we contribute CONTEXTCONFLICT, a comprehensive multi-domain dataset of contextual knowledge conflicts comprising 5,781 samples. We define a taxonomy of six conflict types, including *misinformation*, *inferential*, *temporal*, *granularity*, *perspective*, *ambiguity* conflicts. The first three correspond to ***reasoning*** tasks, where LLMs are required to perform reasoning and identify the correct answer from multiple conflicting evidences. The latter three correspond to ***summarization*** tasks, where LLMs are expected to produce balanced summaries that preserve divergent perspectives. Within each conflict type, we further distinguish *explicit* conflicts that are identifiable by direct surface-level comparison, and *implicit* conflicts that require multi-step cross-evidence reasoning to resolve. An evaluation of seven LLMs covering both closed-source and open-source models on this dataset show that modern LLMs still fall short in resolving contextual knowledge conflicts.

We also provide mechanistic interpretability insights into how LLMs process contextual conflicts. Specifically, we investigate the latent awareness and representational geometry of each conflict type using concept activation vectors (Kim et al., 2018) and spectral energy decomposition (Eckart and Young, 1936), respectively. We also examine evidence attributions at both representation and output levels: at the representation level, we compare the activation geometry induced by single evidence versus combined evidence; at the output level, we quantify evidence contributions using Shapley-based attribution scores (Lundberg and Lee, 2017). Our analysis reveals strong awareness of contextual conflicts internalized in models, with different conflict types emerging at different layer depths and represented as distinct geometry in model's latent space. We also uncover a consistent model bias favoring earlier-positioned evidences, suggesting that this positional preference is a key obstacle to comprehensive evidence integration.

To address this issue, we then design a training-free, label-free steering method that mitigates positional preference and encourages more comprehensive consideration of conflicting evidences. Specifically, we construct a steering direction by computing the centroid of LLMs activations when processing each piece of evidence individually. During inference, we nudge the model's activations along this steering direction, guiding it to attend more evenly across evidence positions and thereby integrate evidence more comprehensively. The results show that our simple method effectively improves LLMs' ability to resolve knowledge conflicts, yielding more balanced summaries in summarization tasks and higher accuracy on reasoning tasks.

Our main contributions are summarized below.

- We contribute a comprehensive multi-domain dataset covering six types of contextual knowledge conflicts, spanning reasoning and summarization tasks across diverse domains
- We present a mechanistic interpretability analysis of contextual conflict processing, revealing how each conflict are detected, geometrically represented, and processed within LLMs
- We design a training-free, label-free steering method that encourages more comprehensive integration of conflicting evidences, yielding better performance in conflict resolution

2 CONTEXTCONFLICT: Contextual Knowledge Conflict Dataset

2.1 Overview

We contribute a multi-domain contextual knowledge conflict dataset to address the limitations of existing datasets. We define a taxonomy of six conflict types and each conflict type probes a distinct capability, including *inferential* reasoning (multi-step reasoning over conflicting premises), *misinformation* robustness (resistance to plausible disinformation), *temporal* reasoning (tracking claim validity over time), *granularity* alignment (reconciling information at different abstraction levels), *perspective* integration (balancing diverse viewpoints), and *ambiguity* resolution (disambiguating co-referential entities across documents). These six conflict types can be organized into two task families: **Reasoning** tasks (*inferential*, *misinformation*, *temporal*) require models to identify the correct answer under conflicting evidence, and **Summarization** tasks (*ambiguity*, *perspective*, *granularity*) require models to produce balanced summaries that preserve divergent viewpoints. Within each family, conflicts are further divided into **explicit** and **implicit** cases. Explicit cases are detectable by direct surface-level comparison and require at most one inferential step. Implicit cases emerge through multi-step cross-evidence reasoning and require at least two inferential steps. We provide the per-type implicit proportions in Table 1.

Type	Size	Avg.#Ev.	Implicit (%)	Base source(s)
Misinformation	1004	2.66	37.1	SciFact; ConflictBank
Inferential	787	5.66	25.4	ENTAILMENTBANK; NEJM-MedQA; FOLIO
Temporal	960	2.26	93.5	ConflictBank; CONFLICTS
Granularity	1020	5.36	7.4	ROAST-ABSA; NEJM-MedQA
Perspective	1010	4.48	49.5	AllSides; Perspectrum
Ambiguity	1000	2.94	0.0	AmbigDocs
Total	**5781**			

Table 1: CONTEXTCONFLICT statistics: size, average number of evidence pieces, implicit conflict proportion, and base sources for each conflict type.

The desired model output differs by task family. In Reasoning tasks, each instance has a verifiable ground-truth answer, so the model is expected to identify the correct conclusion under the conflicting evidence, and Accuracy serve as evaluation metric. In Summarization tasks, each instance admit multiple valid responses: *perspective* and *ambiguity* often involve diverging political opinions or social questions, while *granularity* allows compatible answers at different specificity levels. The model is expected to comprehensively integrate the divergent viewpoints and generate a balanced summary: for *perspective* conflicts, it should present and contrast all stances without privileging any; for *ambiguity* conflicts, it should identify the underlying name collision and cover each referenced entity; for *granularity* conflicts, it should integrate the different levels of specificity and state their compatibility. The Shapley-based Balance score is employed as evaluation metric to measure evenness of evidence use. More evaluation details are in 3.1.

2.2 Dataset Composition

We construct the dataset from ten base datasets spanning diverse domains; Appendix Table 5 lists each source with its domain, license, and per-conflict-type usage. The dataset consists of 1,734 semi-synthetic instances (29.99%) and 4,047 preserved-original instances manually categorized by conflict type (70.01%), for a total of 5,781.

Unlike prior synthetic conflict datasets that build conflicts via entity replacement (Su et al., 2024; Longpre et al., 2021), which keeps the surrounding context unchanged, swaps only the entity, and produces shallow lexical contradictions, our semi-synthetic instances use GPT-5 (Singh et al., 2025) only as an auxiliary generator for conflicting evidence or timestamps; gold labels are inherited from the source dataset or human re-annotated, never produced by GPT-5. This preserves discourse-level coherence in the conflicting evidence while keeping label integrity independent of the generator.

Quality assurance is risk-proportional, scaling with how much each construction step can corrupt labels. Subsets whose perturbations may flip the gold label receive full re-annotation by two independent annotators; subsets whose augmentations rarely change the label are audited on a 30% sample; preserved-original subsets are spot-checked at 10%. Per-source provenance counts are released in the metadata; Appendix B.1 reports the full GPT-5 prompts and validation rules, and Appendix B.3 provides representative samples.

Granularity and Inferential Conflicts. NEJM-MedQA (Savage et al., 2024) combines U.S. medical-licensing exam questions with real clinical cases from the *New England Journal of Medicine*. Each instance provides clinical evidence (symptoms, lab results, patient history) and several candidate reasoning chains. Each chain is generated under a distinct diagnostic-reasoning prompt strategy from the source dataset and carries a binary gold-correctness flag adjudicated by clinicians. We select cases where two chains disagree on the final diagnosis. When both diagnoses are gold-correct but at different abstraction levels (e.g., a broad syndrome label versus a specific pathogen it subsumes), we label a *granularity* conflict: the two answers are compatible and differ only in diagnostic specificity. When only one diagnosis is gold-correct and the other reaches a wrong conclusion through a flawed intermediate step (e.g., a misapplied clinical heuristic), we label an *inferential* conflict: the disagreement is at the reasoning-process level. The original question and reasoning chains are preserved as evidence.

Inferential Conflicts. FOLIO (Han et al., 2024) and ENTAILMENTBANK (Dalvi et al., 2021) are logical inference datasets. Each instance gives a set of premises and a hypothesis, labeled by whether the premises entail, contradict, or are neutral to the hypothesis. We add one or two extra statements that perturb the original reasoning chain. Because the added evidence can change label validity, every resulting instance is re-annotated by two independent annotators, with disagreements resolved by discussion (100% double-annotation coverage; annotation interface in Appendix Figure 8).

Misinformation Conflicts. SciFact (Wadden et al., 2020) is a fact-verification dataset. Each instance pairs a claim with evidence sentences from research abstracts, labeled as supporting or refuting the claim. We generate conflicting evidence in a matching writing style, including experimental-style citations, to build *misinformation* conflicts. Because this augmentation rarely changes gold labels, we audit a 30% sample for stylistic and argumentative consistency.

Temporal Conflicts. ConflictBank-temporal (Su et al., 2024) instance contains evidence statements with an implicit chronological order and a time-sensitive question. We attach an explicit timestamp to each statement, without modifying the original evidence, so that temporal relationships become unambiguous. Because timestamps can change gold labels, each instance is verified by two independent annotators with discussion-based adjudication.

Curation of Remaining Datasets. For datasets used without synthetic generation (AmbigDocs, ROAST-ABSA, AllSides, Perspectrum, CONFLICTS), we randomly sample 10% of each source for quality review and conflict-category validation.

3 Evaluation

3.1 Evaluation Metrics

We evaluate model performance using complementary metrics with task-specific emphasis.

Accuracy. For **reasoning tasks** (*inferential*, *misinformation*, and *temporal* conflicts), we measure accuracy as the proportion of model outputs that match the gold-standard labels.

Evidence Balance. For **summarization tasks**, all evidence sources are equally valid despite their conflicting content, so we quantify how evenly a response integrates them with a Shapley-based attribution framework. We assign each evidence piece a Shapley-value contribution to the response's likelihood and then measure how unequally these contributions are distributed. Given n evidence pieces, index set $N = \{1, \ldots, n\}$, and model response R, the marginal contribution of piece i is

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{1}{n\binom{n-1}{|S|}} \big(v(S \cup \{i\}) - v(S)\big), \quad (1)$$

where $v(S) = \frac{1}{|R|} \sum_{t=1}^{|R|} \log p_{\text{scorer}}(r_t \mid r_{<t}, \mathcal{E}_S)$ is the length-normalized log-likelihood of R under a frozen external scorer (default: Llama-3.2-1B; scorer-size sensitivity is examined in Section 3.2). We clip negative contributions, which arise when the response contradicts a piece, and normalize the non-negative mass into a share distribution $p_i = \max(0, \phi_i) / \sum_j \max(0, \phi_j)$. We then report Balance as the normalized Gini coefficient of $\mathbf{p}$:

$$\text{Balance}(R) = \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^{n} (2i - n - 1)\, p_{[i]}, \quad (2)$$

where $p_{[1]} \leq \cdots \leq p_{[n]}$ are sorted contributions. Lower scores indicate more balanced integration (0 = perfect equality; 1 = maximum inequality).

Faithfulness.[3] To detect hallucinated content that goes beyond the given evidence in conflict scenarios, we report Faithfulness, applicable to all conflict types. It decomposes a response into atomic claims and reports the proportion judged supported by the given evidence. Faithfulness measures grounding rather than factual correctness: a factually accurate response still scores low if some claims are not grounded in the provided evidence.

3.2 Validating the Balance Metric

Unlike established metrics such as Accuracy and Faithfulness, Evidence Balance relies on our Shapley-based attribution framework. We confirm its reliability with three independent checks: a human and LLM agreement study, a causal intervention on the same human-verified subset, and a scorer-size robustness check.

Human Evaluation and LLM-as-a-Judge. To assess agreement between our automatic metric and human judgment of evidence balance, we collect annotations from two independent humans and a GPT-5 LLM-as-a-Judge on 108 samples spanning three models and three summarization tasks. Our metric reaches $\kappa = 0.50$–0.52 against the human annotators (Table 2), comparable to inter-human agreement ($\kappa = 0.52$) and to the LLM-as-a-Judge ($\kappa = 0.51$). Annotation setup, interface, and per-model Balance appear in Appendix C.2.

Causal Intervention. To probe whether Shapley rankings reflect actual evidence dependence, we run a causal-intervention check on the human-annotated subset above, restricted to the two open-weight models where each model's own log-probabilities are accessible (Llama-3.1-8B-Instruct and GPT-OSS-20B), giving 72 already human-verified samples. We remove the highest- and

[3] Computed using RAGAS (Es et al., 2024): https://github.com/vibrantlabsai/ragas

Annotator Pair	Cohen's Kappa
Human ↔ Human	
Human 1 vs Human 2	0.52
Human ↔ Ours	
Human 1 vs Ours	0.52
Human 2 vs Ours	0.50
Human ↔ LLM-as-a-Judge	
Human 1 vs LLM-as-a-Judge	0.55
Human 2 vs LLM-as-a-Judge	0.56
Ours ↔ LLM-as-a-Judge	
LLM-as-a-Judge vs Ours	0.51

Table 2: Inter-annotator agreement (Cohen's Kappa).

Model	ℓ_{full}	ℓ_{high}	ℓ_{low}	Δ_{high}	Δ_{low}
Llama-3.1-8B	-0.656	-0.977	-0.732	0.321	0.076
GPT-OSS-20B	-6.557	-7.442	-6.930	0.885	0.373

Table 3: Causal intervention on the 72 human-verified samples. $\Delta_{\text{high}}=\ell_{\text{full}}-\ell_{\text{high}}$, $\Delta_{\text{low}}=\ell_{\text{full}}-\ell_{\text{low}}$, using each model's own log-likelihood.

lowest-contributing evidence piece and measure the change in each model's own log-likelihood. As Table 3 shows, removing the highest-contributing evidence causes substantially larger drops than removing the lowest, indicating that attribution rankings align with actual evidence dependence.

Scorer-Size Robustness. To rule out artifacts of the 1B default scorer, we rerun the Shapley computation with two alternative scorers on the three summarization tasks: one larger from the same family (Llama-3.1-8B) and one comparable in size from a different family (Gemma-2B). As shown in Appendix Table 6, absolute Balance values shift but the ranking of evaluated models is essentially unchanged, confirming that our results do not depend on the default scorer.

3.3 Evaluation Results

Task Difficulty and Model Scaling. Table 4 reports performance across seven base models and six conflict categories. On reasoning tasks, we observe a difficulty hierarchy: Temporal reaches the highest accuracy (up to 68.0%), followed by Misinformation (up to 61.5%) and Inferential (up to 44.3%), indicating increasing difficulty. Performance scales with model capacity (Inferential: 24.8% on llama-3.1-8b-instruct vs. 44.3% on claude-4.5-sonnet), yet Inferential stays below 50% even for the strongest model, suggesting multi-step reasoning under conflict is a shared SOTA bottleneck.

Pervasive Position Bias. Our Shapley-based attribution reveals strong evidence-position bias in summarization tasks. As shown in Appendix Figure 23 for Llama-3.1-8B-Instruct, Ambiguity shows the strongest bias, with the first evidence contributing 69.0%, while Perspective and Granularity are less extreme but still imbalanced. Even top-performing models remain far from perfectly balanced integration. Detailed per-model evidence distributions are reported in Appendix D.5. Balance is also not monotonic in capacity (GPT-OSS-120B 31.9 vs. 20B 39.8), suggesting that scale alone cannot eliminate position bias and motivating the correction in Section 5.

Faithfulness Patterns. Faithfulness is higher on summarization than on reasoning and is not tightly coupled with Balance (Table 4): GPT-OSS-20B on Perspective reaches 92.5% Faithfulness while its Balance stays at 34.5, showing that a model can be simultaneously well-grounded in the provided evidence and positionally biased, so the bias operates primarily at the evidence-selection stage.

4 Analysis

We conduct three complementary analyses to understand how LLMs internally process conflicting knowledge. (i) **Conflict awareness** measures the initial detection point via the linear separability of hidden states. (ii) **Representational geometry** examines structural segregation via spectral energy analysis. (iii) **Position bias** traces how evidence order skews latent representations and final outputs via Shapley attribution. Together, these lenses provide a layered map of how conflict is encoded, organized, and resolved.

Our mechanistic analyses focus on Llama-3.1-8B-Instruct. To ensure our findings reflect fundamental cognitive mechanisms rather than architectural quirks, we replicate key trends on GPT-OSS-20B (OpenAI et al., 2025), which differs in both lineage and scale. Appendix D.4 reports the layer-wise AUC, ΔER curves, and projections for this model. The consistent conflict-type ordering and emergence patterns across both architectures confirm that these conflict-processing mechanisms are robust, inherent behaviors of modern LLMs.

4.1 Conflict Awareness via Concept Vectors

To understand how LLMs process confliction, we must first verify if they internally register it. Our core premise is that if a model possesses latent conflict awareness, its hidden states for conflicting

Model	Reasoning-based Tasks						Summarization-based Tasks					
	Infer.		Misinfo.		Temp.		Ambig.		Gran.		Persp.	
	Acc▲	Fth▲	Acc▲	Fth▲	Acc▲	Fth▲	Bal▼	Fth▲	Bal▼	Fth▲	Bal▼	Fth▲
gpt-5	43.6	76.8	**61.5**	83.2	**68.0**	88.5	26.1	81.8	**20.2**	**91.1**	**29.8**	71.5
claude-4.5-sonnet	**44.3**	72.8	54.6	81.7	61.9	**90.7**	24.3	83.8	22.9	87.5	31.0	63.7
gemini-2.5-pro	42.4	**80.0**	38.6	**89.9**	65.7	87.1	**20.5**	**92.2**	25.9	85.4	33.9	70.4
gpt-oss-120b	44.1	74.2	51.4	84.4	64.5	82.8	31.9	74.4	23.8	81.6	31.0	79.7
gpt-oss-20b	35.7	72.3	51.8	82.8	63.2	84.5	39.8	75.3	21.8	88.5	34.5	**92.5**
llama-3.1-70b-instruct	38.4	79.5	46.0	80.3	49.7	84.6	27.7	81.2	30.3	77.1	32.4	79.4
llama-3.1-8b-instruct	24.8	61.6	27.5	68.2	19.5	53.7	41.5	78.5	34.4	73.3	36.9	72.3
Llama-3.1-8B-Instruct												
+ CAS	34.2	58.7	31.8	66.1	34.7	49.8	35.7	**91.4**	28.7	84.3	33.5	80.8
+ All-Tokens	**37.0**	59.4	**35.4**	**66.9**	**38.7**	49.0	**30.8**	89.3	25.8	86.1	**27.7**	**82.6**
+ First-Generated	32.0	**63.2**	32.4	65.0	25.7	**51.2**	31.4	89.7	**25.7**	**88.6**	29.0	81.8
GPT-OSS-20B												
+ CAS	42.1	68.9	52.6	84.9	**65.5**	83.5	33.4	84.9	**23.9**	85.5	31.3	85.6
+ All-Tokens	45.7	68.5	**56.0**	81.4	64.1	83.2	28.8	84.3	24.0	83.0	32.0	**88.6**
+ First-Generated	**46.8**	**69.4**	51.9	**85.5**	59.8	**85.3**	**28.4**	**86.1**	24.8	**87.7**	**30.9**	83.5

Table 4: Performance across all conflict tasks (%). Acc ▲ and Fth ▲ (higher is better) are reported for reasoning; Bal ▼ (lower is better) and Fth ▲ for summarization. All-Tokens and First-Generated are our two injection schedules of $u^{(l)}$ (§5); CAS is the Context-Aware Steering alternative direction we compare against (§5). Best and second-best within each model block are **bold** / underlined.

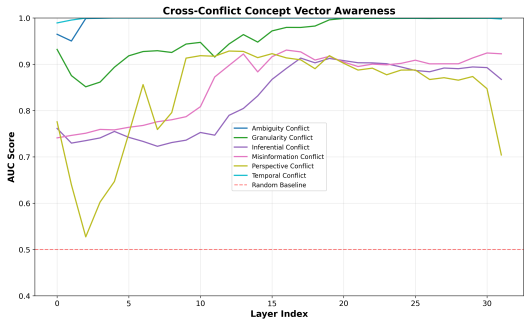

Figure 2: Layer-wise AUC for conflict vs. consistent sample classification. Higher AUC indicates stronger conflict awareness.

inputs should be linearly separable from consistent ones. We quantify this by constructing paired datasets for each conflict type t, consisting of a conflict version $x^{(t)}_{\text{conf}}$ and a consistent version $x^{(t)}_{\text{cons}}$ (details in Appendix D.1). At each layer l, we train a linear logistic-regression probe on $h_l(x)$ and measure the classification performance via AUC:

$$\text{AUC}^{(t)}_l = \text{AUC}\big(\{(h_l(x^{(t)}_{\text{conf}}), 1)\}, \{(h_l(x^{(t)}_{\text{cons}}), 0)\}\big) \quad (3)$$

This approach allows us to map the precise trajectory of conflict awareness. We reveal not only if the model detects a contradiction, but exactly where and how it emerges within the architecture.

Our findings show that models exhibit robust, layer-wise conflict awareness, with signal emergence tied to semantic complexity. Figure 2 illustrates that most conflict types reach high separability (AUC > 0.85) in mid-to-late layers, yet their developmental paths differ. *Temporal* and *ambiguity* conflicts saturate earliest, as they rely on explicit markers handled by early syntactic processing (Tenney et al., 2019). Conversely, *inferential* and *misinformation* conflicts emerge later. These require multi-step reasoning or world knowledge, which rely on abstract semantic representations from deeper layers. Other types show non-monotonic patterns: *granularity* follows a U-shape, while *perspective* conflicts peak in middle layers before declining during viewpoint reconciliation. These trends demonstrate that conflict awareness is not a single trigger. It is a dynamic, heterogeneous process that aligns with the model's progressive semantic refinement, as confirmed across multiple models in Appendix D.2.

4.2 Spectral Energy Analysis Reveals Conflict Dimensionality

Spectral energy analysis allows us to look beyond *whether* a model detects a conflict, and instead uncover *how* that information is geometrically organized in the latent space. Our core motivation is to determine the rank structure of hidden states: do conflict representations concentrate along a few dominant directions, or do they disperse across many dimensions? Answering this is vital, as it dictates the optimal subspace for targeted steering interventions. To capture this geometry, we analyze the rank structure via the spectral energy of the activations.

For each conflict type, we extract hidden states at the last non-padding token, forming matrices $\mathbf{H}^{(l)}_{\text{conf}}, \mathbf{H}^{(l)}_{\text{cons}} \in \mathbb{R}^{n\times d}$. We center each matrix as

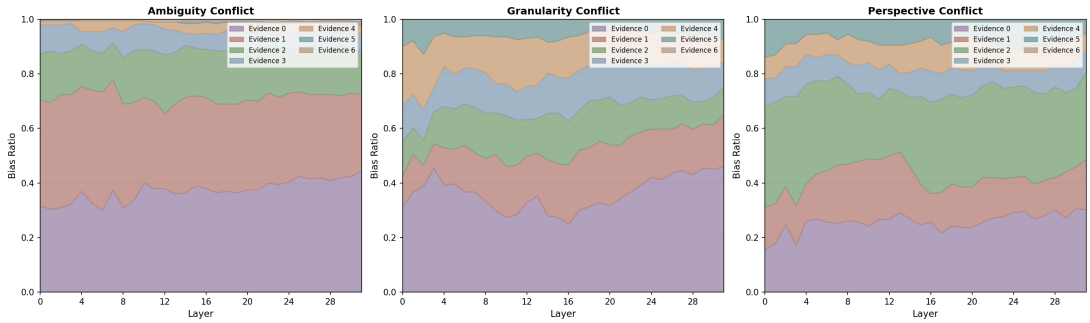

Figure 3: Layer-wise directional projection. $c^{(l)}$ consistently aligns with specific evidence directions, deviating from uniform.

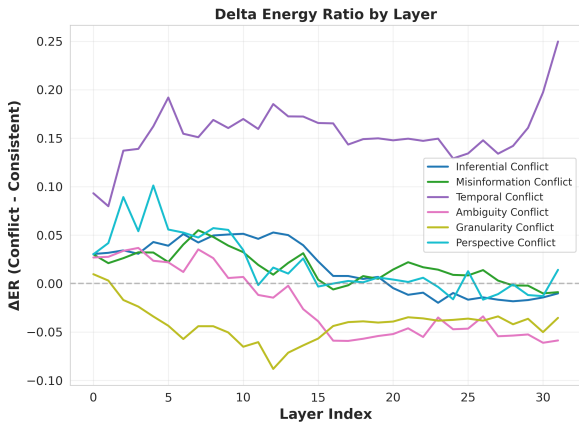

Figure 4: Delta energy ratio across layers. $\Delta\text{ER} > 0$ indicates concentrated representations; $\Delta\text{ER} < 0$ indicates dispersed ones.

$\tilde{\mathbf{H}} = \mathbf{H} - \bar{\mathbf{H}}$ and calculate the energy ratio (ER) and its delta:

$$\text{ER} = \frac{\sum_{i=1}^{k} \sigma_i^2}{\|\tilde{\mathbf{H}}\|_F^2}, \ \Delta\text{ER}_l^{(t)} = \text{ER}_{l,\text{conf}}^{(t)} - \text{ER}_{l,\text{cons}}^{(t)} \tag{4}$$

Here, $\sigma_1, \ldots, \sigma_k$ represent the top-k singular values (with k=10). A positive ΔER signifies low-rank compression in dominant directions, while a negative ΔER indicates dispersion across the broader dimensional space. Further implementation details are provided in Appendix D.3.

The resulting geometric patterns in Figure 4 demonstrate that models handle different conflict types through distinct dimensional strategies. *Temporal* conflicts compress into low-rank representations ($\Delta\text{ER} > 0$), effectively reducing the conflict to the singular dimension of event sequencing. Conversely, *granularity* and *ambiguity* conflicts disperse across dimensions ($\Delta\text{ER} < 0$); the former spans multiple levels of specificity, while the latter activates parallel lexical interpretations. *Inferential* and *misinformation* conflicts hover near zero, suggesting that reasoning processes reweight existing features rather than reorganizing the latent geometry. Finally, *perspective* conflicts exhibit strong layer-dependence, reflecting the gradual emergence of stance as an abstract property. These distinctions indicate that the effective rank change is not merely a byproduct of conflict presence, but a diagnostic signal of how each conflict type is internally structured, represented, and resolved. Collectively, these trends are qualitatively consistent across models in Appendix D.4 and reveal that conflict resolution is not a one-size-fits-all geometric process, but is instead dynamically tailored to the nature of the contradiction.

4.3 Evidence Position Bias in Internal Representations

We investigate "position bias" in LLMs during conflict resolution. Unlike standard retrieval tasks where bias often stems from irrelevant noise, our setting involves multiple legitimate but contradictory evidence pieces. The observed pattern therefore reveals an *implicit trust allocation*: the model's priority under direct competition.

Standard mitigations like reordering or positional adjustments are ineffective here, as they assume a single correct position or noise. Our empirical tests in Appendix D.8 confirm that random reordering fails to override the dominance of Evidence 1. As shown in Section 3.3, models disproportionately rely on the first evidence regardless of content. We analyze this systematic bias through two lenses: activation-space geometry and output-level Shapley attribution.

Representation-Level Analysis. To test whether the bias originates at the representation level rather than at decoding, we compare the model's combined-evidence activation against an unbiased reference built from single-evidence activa-

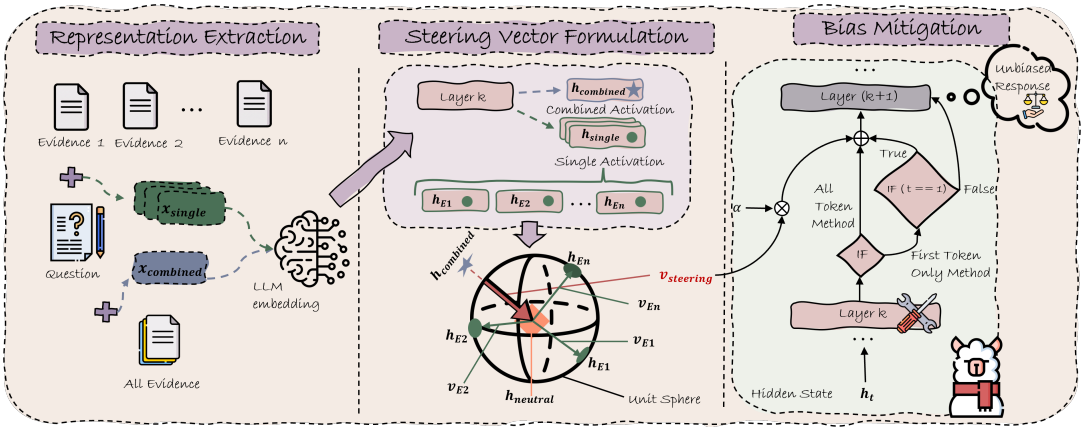

Figure 5: Three-stage activation steering method: (1) collect single-evidence and combined-evidence activations from calibration samples, (2) compute layer-wise steering directions toward neutral integration, (3) apply corrections during generation to mitigate positional bias.

tions. For a sample with K pieces, we build one combined-evidence prompt and K single-evidence prompts. At each layer l, we extract the final-token residual-stream activations: $c^{(l)}$ from the combined prompt, and $a_i^{(l)}$ from the i-th single-evidence prompt. If the model integrates all evidence neutrally, $c^{(l)}$ should align with the *neutral center* $\mu^{(l)} = \frac{1}{K}\sum a_i^{(l)}$. We quantify this alignment using the projection:

$$b_i^{(l)} = \frac{(c^{(l)} - \mu^{(l)}) \cdot (a_i^{(l)} - \mu^{(l)})}{\left(\|c^{(l)} - \mu^{(l)}\| + \epsilon_b\right)\left(\|a_i^{(l)} - \mu^{(l)}\| + \epsilon_b\right)} \tag{5}$$

As shown in Figure 3, this geometric bias persists across all layers. The combined representation consistently shifts toward the first evidence. This suggests the bias is a fundamental representation-level effect rather than a late-stage decoding artifact.

Output-Level Corroboration. Our Shapley-based attribution framework further confirms this first-evidence dominance. As seen in Appendix Figure 23, Evidence 1 disproportionately drives model outputs: accounting for 69.0% of contributions in Ambiguity, 38.4% in Granularity, and 35.0% in Perspective. These values significantly exceed uniform baselines. Across all seven tested models, ranging from 8B to 120B parameters, Evidence 1 dominance remains a consistent, systematic artifact. This reinforces the finding that models possess an inherent architectural preference for early evidence.

5 Activation Steering for Effective Conflict Resolution

We propose a training-free, label-free activation steering method to mitigate position bias. Our previous analysis (Section 4) demonstrates that position bias is encoded as a geometric asymmetry in the residual stream. This suggests that the intervention must operate directly within the hidden-state space rather than at the input or decoding stages. If evidence were integrated uniformly, the combined-prompt activation $c^{(l)}$ would align with a position-agnostic neutral center $\mu^{(l)}$. Since the observed gap $c^{(l)} - \mu^{(l)}$ consistently points toward the first evidence, we steer the representation by translating $c^{(l)}$ back toward the neutral center. As illustrated in Figure 5, our method comprises three stages: activation collection, direction computation, and additive injection.

Activation Collection and Direction Computation. We partition the data into calibration (20%) and test sets (80%). For each calibration sample n, we extract the final-token residual-stream activations at layer l: $\{a_{n,i}^{(l)}\}$ from K single-evidence prompts and $c_n^{(l)}$ from the combined prompt. We calculate the neutral center $\mu_n^{(l)} = \frac{1}{K}\sum_{i=1}^{K} a_{n,i}^{(l)}$ and average these values across the calibration set into $\bar{c}^{(l)}$ and $\bar{\mu}^{(l)}$. We define our steering direction as:

$$u^{(l)} = \frac{\bar{\mu}^{(l)} - \bar{c}^{(l)}}{\max\left(\|\bar{\mu}^{(l)} - \bar{c}^{(l)}\|,\ \epsilon_u\right)} \tag{6}$$

This fixed direction $u^{(l)}$ is then applied to all test instances.

Steered Inference. At inference time, we apply the precomputed direction to nudge the residual stream toward the neutral center: $h_t^{(l)} \leftarrow h_t^{(l)} + \alpha \cdot u^{(l)}$ with $\alpha = 1.0$. We evaluate two injection schedules: *First-Generated* (first step only) and *All-Tokens* (every step). As shown in Table 4, both models

achieve significant performance gains under All-Tokens. Llama-3.1-8B-Instruct shows improved summarization balance and substantial accuracy gains in reasoning (e.g., +19.2 points on Temporal). These improvements provide causal and targeted validation of our claim that the observed bias is rooted in the internal representation.

Analysis and Comparisons. To verify that the gains come from our contrastive design rather than generic context steering, we compare $u^{(l)}$ with *Context-Aware Steering (CAS)*. CAS amplifies the model's overall reliance on context, whereas $u^{(l)}$ targets the evidence-integration axis identified in our analysis. Our direction wins five of six model$\times$reasoning comparisons and remains robust on summarization balance, showing that mechanistically grounded representation-level rebalancing is more effective for tasks requiring selective cross-evidence integration.

The one exception is GPT-OSS-20B Temporal, where CAS outperforms our $u^{(l)}$ (Accuracy 65.5 vs. 64.1). We attribute this to task-mechanism alignment: temporal reasoning requires using every timestamped piece uniformly, so the task itself rewards CAS's blanket amplification of context reliance and offers little headroom for the selective rebalancing our direction is designed for. The boundary clarifies when each direction is preferred: CAS suits tasks demanding uniform integration; our $u^{(l)}$ suits tasks requiring selective cross-evidence integration, which covers most reasoning conflicts.

The All-Tokens schedule yields the largest accuracy gains; the minor decrease in reasoning faithfulness reflects a shift toward multi-evidence reasoning that exceeds verbatim grounding.

6 Conclusion

We release a 5,781-sample multi-domain dataset of contextual knowledge conflicts spanning six conflict types across reasoning and summarization tasks. Our mechanistic analysis shows that LLMs often detect conflicts internally, yet still allocate disproportionate trust to earlier evidence, producing representation-level positional bias. Based on this finding, we propose a training-free, label-free activation steering method that mitigates this bias and improves evidence integration under conflict.

Limitations and Future Work

We acknowledge several limitations. First, our method requires white-box access to residual-stream activations, so it applies only to open-weight models and adds inference-time overhead. Appendix D.5 shows the same position bias in closed-source systems, so the phenomenon itself is not limited to white-box settings. Second, compute constraints restrict our steering experiments to GPT and LLaMA models at the 8B and 20B scale; scaling to 70B or 120B is left for future work. Third, our dataset uses synthetic samples for controlled manipulation; evaluating on noisy, real-world retrieval with heterogeneous evidence remains open. Fourth, our Shapley-based attribution assumes each evidence piece contributes equally in expectation, which may not hold when evidence quality varies in real-world pipelines. A weighted Shapley formulation with source-reliability priors, used alongside Faithfulness, is a promising extension.

Ethical considerations

Our dataset contains synthetically generated misinformation designed to test model robustness against false information. The dataset does not include personally identifying information or offensive content. To mitigate potential risks of misuse, we clearly label each evidence piece as factually correct or incorrect in the dataset metadata. We emphasize that the misinformation samples are constructed solely for research purposes to evaluate conflict handling capabilities and should not be used to train models for generating misleading content. We will include explicit usage guidelines with the dataset release to prevent misuse. All base datasets used in constructing this dataset are publicly available and published through established academic channels, governed by open-access licenses that permit research reuse.

Acknowledgments

We thank the University of Florida Research Computing HiPerGator for providing computational resources and UF NaviGator for providing access to LLM APIs. We also acknowledge Delta at the National Center for Supercomputing Applications through allocation CIS251209 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2024. Make your llm fully utilize the context. In *Advances in Neural Information Processing Systems*, volume 37, pages 62160–62188. Curran Associates, Inc.

Arie Cattan, Alon Jacovi, Ori Ram, Jonathan Herzig, Roee Aharoni, Sasha Goldshtein, Eran Ofek, Idan Szpektor, and Avi Caciularu. 2025. Dragged into conflicts: Detecting and addressing conflicting sources in search-augmented llms. *Preprint*, arXiv:2506.08500.

Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. 2024. Roast: Review-level opinion aspect sentiment target joint detection for absa. *Preprint*, arXiv:2405.20274.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. 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.

Juan Chen, Baolong Bi, Wei Zhang, Jingyan Sui, Xiaofei Zhu, Yuanzhuo Wang, Lingrui Mei, and Shenghua Liu. 2025. Rethinking all evidence: Enhancing trustworthy retrieval-augmented generation via conflict-driven summarization. *Preprint*, arXiv:2507.01281.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle:discovering diverse perspectives about claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 542–557, Minneapolis, Minnesota. Association for Computational Linguistics.

Eunseong Choi, June Park, Hyeri Lee, and Jongwuk Lee. 2025. Conflict-aware soft prompting for retrieval-augmented generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26969–26983, Suzhou, China. Association for Computational Linguistics.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7358–7370, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yibing Du, Antoine Bosselut, and Christopher D. Manning. 2022. Synthetic disinformation attacks on automated fact verification systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10581–10589.

Carl Eckart and Gale Young. 1936. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, and 16 others. 2024. FOLIO: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22017–22031, Miami, Florida, USA. Association for Computational Linguistics.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

Saeedeh Javadi, Sara Mirabi, Manan Gangar, and Bahadorreza Ofoghi. 2026. Contradictions innbsp;context: Challenges fornbsp;retrieval-augmented generation innbsp;healthcare. In *Advances in Information Retrieval: 48th European Conference on Information Retrieval, ECIR 2026, Delft, The Netherlands, March 29 – April 2, 2026, Proceedings, Part I*, page 34–48, Berlin, Heidelberg. Springer-Verlag.

Jing Jin, Yuhan Song, Wen Luo, and Houfeng Wang. 2025. From bias to benefit: Place good documents in good positions.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2668–2677. PMLR.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liška, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: assessing temporal generalization in neural language models. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

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*.

Gaotang Li, Yuzhong Chen, and Hanghang Tong. 2025. Taming knowledge conflicts in language models. In *Forty-second International Conference on Machine Learning*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

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.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Sara Patel, Mingxun Zhou, and Giulia Fanti. 2025. Maxshapley: Towards incentive-compatible generative search with fair context attribution. *Preprint*, arXiv:2512.05958.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H. Chen. 2024. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *npj Digital Medicine*, 7(1):20.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. ConflictBank: A benchmark for evaluating the influence of knowledge conflicts in llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 103242–103268. Curran Associates, Inc.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025. Retrieval-augmented generation with conflicting evidence. In *Second Conference on Language Modeling*.

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 *International Conference on Learning Representations*, volume 2024, pages 35623–35646.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

Zikun Ye and Hema Yoganarasimhan. 2025. Fair document valuation in llm summaries via shapley values. *Preprint*, arXiv:2505.23842.

Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, and

Zhangyang Wang. 2024. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to ai transparency. *CoRR*, abs/2310.01405.

A Extended Related Work

Knowledge Conflicts and Datasets. Large language models often produce inconsistent or hallucinatory outputs when confronted with conflicting knowledge (Xie et al., 2024; Xu et al., 2024). Prior datasets primarily focus on context-memory conflicts (Longpre et al., 2021), often constructed through entity replacement (Su et al., 2024). However, these resources exhibit four major limitations. They rely heavily on synthetic conflicts with limited real-world complexity, emphasize explicit surface-level contradictions while under-covering implicit multi-step conflicts, provide limited domain diversity, and suffer from severe class imbalance that weakens category-level analysis. In contrast, we propose six fine-grained conflict categories spanning multiple domains, with explicit and implicit coverage across reasoning and summarization task types and type-specific implicit proportions. This design addresses key gaps in conflict complexity, diversity, and evaluation fairness. Existing datasets such as ConflictBank (Su et al., 2024) focus on entity-substitution-based factual conflicts, and AmbigDocs (Lee et al., 2024) targets a single ambiguity type. To the best of our knowledge, few existing datasets jointly cover all six conflict types and two task paradigms within one multi-domain dataset.

Conflict Mitigation in RAG. Recent work addresses multi-source conflicts through several approaches. These include multi-agent deliberation (Li et al., 2025; Wang et al., 2025), conflict-driven summarization (Chen et al., 2025), and adversarial-trained assessors (Choi et al., 2025; Javadi et al., 2026). A separate line of work uses contrastive decoding to amplify the contribution of context relative to parametric memory, most notably context-aware decoding (CAD) (Shi et al., 2024). CAD contrasts $\log p(y \mid \text{ctx})$ with $\log p(y \mid \emptyset)$, where the counterfactual baseline is the empty context. This binary contrast treats the K evidence pieces as a single block: it amplifies trust in context as a whole but cannot redistribute attention or representational contribution among the individual evidence pieces. The position bias we identify is precisely an intra-context asymmetry ($b_1^{(l)} \gg b_{i>1}^{(l)}$), which CAD's reference frame cannot express; the two methods therefore address orthogonal problems and cannot be directly compared as alternatives. These methods improve QA performance but largely treat conflict resolution as a black-box problem, often requiring additional models or training and offering limited insight into internal mechanisms. We instead focus on internal conflict representations and propose a training-free activation steering method without external models. More importantly, to the best of our knowledge, few dataset-based studies provide a comparably broad mechanistic account of when and why LLMs fail in contextual conflict resolution.

Evidence Attribution for Multi-Document Summarization. Shapley values have been used to quantify document importance in LLM-generated summaries (Ye and Yoganarasimhan, 2025). Recent work improves efficiency through semantic clustering and decomposable utility functions (Ye and Yoganarasimhan, 2025; Patel et al., 2025). However, these methods rely on LLM-as-a-judge scoring and focus on content provider compensation scenarios rather than conflict analysis. We introduce Shapley attribution specifically for fairness analysis in conflict scenarios, proposing the Balance metric to quantify evidence-integration bias via a normalized Gini coefficient.

Mechanistic Interpretability of LLMs. Prior work has probed model representations using concept activation vectors (Kim et al., 2018), spectral decomposition (Raghu et al., 2017), and sparse autoencoders (Huben et al., 2024), primarily to study factual recall, reasoning, or sentiment. Yet mechanistic analysis of how LLMs internally handle contextual knowledge conflicts remains scarce (Xu et al., 2024). To the best of our knowledge, few dataset-based studies jointly analyze conflict awareness, representational geometry, shared feature organization, and evidence-integration bias across conflict types and tested model architectures.

Positional Bias in LLMs. LLMs exhibit positional bias over long contexts (Liu et al., 2024), which has been addressed through positional-encoding adjustments (Zhang et al., 2024), document reordering (Jin et al., 2025), or training-time augmentation (An et al., 2024). However, prior studies mainly examine retrieval under noisy context, where most evidence is irrelevant and positional bias determines whether one relevant item is recovered from surrounding noise. In conflict scenarios, by contrast, all evidence is relevant but semantically competing and often logically incompatible. Positional bias therefore becomes a selective-integration problem, reflecting the model's tendency to favor specific evidence positions under genuine competition. To our knowledge, this setting remains underexplored. We provide a systematic analysis of positional bias in this regime and propose a training-free activation steering method (Zou et al., 2023; Rimsky et al., 2024) that improves reasoning performance and maintains strong summarization evidence-integration balance on tested models.

B Dataset Details

B.1 Construction Process

Table 5 lists the ten base datasets used to construct our dataset. The construction procedure for each conflict type is described in Section 2.2; here we provide the GPT-5 prompts used for semi-synthetic data generation. Figure 6 shows the two-step prompt for misinformation conflicts (SciFact subset): question generation followed by conflicting evidence generation. Figure 7 shows the prompt for inferential conflicts (FOLIO and ENTAILMENT-BANK subsets), which augments original premises with conflicting reasoning branches. The annotation interface for manual re-annotation of FOLIO instances is shown in Figure 8. Figure 9 shows the prompt for temporal conflicts (ConflictBank subset), which attaches explicit temporal markers to generate time-dependent questions. Figure 10 shows the prompt for generating factual questions for other ConflictBank conflict types.

B.2 Decision Rules for Borderline Conflict-Type Overlaps

Our taxonomy is defined by operational, structural criteria rather than surface intuition. We document the decision rule for each pair of conflict types whose surface descriptions could otherwise be confused.

Temporal vs. Misinformation. The criterion is whether each evidence piece holds true at some time point. In *temporal* conflicts, every evidence piece is true at its own time point; the conflict arises because claim validity changes over time. In *misinformation* conflicts, some evidence is factually incorrect, and each piece carries an accuracy label in the dataset metadata.

Ambiguity vs. Inferential. The criterion is the task goal, not merely the number of reasoning steps. In *ambiguity* conflicts, each evidence piece describes a different real entity that shares the same name (e.g., "Jordan" can refer to the basketball player or to a university teacher with the same name); each piece is understandable on its own, no cross-evidence reasoning is needed, and the goal is to integrate and present every entity fairly rather than let the more popular referent dominate. Ambiguity instances are identified directly from the entity-disambiguation metadata of the source dataset. In *inferential* conflicts, the conflict is not limited to entities: it only emerges after combining multiple pieces of evidence through reasoning, and the goal is to derive the single correct conclusion.

Perspective vs. Granularity. The criterion is compatibility. *Perspective* conflicts involve incompatible stances on the same question. *Granularity* conflicts involve compatible answers at different levels of specificity.

B.3 Data Samples

We provide representative examples for each conflict type in our dataset. Figure 11 illustrates an inferential conflict where logical reasoning leads to an uncertain conclusion. Figure 12 presents a misinformation conflict involving contradictory scientific claims with accuracy labels. Figure 13 demonstrates a temporal conflict requiring temporal reasoning across events. Figure 14 shows an ambiguity conflict where one name refers to multiple distinct entities. Figure 15 provides a granularity conflict example with varying levels of detail, and Figure 16 highlights a perspective conflict involving diverse viewpoints on a policy issue.

C Evaluation Details

C.1 Evaluation via Different Scorer Models

Our Balance metric relies on a scorer model to compute the value function. As defined in Eq. 2 of Sec-

Base dataset	Type	Description
AmbigDocs (Lee et al., 2024)	Synthetic	A multi-document reasoning dataset with entities sharing the same name. Tests models' ability to disambiguate entities and produce complete answers under ambiguous evidence.
ENTAILMENTBANK (Dalvi et al., 2021)	Non-synthetic	A science QA dataset with multi-step entailment chains. Shows how answers can be derived step-by-step from foundational facts.
NEJM-MedQA (Savage et al., 2024)	Non-synthetic	Combines MedQA (USMLE-style questions) with real NEJM clinical cases. Evaluates diagnostic reasoning in both exam and real-world clinical contexts.
FOLIO (Han et al., 2024)	Non-synthetic	An expert-written logical reasoning dataset annotated with first-order logic. Contains diverse reasoning structures for evaluating formal reasoning in realistic settings.
AllSides	Non-synthetic	We collect recent news articles from the AllSides website. For each topic, we gather articles from media outlets representing different political stances (left, center, right), providing source material for perspective-conflict construction.
Perspectrum (Chen et al., 2019)	Non-synthetic	Contains controversial claims with stance sentences and supporting evidence. Designed for multi-perspective discovery and understanding.
ROAST-ABSA (Chebolu et al., 2024)	Non-synthetic	A review-level aspect-based sentiment analysis dataset across multiple languages and domains. Includes aspect-opinion-sentiment-target quadruples for fine-grained analysis.
SciFact (Wadden et al., 2020)	Non-synthetic	A scientific claim verification dataset with expert-labeled evidence as supporting or refuting. Evaluates evidence retrieval, verification, and explanation.
ConflictBank (Su et al., 2024)	Synthetic	A systematically constructed dataset that injects plausible but incorrect distracting knowledge. Tests model robustness and truthfulness under conflicting evidence.
CONFLICTS (Cattan et al., 2025)	Non-synthetic	A RAG dataset built on real web search results with expert-annotated conflict types. Evaluates conflict recognition and response quality under multi-source knowledge conflicts.

Table 5: Base datasets used to construct our dataset.

tion 3.1, this function maps each evidence subset to a normalized utility score. To verify the robustness of our metric, we test two different scorer models: Llama-3.1-8B and Gemma-2B. These models differ in both scale and pretraining approach.

Table 6 shows Balance scores computed using both scorer models. While absolute values differ between scorers, relative performance patterns remain highly consistent across tested models and conflict types. Crucially, model rankings by Balance remain nearly identical across scorers. For instance, `gemini-2.5-pro` consistently achieves the best Balance scores in Ambiguity under both scorers, while `llama-3.1-8b-instruct` consistently shows the highest Balance scores across conflict types. These results suggest that our Shapley-based Balance metric is robust to scorer choice and less likely to be driven by scorer-specific artifacts.

C.2 Human Annotation and Metric Validation

To validate our Shapley-based metric, we conduct a human annotation study. Two independent annotators evaluate evidence contributions for a subset of samples from summarization tasks. Figure 17 shows the annotation interface. Annotators rate each evidence source's contribution to the response on a 1–5 scale. We provide detailed guidelines in Figure 18. Table 2 reports Cohen's Kappa coefficients across 108 annotated samples. Human annotators achieve moderate agreement ($\kappa = 0.5199$). Our automatic metric also agrees with both annotators ($\kappa = 0.5208$ and 0.5014). These results suggest that the metric is broadly consistent with human judgments of evidence contribution. Table 7 reports per-model Balance scores from both human raters and the LLM-as-a-Judge across the three summarization tasks.

Misinformation: Question Generation

Claim: {claim}
Original Evidence: {evidence_context}

Generate a Clear Question: Create a direct, focused question based on the claim that can be answered using scientific evidence. The question should be specific and answerable.

Misinformation: Conflicting Evidence Generation

Task: Generate conflicting evidence statements that oppose the original findings.

Requirements:
1. **Quantity & Style:** Generate the exact same number of statements as provided. Match the language style, professional tone, and complexity perfectly.
2. **Conflict Strategy:** Create plausible rebuttals. Do not use simple "not" negations; instead: modify findings, cite different methodological outcomes, or focus on alternative variables that lead to opposite conclusions.
3. **Diversity:** Ensure each statement uses a unique sentence structure.
4. **Output:** Return ONLY the numbered list of conflicting evidence statements.

Claim: {claim}
Original Evidence: {evidence_context}

Figure 6: Prompts for misinformation conflict construction.

FOLIO: Conflicting Premise Generation

Your task:
1. **Decision:** With 50% probability, choose either to:
- **Option A:** Add 1–2 plausible statements that create a conflicting reasoning branch (requires at least one inference step; avoid trivial "X vs. not-X").
- **Option B:** Keep the premises as they are.
2. **Output:** Return ONLY the final numbered list of statements.
3. **Constraints:** No meta-talk, no "source_#" references, no explanations.

Premises: {premises_raw_json}
Hypothesis: {conclusion}

Figure 7: Prompt for FOLIO inferential conflict construction.

D Mechanistic Analysis Details

D.1 Conflict-Consistent Pair Construction and Cross-Validation Statistics

The concept-vector analysis in Section 4 relies on pairing each conflict instance $x^{(t)}_{\text{conf}}$ with a consistent counterpart $x^{(t)}_{\text{cons}}$ in which the evidence no longer produces a conflict. We construct each consistent counterpart by preserving the question and surrounding context and replacing or filtering only the evidence set, following per-type rules:

- **Misinformation.** We retain only the evidence pieces labeled factually correct in the source annotation and discard the conflicting ones, so that every remaining piece supports the same gold conclusion.
- **Inferential (FOLIO and EntailmentBank).** We keep the original entailment chain and remove the GPT-5-generated conflicting branch, so that all premises jointly support the original hypothesis label.
- **Temporal.** We anchor on a single timestamp and retain only the evidence pieces consistent with that anchor; we introduce no new content.
- **Granularity.** We keep evidence pieces drawn from the same diagnostic specificity level (either all broad-syndrome or all specific-disease), without mixing levels.
- **Perspective.** We retain evidence pieces that share the same stance label and discard those expressing the opposing stance.

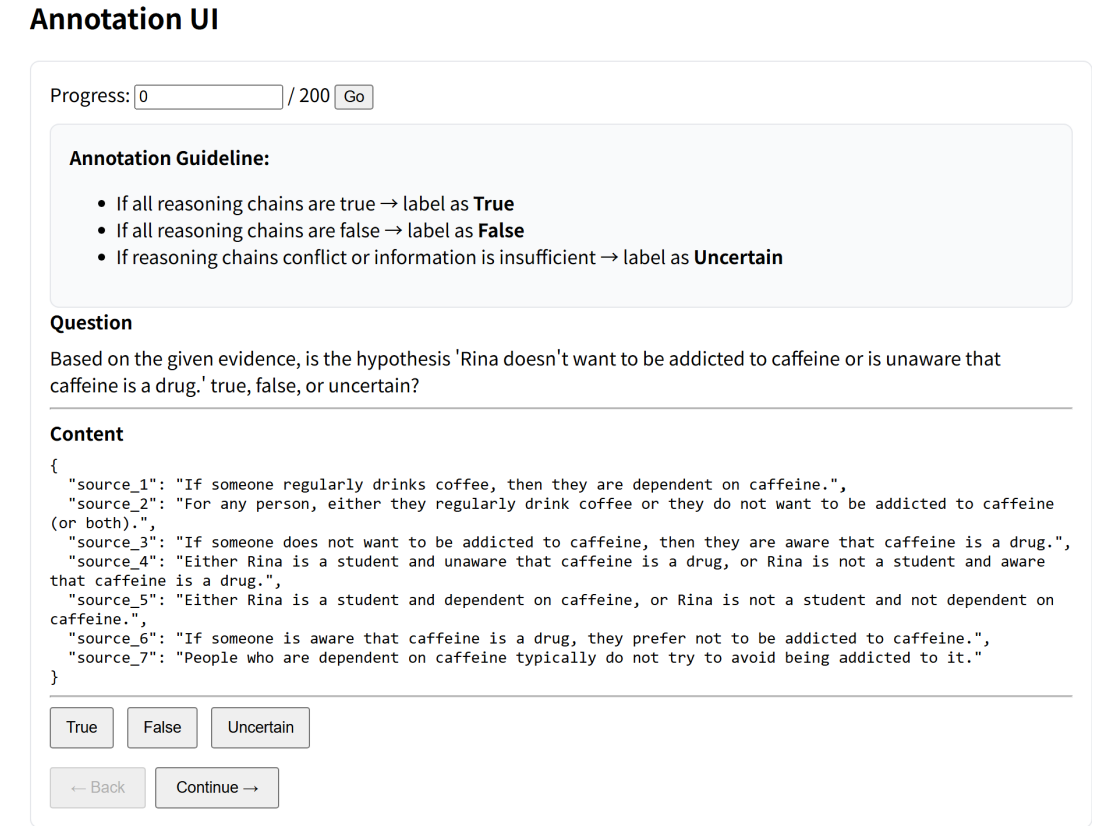

Figure 8: Annotation interface for FOLIO manual verification. Annotators select True/False if reasoning chains lead to a unique answer, or Uncertain if chains produce conflicting conclusions.

- **Ambiguity.** We retain evidence pieces that refer to a single entity, using the source dataset's disambiguation metadata.

Each conflict instance is paired with exactly one consistent counterpart, and we use these conf-versus-cons pairs as the labeled inputs to the linear probes in Section 4. Table 8 reports per-type sample counts and 5-fold stratified cross-validation statistics. Probes are linear logistic-regression classifiers with inverse-regularization strength C=1.0 and a maximum of 1000 iterations, fit on residual-stream activations at each layer.

D.2 Concept Vector Analysis on Additional Models and Implicit Conflicts

We extend concept vector analysis to additional models and examine implicit versus explicit conflicts.

Analysis on Additional Models. Figure 19 shows that GPT-OSS-20B exhibits awareness trends similar to those of Llama models despite having fewer layers. Temporal and ambiguity conflicts reach saturation quickly, while other conflict types show gradual emergence. The model exhibits strong conflict awareness across all types. These findings suggest that our observations generalize across the tested model scales and architectures.

Implicit vs. Explicit Conflicts. We analyze perspective and inferential conflicts from different source datasets. We operationalize the explicit/implicit distinction using the structural criterion in Section 2.1. A conflict is **explicit** if it requires at most one inferential step to detect, and **implicit** if it requires at least two cross-evidence inferential steps.

Inferential conflicts. ENTAILMENTBANK (explicit): Each reasoning chain presents a step-by-step entailment that directly yields a stated conclusion. When a second evidence piece asserts an incompatible conclusion, the contradiction is identifiable by direct comparison of their final claims, requiring a single inferential step. FOLIO (implicit): GPT-5-generated premises introduce a conflicting reasoning branch by altering conditional or logical dependencies. To identify the conflict, a reader must trace causal relationships through

Temporal Conflicts: Question Generation

Based on two conflicting claims about the same subject, generate ONE concise open-ended question with a temporal marker.

Inputs: Subject, default claim, temporal claim, default evidence, temporal evidence

Requirements:
1. Extract ALL timestamps/dates from both evidences.
2. Identify the attribute being described (e.g., employer, award, title).
3. Randomly choose to ask about a SPECIFIC YEAR before, between, or after identified timestamps.
4. Use patterns like: "What was {subject}'s [attribute] in [year]?" or "Who was {subject}'s [attribute] as of [year]?"
5. Prefer specific years over time ranges.
6. Do NOT use yes/no questions or multiple-choice formats.
7. Self-contained, one sentence, $\leq$ 20 words.
8. Output only the question (no explanations, no metadata).

Figure 9: Prompt for temporal conflict question generation.

General ConflictBank: Question Generation

Given a subject with a default claim (assumed correct) and a conflicting claim, generate ONE concise factual question about the subject without revealing either claim explicitly.

Inputs: Subject, default claim, conflicting claim

Requirements:
- Use simple WH-questions like "Where did {subject} work?", "When was {subject} born?", "Who is {subject} married to?"
- Do NOT include claim contents in the question.
- Self-contained, one sentence.

Figure 10: Prompt for general ConflictBank question generation.

Inferential Conflict Example

Question: Based on the given evidence, is the hypothesis 'Miroslav Venhoda loved music.' true, false, or uncertain? The final answer should be 'True', 'False', or 'Uncertain'. Only answer 'Uncertain' if the reasoning results from different evidence pieces conflict with each other.
Evidence 1: Miroslav Venhoda was a Czech choral conductor.
Evidence 2: Miroslav Venhoda specialized in the performance of Renaissance and Baroque music.
Evidence 3: All choral conductors are musicians.
Evidence 4: Some musicians love music.
Evidence 5: In 1946, Miroslav Venhoda published a book titled Method of Studying Gregorian Chant.
Evidence 6: Authors of instructional books on musical practice often have strong enthusiasm for music.
Evidence 7: A scholar's focus on liturgical chant can be motivated by historical or religious interest rather than a personal love of music.
Ground Truth: Uncertain

Figure 11: Example of an Inferential Conflict, demonstrating logical uncertainty.

multiple premises and compare derivations across evidence chains. This process requires combining at least two evidence pieces through conditional logic before the incompatibility surfaces, satisfying our implicit criterion.

Perspective conflicts. Perspectrum (explicit): Evidence pieces contain explicit stance sentences that directly affirm or negate the same claim (e.g.,

Misinformation Conflict Example

Question: In a bleomycin-induced mouse model of pulmonary fibrosis, does pharmacological activation of AMPK (e.g., with metformin) reduce established lung fibrosis in an AMPK-dependent manner? Answer with Yes or No.
Evidence 1: In a bleomycin model of lung fibrosis in mice, metformin therapeutically accelerates the resolution of well-established fibrosis in an AMPK-dependent manner.
Evidence 2: In primary myofibroblasts isolated from IPF lungs, pharmacological activation of AMPK did not attenuate fibrotic activity; instead, it was accompanied by diminished mitochondrial biogenesis and persistence of apoptosis resistance.
Evidence 3: Contrary to expectations, these investigations indicate that heightened AMPK signaling underlies non-resolving, pathologic fibrotic remodeling, and argue that metformin (and related AMPK agonists) are unlikely to reverse established fibrosis because they maintain myofibroblasts in an activated, apoptosis-resistant state.
Accuracy Labels: [True, False, False]
Ground Truth: Yes

Figure 12: Example of a Misinformation Conflict with specific accuracy labels for each evidence.

Temporal Conflict Example

Question: Which university did Alvaro Mata attend in 2030?
Note: Evidence snippets are truncated for readability; we use [...] as the truncation marker.
Evidence 1: In the heart of Cleveland, Ohio[...] lies Cleveland State University. It is here that Alvaro Mata, a researcher of unyielding curiosity and dedication, embarked on a transformative journey[...] It is the early 2000s, and Cleveland State University is thriving[...] Alvaro Mata emerges as a bright and ambitious individual[...] Evidence of Alvaro's attendance at Cleveland State University can be found in the university's library archives, where academic papers and theses reveal his prolific output as a student researcher.
Evidence 2: Renowned Researcher Alvaro Mata Celebrates 14-Year Milestone at Macquarie University. Sydney, Australia - December 15, 2038. Today, Macquarie University in Sydney marked a significant milestone as esteemed researcher Alvaro Mata completed an impressive 14-year tenure at the institution[...] "It's hard to believe it's been 14 years since I first stepped onto this campus," Mata reflected.
Ground Truth: Macquarie University

Figure 13: Example of a Temporal Conflict requiring chronological reasoning.

"X is beneficial" vs. "X is harmful"). The viewpoint conflict is identifiable by direct comparison of these surface propositions in a single inferential step. AllSides (implicit): Evidence pieces describe the same event through selective emphasis, differential fact selection, and divergent rhetorical framing, without any single statement directly contradicting another. Detecting the underlying viewpoint conflict requires integrating implicit stances across multiple documents. The process demands multi-step cross-document synthesis to identify what each article implies but does not state, which satisfies our implicit criterion on structural grounds independent of the data source.

Figure 20 compares inferential conflicts. ENTAILMENTBANK (explicit) maintains consistently high AUC throughout all layers. FOLIO (implicit) shows lower AUC in early layers and continues to decline in final layers. The gap suggests that models rely heavily on surface-level signals for conflict detection and struggle when contradiction requires multi-step cross-evidence inference.

Figure 21 compares perspective conflicts. The difference is even more striking: Perspectrum (ex-

Ambiguity Conflict Example

Question: What is Peter Cusack known for?
Evidence 1: Peter Cusack (musician) is an artist and musician who is a member of CRiSAP (Creative Research in Sound Arts Practice), and is a research staff member at the London College of Communication. He was a founding member and director of the London Musicians' Collective. He is best known as a member of the avant garde musical quartet, Alterations (1978–1986; with Steve Beresford, David Toop, and Terry Day), and the creator of field and wildlife recording-based albums.
Evidence 2: Peter Cusack made his premiership début with the Sydney Roosters in the 1998 NRL season. A front-rower, Cusack was one of the last remaining top level league players to hold a job outside football, working part-time as a plumber. He played in the 2000 NRL Grand Final loss to the Brisbane Broncos. Cusack was awarded the 2002 Sydney Roosters season's "Clubman of the Year" and played in their 2002 NRL Grand Final victory over the New Zealand Warriors.

Figure 14: Example of an Ambiguity Conflict involving two different individuals with the same name.

Granularity Conflict Example

Task: Summarize the following evidence in 2-3 sentences.
Evidence 1: I'm only here for taste and texture, and the recent cups haven't hit the same mark. The broth turns thin unless I dump the entire seasoning in, and even then I reach for soy sauce to wake it up. The noodles rehydrate fine—springy enough after five minutes under a kettle—but the base lacks that savory backbone it used to have[...] If you're chasing a rich sip straight from the cup, this version feels watered down.
Evidence 2: The "lower sodium" splash on the front sent me straight to the Nutrition Facts, and the numbers tell a different story than I expected. The serving size is smaller than my older cups[...] Net weight is down, the water line looks lower, and the sodium per serving drops mostly because the serving itself is lighter. It reads like less soup rather than a smarter recipe.
Evidence 3: As an overall lunch option, these are still easy: boil water, fill to the line, lid on for five minutes, stir hard, and you've got something warm at your desk. The flavor is dependable once you learn your routine[...] The convenience is great for busy days, yet the front-of-pack claims don't line up neatly with what I taste and how much I'm actually eating.
Evidence 4: I prefer the large McDougall cups because one can carry me through an afternoon of meetings without snacks. Filled to the line, it's closer to a bowl than a snack cup, and sometimes I can't finish it in one sitting.
Evidence 5: For anyone sensitive to salt, using about three-quarters of the seasoning packet has been the sweet spot. I stir in the dry mix gradually, taste, and stop right before it tips into that heavy-salty zone; a squeeze of lemon and some scallions round it out.

Figure 15: Example of a Granularity Conflict where evidence varies in detail and focus.

plicit) achieves stable high AUC across layers, whereas AllSides (implicit) fluctuates near chance level. This suggests that model representations are less sensitive to conflicts that require multi-step cross-document synthesis, regardless of whether those conflicts arise from logical structure (FOLIO) or selective framing (AllSides). The shared structural factor is the number of inferential steps required.

D.3 Spectral Energy Analysis Implementation Notes

The three-stage pipeline and the ER and ΔER formulas are defined in Section 4.2. We list here the numerical and software details omitted from the main text. We compute the top-k singular values

Perspective Conflict Example

Question: Should the school day be extended?
Evidence 1: Evidence from Chicago Public Schools shows that assigning underprepared 9th graders to a double-period (extended time) algebra course raised math achievement, increased credit accumulation, and led to higher graduation and college enrollment rates—indicating targeted extensions of the school day for core instruction can produce substantial gains for struggling students.
Evidence 2: The AAP recommends middle and high schools start at 8:30 a.m. or later because adolescents need 8–10 hours of sleep; insufficient sleep is linked to worse grades, depression, and accident risk. Extending the school day without shifting to later start times may exacerbate sleep deprivation and harm learning and health.
Evidence 3: National survey data show teachers report substantially higher job-related stress and burnout than other working adults, with time pressure and workload as leading contributors. Lengthening the school day risks worsening burnout and turnover unless paired with added staffing, planning time, and compensation.
Evidence 4: On school days, juvenile crime and victimization rates peak in the hours immediately after school (roughly 2–6 p.m.). Extending the school day to provide structured supervision during these hours is argued to reduce delinquency and enhance community safety.
Evidence 5: A package of reforms that included substantially increased instructional time produced large gains in math and reading. The study suggests extended time can be effective when embedded in a broader model, implying time alone is unlikely to deliver similar results.
Evidence 6: The EEF estimates extending school time yields a small average impact on learning (around +2 months' progress per year) at moderate cost; benefits are larger when time is tightly targeted (e.g., tutoring), raising questions about the cost-effectiveness of blanket school-day extensions.

Figure 16: Example of a Perspective Conflict featuring diverse viewpoints on school hour extensions.

with PyTorch's `svd_lowrank` on the centered activation matrix $\tilde{\mathbf{H}}$. We set k=10 and $\epsilon_{\text{ER}} = 10^{-12}$ for numerical stability. Activations are taken at the last non-padding token of each sample, matching the protocol used in Section 4.2.

D.4 Spectral Energy Analysis in Other Models

Figure 22 shows delta energy-ratio patterns in `openai/gpt-oss-20b`. The type-specific geometric patterns remain consistent with Llama-3.1-8B, suggesting that similar geometric trends appear across tested models.

D.5 Evidence Position Bias Across Tested Models

We analyze evidence-position bias across seven tested models spanning different scales and training approaches. Figures 23–29 show evidence-contribution distributions for these tested models. The bias is pervasive within this dataset: earlier evidence consistently receives larger contributions. This pattern appears in both small models (Llama-3.1-8B) and large models (GPT-5, Claude-4.5-Sonnet), and in both proprietary and open-source systems. In our experiments, this bias is not removed by model scale or training approach.

D.6 Bias Measurement Implementation Notes

The measurement setup, neutral center $\mu^{(l)}$, and normalized directional projection $b_i^{(l)}$ are defined in Section 4.2's sibling subsection on representation-level position bias. We list here the numerical and edge-case details omitted from the main text.

Numerical constants. We set $\epsilon_b = 10^{-8}$ inside the projection denominator for numerical stability. If $\|d^{(l)}\|$ or $\|v_i^{(l)}\|$ falls below the degeneracy threshold $\tau_b = 10^{-12}$, we set $b_i^{(l)} = 0$ for that sample-layer pair and log it as a skipped projection.

Interpretation. The projection $b_i^{(l)} \in [-1, 1]$ measures cosine alignment between the deviation

Model	Ambiguity		Granularity		Perspective	
	Llama-3.1-8B	Gemma-2B	Llama-3.1-8B	Gemma-2B	Llama-3.1-8B	Gemma-2B
gpt-5	24.7	20.7	16.7	15.2	26.3	19.9
claude-4.5-sonnet	23.6	19.4	18.9	15.7	26.2	18.1
gemini-2.5-pro	19.3	15.9	20.0	16.3	27.5	17.3
gpt-oss-120b	30.0	25.1	20.0	15.6	27.3	22.0
gpt-oss-20b	38.4	27.0	18.5	16.5	26.7	17.0
llama-3.1-70b-instruct	26.3	20.5	24.6	23.0	30.2	26.1
llama-3.1-8b-instruct	39.9	27.7	29.9	28.5	33.1	26.8

Table 6: Balance scores under two different scorer models. The relative rankings and performance patterns remain consistent across scorers, demonstrating robustness of the Shapley-based Balance metric to scorer choice.

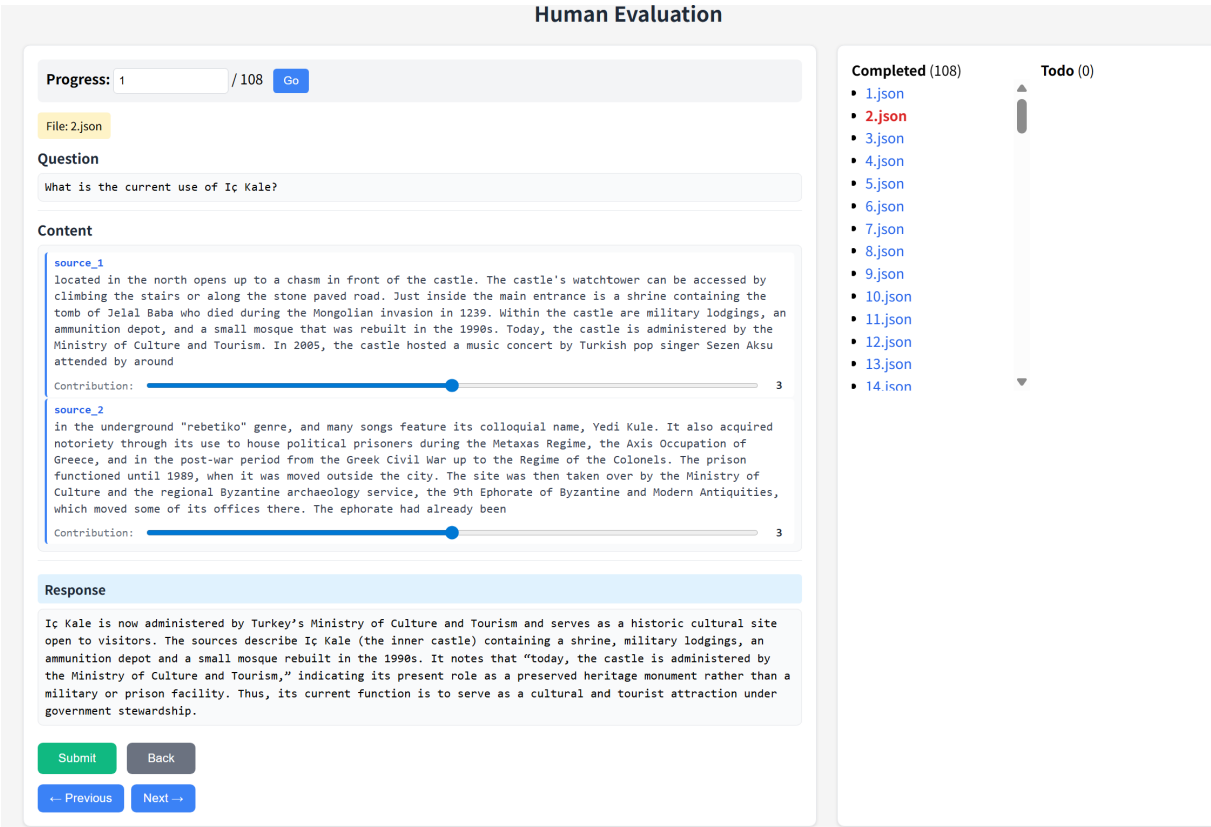

Figure 17: Annotation interface for rating evidence contributions on a 1–5 scale.

direction $d^{(l)}$ and the direction toward evidence i. A value of 1 means $c^{(l)}$ aligns perfectly with evidence i, 0 means orthogonality, and -1 means opposing alignment. Because cosine projections can be negative and do not sum to one, we treat $b_i^{(l)}$ as directional alignment strength rather than probability mass. Larger gaps (e.g., $b_1^{(l)} \gg b_2^{(l)}$) indicate stronger positional asymmetry.

Layer-wise computation. We compute $b_i^{(l)}$ across all layers $l \in \{1, \ldots, L\}$ to track how bias evolves through depth, allowing us to distinguish bias that emerges in lower layers from bias that accumulates gradually across the network.

D.7 Representation-Level Position Bias in GPT-OSS-20B

To assess whether representation-level positional asymmetry is specific to the Llama architecture, we apply the same directional bias attribution analysis to GPT-OSS-20B. The full methodology is provided in Appendix D.6. Figure 30 shows layer-wise directional projections for all six conflict types on GPT-OSS-20B. The geometric bias toward earlier evidence persists from the lowest to the highest layers, closely mirroring the pattern

Evidence Contribution Annotation Guidelines

Task Overview: Evaluate the contribution of each evidence source to the model's response for a given question. Rate each source on a 1–5 scale.

Rating Scale:
1 – Minimal Contribution: Evidence barely reflected in response; removing it would not affect core content.
2 – Minor Contribution: Some information mentioned but peripheral; plays supporting role.
3 – Moderate Contribution: Key information clearly reflected; important for response completeness.
4 – Major Contribution: Core content prominently featured; significant impact on main arguments.
5 – Dominant Contribution: Response primarily derived from this evidence; plays decisive role.

Core Principles:
1. Focus on relative distribution, not absolute values. Ratings should reflect *relative contribution differences* between sources.
Example A (Significant disparity): Source 1: 5, Source 2: 1 → Response heavily relies on Source 1.
Example B (Balanced): Source 1: 3, Source 2: 3 → Both sources contribute equally.
Example C (Low but balanced): Source 1: 1, Source 2: 1 → Response relies on model's own knowledge.

2. Base judgment on content matching:
– *Direct quotation:* Response uses verbatim or paraphrased content from evidence.
– *Information correspondence:* Facts, data, viewpoints in response traceable to evidence.
– *Structural influence:* Response organization influenced by evidence structure.

3. Evaluate each evidence independently, then use score distribution to reflect relative relationships.

Annotation Steps:
1. Read the **Question** to understand the core inquiry.
2. Read all **Sources** to understand the information each provides.
3. Read the **Response** carefully, analyzing its content and structure.
4. **Rate each source:** Identify which parts of the response come from each evidence, assess the weight of that information, and assign a 1–5 score.
5. **Check consistency:** Ensure score distribution reasonably reflects relative contributions.

Important Notes:
– *Avoid position bias:* Do not favor evidence appearing earlier.
– *Distinguish quality from contribution:* High-quality evidence unused in response should receive a low score.
– *Handle conflicting evidence:* Focus on which evidence's viewpoint dominates in the response.

Figure 18: Annotation guidelines for evaluating evidence contribution to model responses.

Model	Ambig.	Gran.	Persp.
Human Evaluation			
claude-4.5-sonnet	**21.18**	**16.53**	**12.85**
gpt-oss-120b	22.51	17.32	13.97
llama-3.1-8b	28.80	23.89	15.91
LLM-as-a-Judge			
claude-4.5-sonnet	**20.62**	**9.06**	**17.89**
gpt-oss-120b	24.66	13.15	18.30
llama-3.1-8b	30.99	21.02	21.44

Table 7: Balance scores (normalized Gini %, lower is better) from human evaluation and LLM-as-a-Judge on three models spanning diverse capability levels.

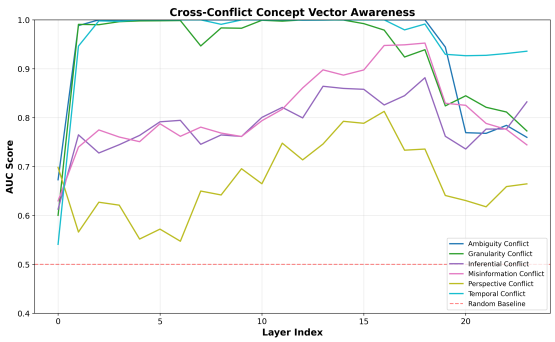

Figure 19: Layer-wise AUC for conflict awareness in GPT-OSS-20B. Despite fewer layers, the model demonstrates consistent awareness patterns across conflict types.

observed in Llama-3.1-8B-Instruct. As shown in Figure 3, this cross-layer tendency is already evident in the Llama model. This cross-model consistency suggests a shared pattern on tested models under our dataset: directional asymmetry in combined-evidence representations appears across both architectures, rather than being limited to a

Conflict type	$\lvert x_{\text{conf}} \rvert$	$\lvert x_{\text{cons}} \rvert$	Fold (train / val)
Misinformation	1004	1004	1606 / 402
Inferential	787	787	1259 / 315
Temporal	960	960	1536 / 384
Granularity	1020	1020	1632 / 408
Perspective	1010	1010	1616 / 404
Ambiguity	1000	1000	1600 / 400

Table 8: Conflict-consistent pair counts and 5-fold stratified cross-validation statistics per conflict type. Each conflict instance is paired with one consistent counterpart, so $|x_{\text{conf}}| = |x_{\text{cons}}|$ by construction; per-fold training and validation sizes are computed from the combined pool $|x_{\text{conf}}| + |x_{\text{cons}}|$ with an 80/20 stratified split.

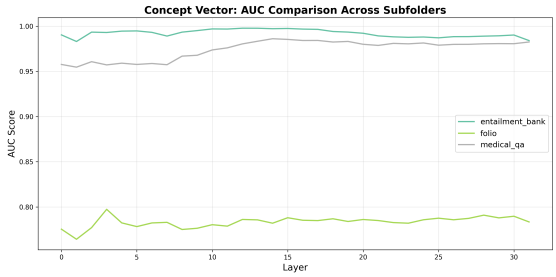

Figure 20: Concept vector AUC comparison for inferential conflicts across datasets. EntailmentBank (explicit) shows consistently higher AUC than FOLIO (implicit), indicating stronger awareness of surface-level logical conflicts.

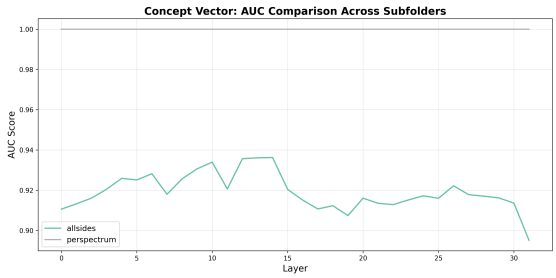

Figure 21: Concept vector AUC comparison for perspective conflicts across datasets. Perspectrum (explicit) achieves stable high AUC, while AllSides (implicit) fluctuates near chance level, indicating reduced sensitivity to subtle viewpoint differences.

Llama-specific artifact.

D.8 Position Shuffling Experiment

To test whether randomly reordering evidence input can mitigate position bias, we permute evidence order at inference time and re-measure evidence contribution distributions using our Shapley-based attribution framework. Tables 9 and 10 show that Evidence 1 continues to dominate after shuffling on both Llama-3.1-8B-Instruct and GPT-OSS-20B.

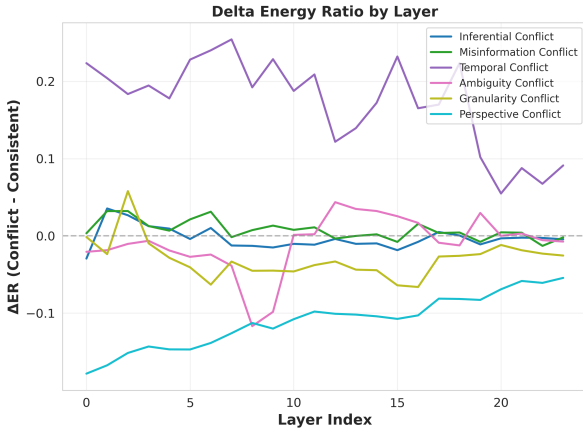

Figure 22: Delta energy ratio in `openai/gpt-oss-20b` across layers. $\Delta\text{ER} > 0$ indicates conflict states are more concentrated (low-rank); $\Delta\text{ER} < 0$ indicates more dispersed (high-rank).

This result suggests that position bias is mechanistically stable and is not fully resolved by simple input-ordering heuristics in our experiments.

Table 9: Evidence contribution distribution after position shuffling (Llama-3.1-8B-Instruct).

Conflict Type	Ev.1	Ev.2	Ev.3	Ev.4	Ev.5	Ev.6
Ambiguity	59.0%	28.9%	5.9%	6.1%	—	—
Granularity	35.3%	18.3%	14.0%	15.2%	12.6%	4.6%
Perspective	47.4%	27.2%	10.9%	5.0%	3.6%	5.9%

Table 10: Evidence contribution distribution after position shuffling (GPT-OSS-20B).

Conflict Type	Ev.1	Ev.2	Ev.3	Ev.4	Ev.5	Ev.6
Ambiguity	55.2%	32.2%	4.9%	7.5%	—	—
Granularity	27.4%	17.4%	16.4%	16.7%	15.5%	6.5%
Perspective	46.5%	22.8%	14.0%	5.7%	4.1%	6.9%

D.9 Steering Strength Sensitivity Analysis

To assess sensitivity to steering strength, we evaluate Llama-3.1-8B-Instruct across $\alpha \in \{0.5, 1.0, 1.5, 2.0\}$ on all six conflict categories. Figure 31 reports Balance scores (summarization tasks) and Accuracy (reasoning tasks) for each value of α. The curves remain stable across this range. Accuracy improves from $\alpha = 0.5$ to $\alpha = 1.5$ for all three reasoning conflict types under both steering variants, and changes only modestly at $\alpha = 2.0$. Temporal conflicts show the largest gain, especially under `first_generated` steering, while inferential and misinformation conflicts follow the same overall trend with smaller variation. These results indicate that the method is robust to

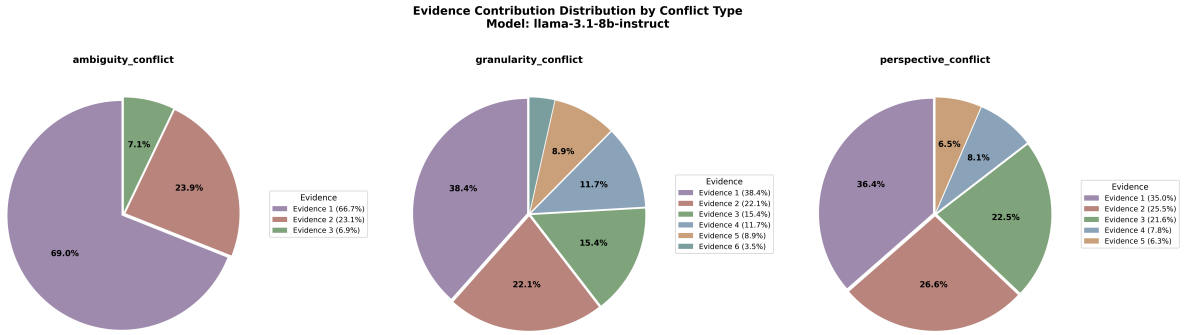

Figure 23: Evidence contribution distribution from output-level analysis. As described in Section 3.1, the distribution is highly non-uniform, with earlier evidence often dominating.

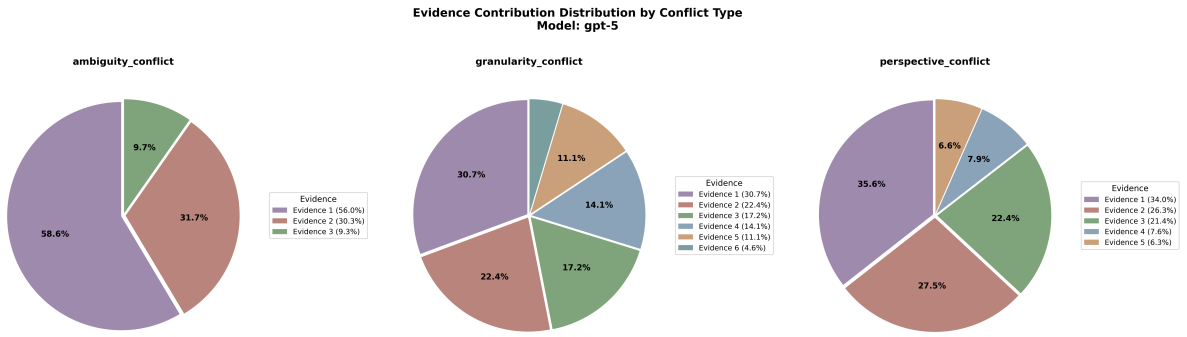

Figure 24: Evidence position bias (pie chart) for `gpt-5`.

the choice of α, with strong performance throughout the tested range and a reliable operating region around $\alpha \in [1.0, 1.5]$.

E Mitigation Method Details

E.1 Prompts

We use different prompts for summarization and reasoning tasks. Each prompt consists of a system message and a user message. All prompts are designed to be concise and task-appropriate.

E.1.1 Summarization Tasks

For summarization tasks (ambiguity, granularity, perspective conflicts), we use three prompt variants shown in Figures 32, 33, and 34.

E.1.2 Reasoning Tasks

For reasoning tasks (inferential, misinformation, temporal conflicts), we use prompts shown in Figures 35 and 36.

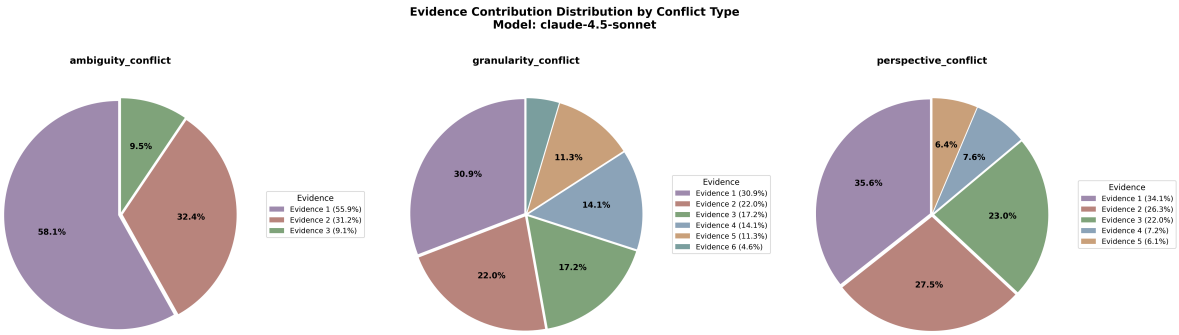

Figure 25: Evidence position bias (pie chart) for `claude-4.5-sonnet`.

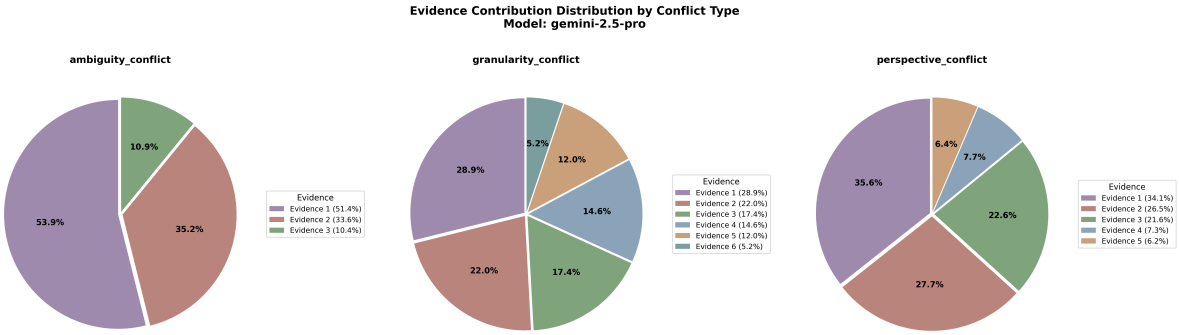

Figure 26: Evidence position bias (pie chart) for `gemini-2.5-pro`.

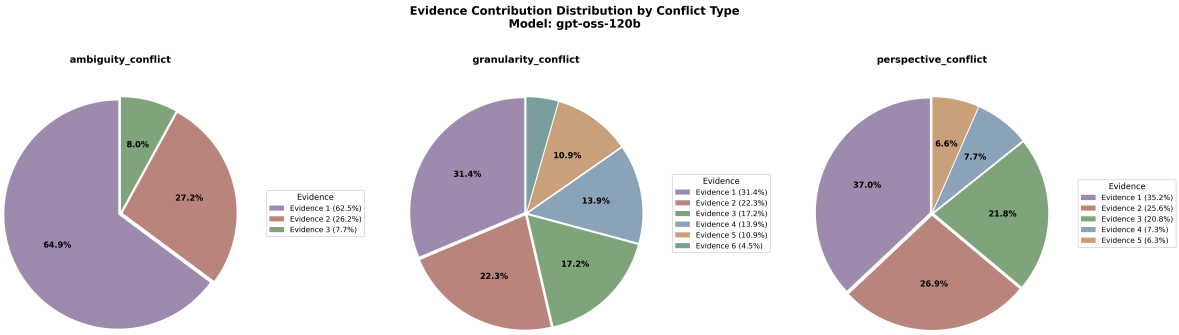

Figure 27: Evidence position bias (pie chart) for `gpt-oss-120b`.

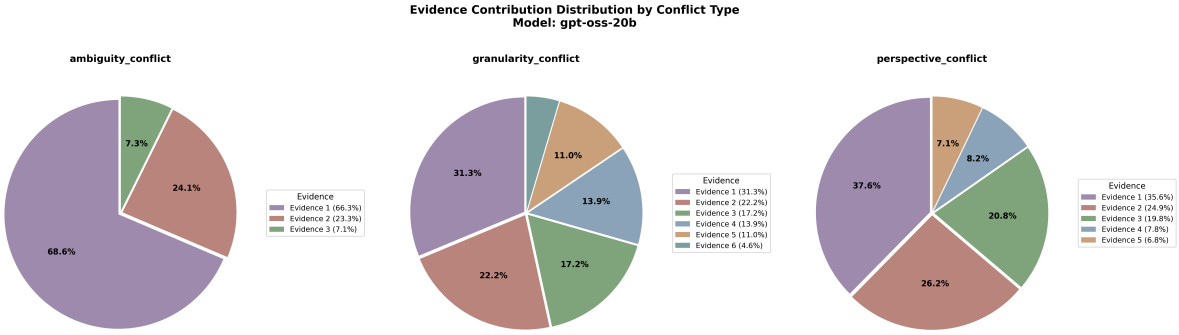

Figure 28: Evidence position bias (pie chart) for `gpt-oss-20b`.

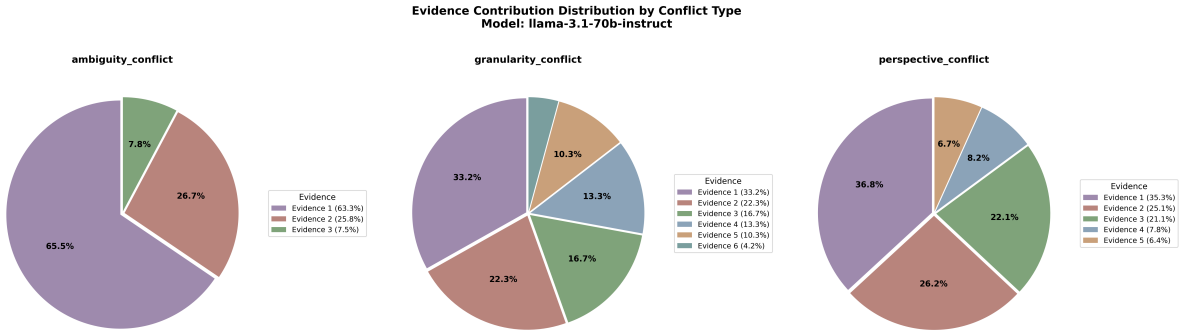

Figure 29: Evidence position bias (pie chart) for `llama-3.1-70b-instruct`.

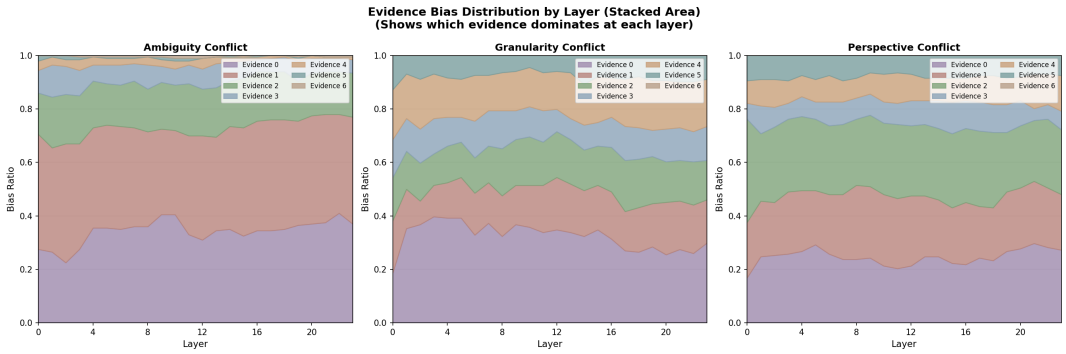

Figure 30: Layer-wise directional projection in GPT-OSS-20B via directional bias attribution. The combined representation consistently aligns with earlier evidence directions across layers and conflict types, replicating the pattern in Llama-3.1-8B and suggesting cross-model consistency of representation-level positional bias on tested models.

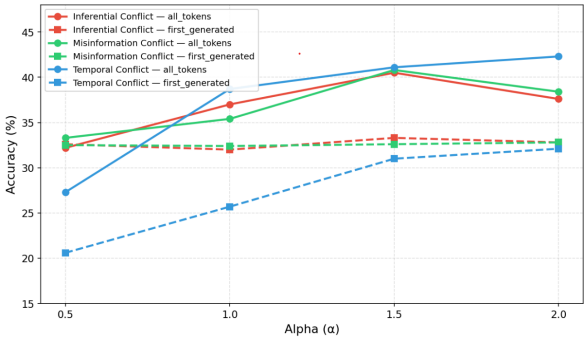

Figure 31: Sensitivity of activation steering to the coefficient α on Llama-3.1-8B-Instruct. Balance scores (lower is better) and Accuracy (higher is better) remain stable across a wide range of α, suggesting robustness to this hyperparameter within the tested range.

Simple Prompt for Summarization

System: You are a helpful assistant that answers questions based on provided evidence. Answer concisely in 2–3 sentences.

User: {evidence_text}
[Question]: {question}

Figure 32: Simple prompt for summarization tasks.

Prompt for Summarization

System: You are a neutral information synthesis assistant. Analyze all sources with equal importance and provide unbiased, balanced answers.

User: {evidence_text}
[Question]: {question}

Figure 33: Prompt for summarization tasks.

Single-Evidence Prompt for Summarization

System: You are a helpful assistant. Answer the question based ONLY on the provided evidence. Be concise and factual.

User: [Evidence]: {source_content}

[Question]: {question}

Figure 34: Single-evidence prompt for summarization tasks.

Single-Evidence Prompt for Reasoning

System: You are a helpful assistant that provides step-by-step reasoning.

User: [Evidence]: {source_content}

[Question]: {question}

Based on the evidence above, think step by step and provide your final answer inside `<answer></answer>` tags.

Figure 35: Single-evidence prompt for reasoning tasks.

Combined-Evidence Prompt for Reasoning

System: You are a helpful assistant that provides step-by-step reasoning.

User: [Evidence 1]: {source_1}

[Evidence 2]: {source_2}

...

[Question]: {question}

Analyze all evidence above and answer the question. Think step by step, then provide your final answer inside `<answer></answer>` tags.

Figure 36: Combined-evidence prompt for reasoning tasks.