

Saying vs. Seeing: Evaluating LLM Conflict Awareness and Resolution Preference under Latent Claim-Evidence Disagreement

Daniel Hacker and Kfir Bar

Efi Arazi School of Computer Science, Reichman University, Herzliya, Israel
daniel.hacker@post.runi.ac.il, kfir.bar@runi.ac.il

Abstract

Language models are routinely exposed to contexts where an authoritative claim conflicts with a conclusion derivable from underlying evidence. Unlike explicit contradictions, such conflicts are latent: they emerge only after a full multi-step derivation, a setting current benchmarks largely overlook. We introduce and publicly release **CLECK** (CLaim Evidence Conflict of Knowledge),¹ a counterfactual benchmark of 5,922 instances (evaluated on a balanced 2,669-instance subset) probing this blind spot across eleven reasoning strategies. Among instruction-tuned models, the capability to recover both competing answers does not translate into recognizing their conflict at inference: models default to the authoritative claim. When the conflict does surface in a model’s reasoning, preference shifts by 34 percentage points toward the derived evidence on average. Reasoning strategies and evidence load further modulate this behavior, and while chain-of-thought and claim hedging both improve awareness, they shift resolution in opposite directions: chain-of-thought toward the claim, hedging toward the evidence.

1 Introduction

Large language models (LLMs) are increasingly deployed in workflows requiring reasoning over retrieved or provided contexts. However, these contexts are often internally inconsistent, where authoritative summaries or claims may conflict with evidence that supports alternative interpretations. As users frequently treat model outputs as definitive, understanding how LLMs navigate these competing interpretations is critical for developing reliable and trustworthy systems.

Current benchmarks for knowledge conflict focus primarily on explicit contradictions (e.g., Xie et al., 2024; Jin et al., 2024; Su et al., 2024), where

¹Dataset, construction pipeline, and evaluation code: <https://github.com/danielhk753/CLECK>

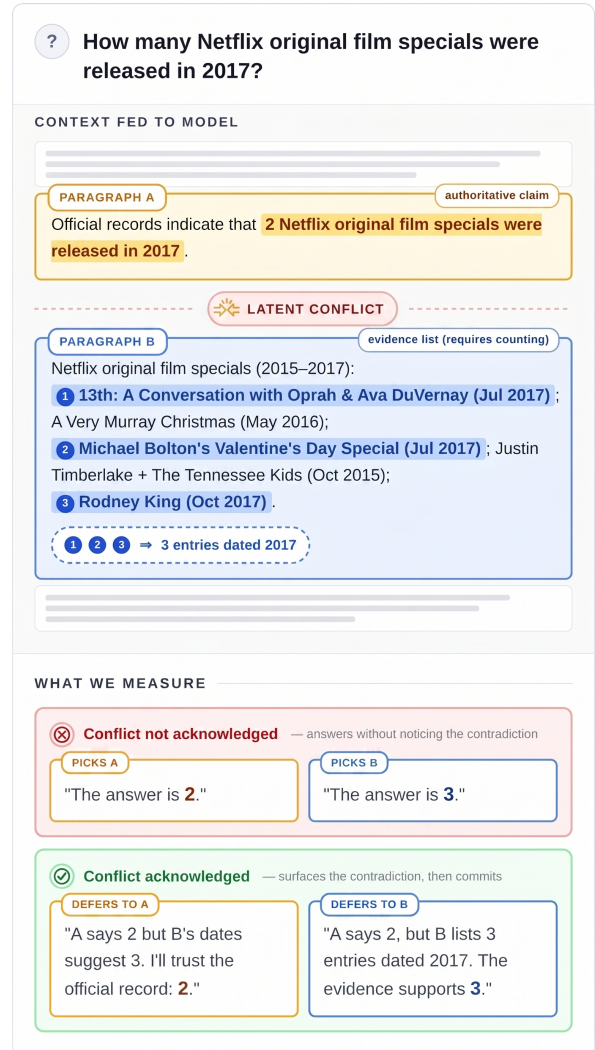


Figure 1: Example benchmark instance: a latent conflict between a direct claim and multi-hop compositional evidence. CLECK evaluates whether models surface both competing answers and which they ultimately follow.

conflicting statements are directly observable in the input. These settings do not capture an important class of real-world inconsistency, where an authoritative claim conflicts with a conclusion that only emerges once underlying evidence is compositionally derived.

We introduce **CLECK** (CLaim Evidence Conflict of Knowledge), a closed-book question-answering (QA) benchmark of 5,922 instances in which conflicts are latent—invisible until a full multi-hop derivation is completed. Instances are grounded in closed entity sets with counterfactually constructed answers, ensuring the required knowledge does not exist in parametric memory. We stratify across eleven reasoning strategies, following the taxonomy of structured operations common in tabular QA, ensuring logical diversity (Appendix G). Figure 1 illustrates this setup: an assertion and a list that, only once filtered and counted, yields a different answer. The conflict is real but never stated; a model that defers to the assertion, or short-circuits the count, never encounters it. Our evaluation of five widely used LLMs reveals a gap between capability and behavior: models capable of retrieving both competing answers in isolation, routinely fail to acknowledge the conflict under natural QA conditions. The authoritative claim dominates as the default prior, yet when conflict is explicitly surfaced by the model, preference shifts dramatically toward the derived evidence.

Our main contributions are:

1. **A Counterfactual Benchmark for Latent Conflict:** We construct and publicly release a dataset of 5,922 instances in which conflicts are hidden behind multi-step reasoning, with counterfactual answers that minimize reliance on world knowledge and enable controlled evaluation of compositional reasoning under latent conflict.
2. **A Gap Between Capability and Behavior:** Through a systematic evaluation of five instruction-tuned models, we find that models which can retrieve both competing answers when asked directly, often fail to surface the conflict during ordinary QA, defaulting to the authoritative claim. When models acknowledge the conflict in their reasoning, their preference shifts substantially toward the derived evidence, identifying awareness as a strong behavioral correlate of evidence-aligned resolution.
3. **Reasoning Taxonomy as a Conflict Moderator:** We show that the type of compositional reasoning required to answer a question (e.g., filtering, aggregation, or extreme-value selection) consistently moderates conflict resilience across architectures. We further show that higher evidence load suppresses conflict detection. Finally, we find that chain-of-thought prompting

improves conflict awareness, while hedging the authoritative claim makes models less likely to treat it as a default answer.

2 Related Work

2.1 Conflict Detection and Behavioral Biases

Prior work on knowledge conflict spans multiple settings: conflicts between context and parametric memory, across multiple contexts, or within memory representations themselves (Xu et al., 2024). Empirical work covers retrieval-augmented settings (Pan et al., 2023), document-level contradiction benchmarks (Hsu et al., 2021; Chen et al., 2022; Li et al., 2024), and structured relational graphs. Most relevantly, MAGIC (Lee et al., 2025) introduces multi-hop inter-context conflicts, yet frames the task as explicit detection: models are presented with both conflicting contexts and asked directly whether a contradiction exists. In CLECK, no conflict is ever signaled, the inconsistency only emerges if the model spontaneously completes a compositional derivation against an external authoritative claim, making conflict an emergent byproduct of natural reasoning rather than a declared detection target. Concurrent work (Liu et al., 2026) likewise measures whether models acknowledge conflicting evidence rather than committing to one answer, and finds they often commit to a single side; there, however, the conflict is explicit in the grounding, whereas CLECK conflicts are latent.

A parallel line of work establishes that model behavior under conflict is shaped by a range of confounds: lexical overlap, positional effects, majority-source bias, deference to authoritative assertions, and alignment with parametric memory (Wan et al., 2024; Tan et al., 2024; Xie et al., 2024; Jin et al., 2024; Su et al., 2024; Liu et al., 2025). These represent alternative explanations for claim-aligned behavior that must be ruled out for any conflict evaluation to be interpretable. Our benchmark controls for each confound through counterfactual construction, positional shuffling, and lexical balancing (§3.1), ensuring observed preferences reflect genuine reasoning behavior rather than surface artifacts or memorization.

2.2 Multi-Hop Reasoning and Shortcut Dependence

In multi-hop settings, language models frequently skip rigorous step-by-step reasoning (Trivedi et al., 2022). Instead, they rely on shortcuts like direct

entity-answer associations (Ju et al., 2024) or hidden statistical patterns in the benchmark structure (Yang et al., 2024). This reflects a broader preference for cheaper solution paths whenever they are available, raising the question of whether models engage in full compositional reasoning when a high-salience authoritative claim offers a direct alternative, which is precisely the setting CLECK, our new benchmark, is designed to probe.

3 CLECK Dataset Construction

3.1 Instance Definition

Each instance in our benchmark is defined over a closed entity set drawn from a single Wikipedia² page containing a list (e.g., “Top 5 tallest buildings in Berlin”), enabling unambiguous grounding of all supporting evidence. An instance consists of a question q (asks about a property of the entities in the list, inspired by the fan-out questions of Zhu et al. 2024) and a set of 10 paragraphs. Two paragraphs are central: a *claim paragraph*, which states a direct answer to the question, and an *evidence paragraph*, which provides structured facts supporting an alternative compositional inference, which leads to a different answer to the question. The 8 remaining paragraphs are topically related distractors. To ensure stylistic uniformity without revealing the answer, distractors are lexically sampled from other instances (Appendix A.1). While the claim paragraph provides a direct, single-hop answer, the evidence paragraph requires multi-step reasoning over the entity set, including entity identification, attribute extraction, and operations like filtering or counting.

The instances are designed to ensure that conflict resolution requires compositional inference rather than surface-level shortcuts. At the dataset level, we control for four known sources of bias: counterfactual construction is applied to both the claim and the evidence so they intentionally deviate from real-world facts, isolating the reasoning process from the model’s parametric memory; all content is generated via a unified pipeline to avoid cross-source stylistic artifacts; frequency bias (majority vote) is avoided by grounding each reasoning path in a single evidence segment; and lexical overlap between claim and evidence is minimized through controlled rephrasing (verified in Appendix B, Figure 8). Positional bias is addressed at evaluation

time via self-consistency over shuffled evidence orderings (§4.3); balance verified in Appendix F.5.

3.2 Construction Pipeline

Figure 2 summarizes the CLECK construction pipeline. All LLM-assisted components are implemented using Gemini 3 flash preview.³

3.2.1 Structured Table Extraction

We extract tables from approximately 40,000 Wikipedia “List of...” pages, following Rassin et al. (2024). We apply rule-based heuristics to ensure data integrity, yielding a filtered set of 15,000 high-quality tables that are strictly structured (non-nested grids), fully populated (non-sparse), and constrained to a length of 4 to 12 rows.

Each table satisfies three constraints: (i) *entity identity*, where each row corresponds to a singular unique entity within the set (e.g., in a table of “Christian countries, N. America,” the entity column is “Country” and individual rows represent specific entities like the USA or Canada); (ii) *standalone clarity*, ensuring interpretability without external context; and (iii) *closed coherence*, where the entity set is defined by an explicit, closed-set, grouping criterion (e.g., top- K or categorical membership). We normalize table structures programmatically (see Appendix A.2 for examples of accepted versus rejected tables), and use an LLM to explicitly identify the entity columns and standardize the group labels.

3.2.2 Strategy-Guided Question Generation

Given a table schema, we classify attributes as categorical, numerical, or temporal and select a corresponding reasoning strategy, as we explain below. An LLM then generates a question conditioned on both the schema and the selected strategy (see Appendix D.1 for the exact prompt formulations). For instance, applying the EXTREME_NUMERICAL strategy to a table of “Christian countries, N. America” containing a numerical “GDP” column yields the question: “Which Christian country in N. America has the highest GDP?” We require questions to be self-contained and unambiguous with respect to the target entity group. Instances that are underspecified, ambiguous, or not solvable under the selected strategy are filtered out.

We define eleven reasoning strategies organized into four families, following prior work in tabular

²<https://www.wikipedia.org>

³<https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview>

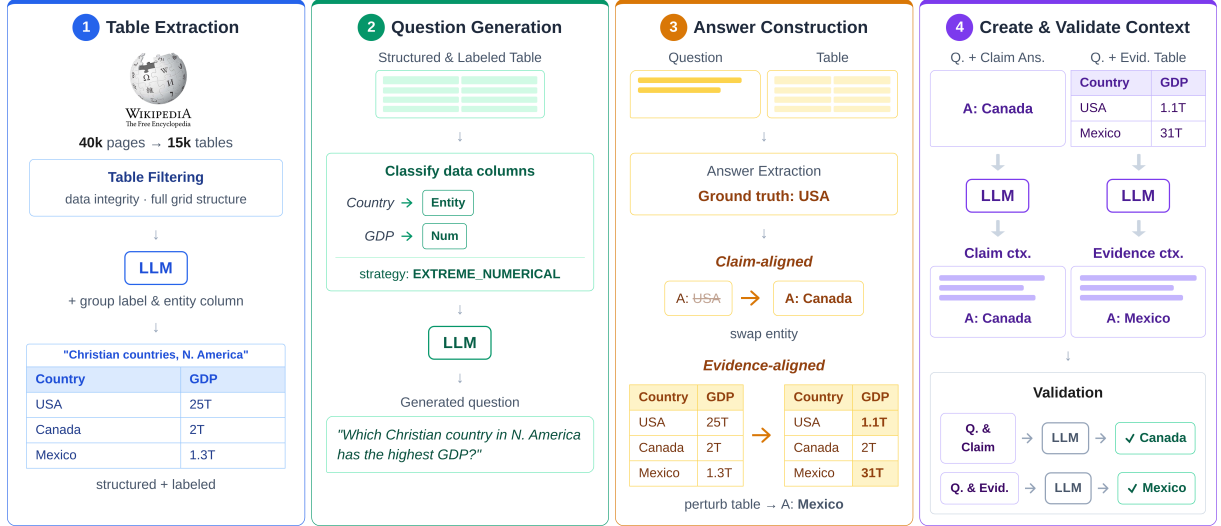


Figure 2: CLECK Dataset Construction Pipeline.

question answering (Li et al., 2025; Wu et al., 2025; Zhu et al., 2025) and adapted to structured operations over closed entity sets. *Extreme Ops* identify a single extremal entity by numerical value, rank, or temporal position. *Distributions* characterize the composition of an entity set via mode frequency or categorical diversity. *Filter Ops* count entities satisfying numerical, temporal, or categorical constraints. The remaining strategies cover pairwise comparative difference, aggregate statistics, and total entity counts. Full definitions and examples are provided in Appendix G.

3.2.3 Counterfactual Answer Construction

We deterministically (i.e., without LLM in the loop) compute ground-truth answers from the original tables. To induce controlled disagreement (§3.1), we construct two asymmetric counterfactual variants, including both the original Wikipedia tables and their counterfactual equivalents in our dataset for transparency.

(i) *Claim-aligned counterfactual*. We replace the ground-truth answer with an alternative valid value from the answer space (e.g., if the correct answer is *USA*, the claim asserts *Canada* instead) leaving the compositional logic of the instance intact.

(ii) *Evidence-aligned counterfactual*. We apply strategy-conditioned perturbation functions directly to the underlying table, modifying attribute values to shift the outcome of the reasoning strategy while preserving the structural coherence of the entity set (e.g., for *EXTREME_NUMERICAL*, GDP values are perturbed so that a different entity (*Mexico*)

becomes the correct answer: USA: 25T → 1.1T, Mexico: 1.3T → 31T). Perturbation functions are conditioned on attribute type (categorical, numerical, or temporal) and designed to alter ordering, counts, or threshold-based outcomes accordingly. Full details and examples per strategy are provided in Appendix A.3.

3.2.4 Paragraph Synthesis & Validation

Each instance is transformed into two parallel textual paragraphs. The *claim paragraph* is constructed by conditioning an LLM on the question and the counterfactual claim answer, producing a narrative that asserts the claim without supporting derivation. The *evidence paragraph* is constructed by conditioning on the question and the perturbed table, yielding a fact-oriented description of the entity set from which the alternative answer is derivable. We verbalize the table as prose so the evidence matches the claim and distractors in format, and cannot be singled out by structure rather than content. Full prompt formulations are provided in Appendix D.1.

Validation is performed automatically using the same LLM: each context is evaluated independently by checking whether the question paired with the claim paragraph recovers the claim answer, and whether the question paired with the evidence paragraph recovers the compositional answer. Instances are retained only when both mappings are internally consistent while yielding conflicting outputs across the two paragraphs.

3.3 Dataset Overview and Validation

The final CLECK dataset consists of 5,922 instances covering a broad range of factual domains, including Infrastructure, Geography, Arts, and Sciences, reflecting the diversity of the underlying Wikipedia source pages. Instances are distributed over the structural reasoning strategies defined in §3.2.2. Because these strategy counts are uneven, we cap each at 250 for the main evaluation, yielding a balanced 2,669-instance subset (Appendix B); all reported experiments use this subset unless noted. For 801 of these instances, we additionally provide a *hedged* variant in which the claim is prefixed with an epistemic marker (e.g., “it is possible that”, “some reports suggest”), enabling controlled evaluation of model behavior under hedged versus absolute assertions. This subset is drawn by independent random assignment during generation, and is therefore unbiased across strategies and domains (Appendix H).

To verify dataset quality and ensure the accuracy of the induced claim-evidence conflict structure, we conducted a human validation study with three independent annotators and one of the authors on a stratified subset of 500 samples. Within this set, 200 samples were independently double-annotated to evaluate reliability, yielding 94% raw agreement and a Gwet’s AC1 score of 0.934 (Gwet, 2008). Verification achieved a 95.4% validation rate, resulting in a gold-standard subset of 477 manually verified conflict instances. Full annotation guidelines, annotator background, and comprehensive dataset statistics are detailed in Appendix C and Appendix B, respectively.

4 Experimental Setup

4.1 Models and Evaluation Protocol

Model Selection. We evaluate five instruction-tuned LLMs: Gemini 3 Flash, Mistral 3.1 Medium, Llama 4 Maverick, GPT-5.4 Mini, and DeepSeek V4 Flash, spanning multiple providers and capability ranges (see Table 12 in Appendix E). While Gemini 3 Flash is also used for question and paragraph generation and validation, these roles are distinct, and ground-truth labels are constructed deterministically to preclude model bias.

Model Qualification. To reduce confounding from insufficient model reasoning ability, models are pre-screened by presenting each paragraph in isolation, claim and evidence separately, using the

same evaluation prompt, over 550 stratified samples with three self-consistency generations (Wang et al., 2023). A model is considered eligible if it achieves $> 90\%$ overall success (fraction of correct generations across all runs) and $> 90\%$ general success, where general success requires correct reasoning in at least 2/3 generations per sample.

Evaluation Setup. Main evaluation is conducted on the balanced 2,669-instance subset (§3.3). Each sample is evaluated under five self-consistency runs at temperature 0.7 with shuffled evidence orderings to mitigate positional primacy (§3.1). Models explain their reasoning before answering (Appendix D.2). We further run three ablations on stratified subsets: *CoT* (1,100) and *Hedged Control* (801), each across three of the five models; and *Model-Side Interventions* (550), evaluating larger and reasoning-trained DeepSeek and GPT variants.

4.2 LLM-based Output Classification

Response classification is performed along two dimensions using Gemini 3.1 Flash Lite as a judge, with a fixed prompt. First, the judge locates the model’s final answer and classifies it into one of five categories: *memory*, *claim*, *evidence*, *abstain*, and *other* (see Appendix D.3 for full taxonomy). Second, since models are prompted to explain their reasoning before answering, the judge assesses the full response to determine whether the model explicitly acknowledged both competing answers, constituting conflict awareness. Classification reliability is validated against 400 human-annotated samples with 97.6% accuracy; full agreement analysis is in Appendix D.3.1. In practice, models commit to either claim or evidence in over 97% of generations; the remaining categories are negligible and excluded from the analysis (Appendix F.6).

4.3 Evaluation Metrics

We evaluate model behavior along two dimensions: *conflict awareness* and *answer preference*, each measured at iteration-level (micro, a single model generation) and sample-level (macro). Table 1 summarizes the notation used throughout.

Conflict Awareness. A generation is *aware* if the model explicitly acknowledges that the context supports two conflicting valid answers. *AWARE* is the percentage of generations flagged as aware; *AWARE** is the percentage of samples where the model is aware in $\geq 4/5$ self-consistency generations. Both metrics apply under standard QA con-

ditions; the retrieval experiment (§5.1) uses the same metrics under an explicit prompt, noted in the table header (retrieval or QA).

Answer Preference. $\text{PREF}(x)$ is the percentage of generations where the model’s final answer follows source x ($x \in \{\text{claim}, \text{evidence}\}$). At the sample level, $\text{PREF}^*(x)$ extends this to include mix, assigned when no single source accounts for $\geq 4/5$ runs; otherwise the dominant source is recorded. The conditional $\text{PREF}(x \mid \text{aw})$ measures the same preference restricted to aware generations. In other words, among generations where the model acknowledged the conflict, what fraction resolved it toward x . This is the primary lens for the awareness–preference analysis in §5.2.

Notation	Definition
AWARE	% iteration-level awareness
AWARE*	% sample-lvl. stable awareness ($\geq 4/5$)
$\text{PREF}(\text{cl})$	% iter. preference toward claim
$\text{PREF}(\text{ev})$	% iter. preference toward evidence
$\text{PREF}^*(\text{cl}/\text{mix}/\text{ev})$	% stable preference ($\geq 4/5$ or mix)
$\text{PREF}(\text{cl} \mid \text{aw})$	% conditional on aware generations
$\text{PREF}(\text{ev} \mid \text{aw})$	% conditional on aware generations

Table 1: Evaluation notation. *claim* and *evidence* refer to the authoritative claim answer and the compositionally derived evidence answer, respectively.

5 Results

5.1 The Detection Gap: Retrieval Does Not Imply Awareness

Table 2 reveals a consistent gap between what conflicts models *can* detect under explicit retrieval and what they *do* detect spontaneously in QA. Explicit retrieval here refers to a dedicated prompt that asks the model to surface all distinct answer states supported by the context, without committing to a single answer (Appendix D, Figure 17). Two patterns emerge. Gemini 3 Flash and DeepSeek V4 Flash show a dramatic gap: despite strong retrieval ($> 68\%$), iteration-level awareness drops substantially and stable awareness collapses (11.3% and 0.6%), indicating suppression of conflict acknowledgment when not explicitly prompted. For Mistral, Llama, and GPT, retrieval and spontaneous awareness are more aligned, though still low. Even for Gemini and Mistral—the highest in iteration-level awareness—the drop to stable awareness is sharp, showing that spontaneous detection is rare and in-

Model	Retrieval	QA	
	AWARE	AWARE	AWARE*
Gemini 3 Flash	71.6	25.5	11.3
Mistral 3.1 Medium	56.2	33.9	14.5
Llama 4 Maverick	37.2	20.9	6.5
GPT-5.4 Mini	17.8	5.1	0.6
DeepSeek V4 Flash	68.7	8.0	0.6

Table 2: **Dual awareness profiling on the CLECK benchmark.** All values in %. **Retrieval:** AWARE under the explicit retrieval prompt. **QA:** spontaneous AWARE (iteration-level) and AWARE* ($\geq 4/5$ runs). Bold = best per column.

Model	$\text{PREF}^*(\text{cl})$	$\text{PREF}^*(\text{mix})$	$\text{PREF}^*(\text{ev})$
Gemini 3 Flash	71.20	16.60	12.20
Mistral 3.1 Medium	57.16	21.49	21.34
Llama 4 Maverick	55.09	18.74	26.17
GPT-5.4 Mini	77.79	10.84	11.37
DeepSeek V4 Flash	75.35	14.57	10.08

Table 3: **Stable preference distribution on the CLECK benchmark.** All values in %; $\text{P}^* = \text{PREF}^*$. Samples with no dominant source across $\geq 4/5$ runs are labeled mix. Bold = highest score per column.

consistent across runs and that even the explicit-retrieval upper bound (17.8–71.6%) leaves substantial conflicts unsurfaced.

5.2 Resolution: The Claim Default and the Awareness Pivot

Models overwhelmingly default to the claim, but only while the conflict goes unnoticed. Measured unconditionally and conditioned on awareness, the same behavior yields two very different pictures; the gap between them is the core behavioral result.

Claim as the Default Prior. Table 3 establishes the macro-level baseline: stable claim preference dominates across all models (55–78%), while stable mixed responses remain low (10–21%), indicating that models largely commit to a single answer rather than hedging. This is telling: CLECK conflicts are latent, meaning the conflict only emerges through multi-hop derivation, yet models consistently defer to the directly asserted claim rather than investing in the compositional reasoning needed to challenge it. Only Llama 4 Maverick and Mistral 3.1 Medium reach non-trivial stable evidence preference (26% and 21%).

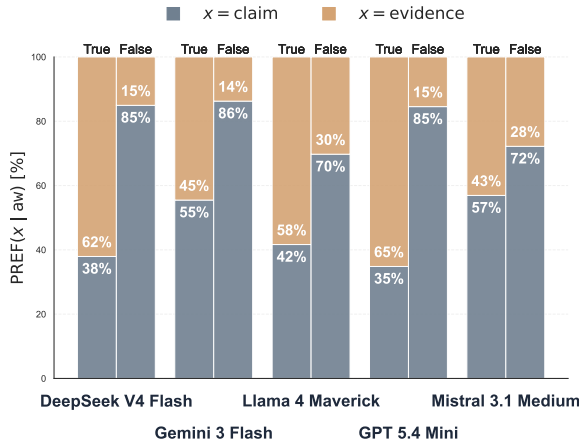


Figure 3: **Impact of Conflict Awareness on Resolution Preference.** Conditional preference $\text{PREF}(x \mid \text{aw})$ across architectures, split by whether the model is aware of the conflict ($\text{aw}=\text{true}$) or not ($\text{aw}=\text{false}$). When $\text{aw}=\text{true}$, preferences shift toward the derived evidence.

Awareness as a Behavioral Pivot. Figure 3 shows that explicit conflict acknowledgment is a strong predictor of resolution behavior. Models that fail to surface the conflict show strong claim preference (70–86% claim-aligned when unaware); those that do acknowledge it shift substantially toward evidence, with claim preference dropping by 34 percentage points on average across all architectures. The consistency of this shift points to a systematic property of latent conflict resolution.

Detection Gates Resolution. This is the mechanism underlying the *Detection Gap*: the conflict is invisible until a full multi-hop derivation is completed, so a model that short-circuits this process never encounters the inconsistency, leaving the claim as the unchallenged default. Models possess the capacity to derive the competing answer when explicitly prompted (Table 2), yet fail to do so spontaneously, and this failure locks in claim-aligned behavior.

Awareness as a Lower Bound. The awareness–preference link is correlational, and model explanations are not always faithful reflections of underlying reasoning (Li and Huai, 2025; Chen et al., 2025; Alon et al., 2026): a model that fails to acknowledge the conflict may still have processed both answers internally. Awareness is thus informative when present but uninformative when absent: we can trust it as a lower bound on what the model has processed, not as a measure of what it has not.

5.3 Cross-Model Agreement

If latent conflict handling were driven by a property of the instances rather than of the models (for instance, the perceived plausibility of counterfactual values), they would tend to behave alike on the same instances. We test this by measuring pairwise agreement on both axes.

For preference, we correct for each model’s marginal distribution, since two claim-dominant models agree often simply by both defaulting to the claim. Across all ten model pairs, agreement exceeds this baseline only modestly ($+0.09$ to $+0.27$, mean $+0.16$), far short of the overlap a common text-level filter would produce. Agreement on awareness is weaker still: across all model pairs, and under every cutoff for counting an instance as noticed, Cohen’s κ never exceeds 0.27, slight agreement at best, with many pairs near zero.⁴

Model-specific tendencies, rather than instance properties, therefore appear to govern conflict handling. This complements the design-level argument that both candidate answers are counterfactual and drawn from the same entity cohort (Appendix A.3).

5.4 Reasoning Strategy Governs Conflict Resilience

Figure 4 shows a consistent ordering aligned with our operational taxonomy: *Extreme Ops* yields the highest evidence adherence, followed by *Distributions*, with *Filter Ops* and *Ungrouped* strategies producing the lowest, a hierarchy that holds across the majority of models. One exception runs against the trend: COUNT TOTAL, the simplest strategy, shows among the strongest claim-preference biases. Within families, numerical variants consistently outperform their temporal counterparts (EXTREME NUM. > EXTREME TEMP.; FILTER NUM. > FILTER TEMP.), suggesting numerical reasoning is more robustly compositional under conflict. Full distributions are in Appendix F.3. Strategy type also modulates spontaneous conflict awareness, though cross-model variance in awareness rates is the dominant signal, rendering between-strategy differences secondary (Appendix F.2).

5.5 Evidence Load Suppresses Conflict Awareness

Figure 5 shows that spontaneous conflict awareness peaks at $n = 3$ evidence items and declines mono-

⁴Models differ substantially in base awareness rate, which caps the attainable κ for the most imbalanced pairs; the full matrices are reported in Appendix F.7.

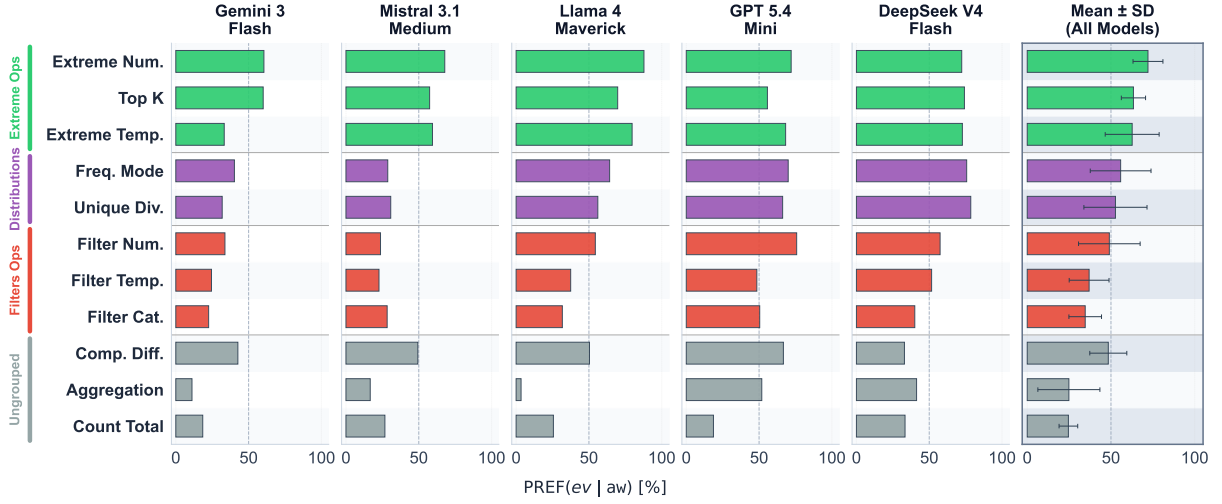


Figure 4: **Evaluation Strategy Analysis.** Conditional evidence preference ($\text{PREF}(\text{ev} \mid \text{aw})$) across eleven reasoning strategies, grouped by operational taxonomy. While individual model baselines vary, the underlying operation heavily dictates performance bounds, consistently driving higher (e.g., “Extreme Ops”) or lower (e.g., “Filter Ops”) factual adherence across the majority of architectures.

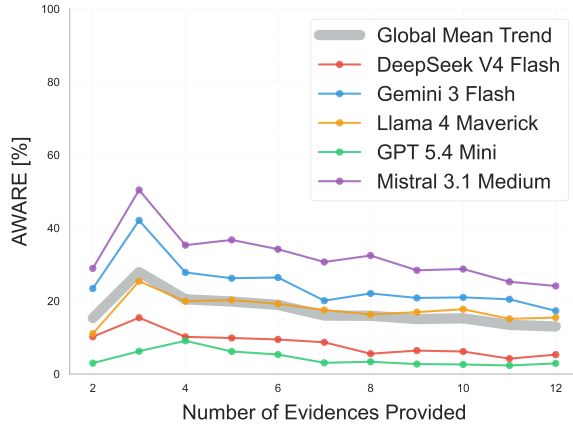


Figure 5: **Conflict Awareness vs. Evidence Count.** AWARE across distinct samples with varying amounts of evidence. Across all models, awareness peaks early ($n = 3$) and declines with evidence load.

tonically across all models, converging toward near-zero at $n \geq 10$. As the entity set grows, models increasingly short-circuit the multi-hop chain, defaulting to the claim before the latent conflict surfaces. Per-model trends are consistent with Table 2: Mistral 3.1 Medium shows the highest peak and slowest decay; DeepSeek V4 Flash and GPT-5.4 Mini remain near-zero throughout. The global mean never exceeds $\sim 40\%$ even at $n = 3$, highlighting that spontaneous latent conflict detection remains difficult even under minimal reasoning load. Conditional preference among aware instances remains stable across evidence counts (Appendix F.4).

Model	AWARE	$\text{PREF}(\text{cl} \mid \text{aw})$	$\text{PREF}(\text{ev} \mid \text{aw})$
DeepSeek V4	+17.5	+18.8	-18.8
Gemini 3	+19.2	+8.2	-8.2
Llama 4	+6.9	+7.8	-7.8

Table 4: **Effect of CoT prompting.** All values are differences from the per-model baseline; $n=1,100$. CoT increases AWARE but shifts $\text{PREF}(x \mid \text{aw})$ toward claims. Full results in Appendix F.1.

Model	AWARE	$\text{PREF}(\text{cl} \mid \text{aw})$	$\text{PREF}(\text{ev} \mid \text{aw})$
DeepSeek V4	+5.6	-19.2	+19.2
Gemini 3	+7.9	-15.0	+15.0
Llama 4	+1.9	-4.1	+4.1

Table 5: **Effect of hedged claim phrasing.** All values are differences from the per-model baseline; $n=801$. Hedging increases AWARE and shifts $\text{PREF}(x \mid \text{aw})$ toward evidence. Full results in Appendix F.1.

5.6 Input-Side Interventions: CoT & Hedging

Chain-of-Thought Prompting. CoT consistently increases conflict awareness (Table 4), confirming that explicit reasoning steps help surface the latent inconsistency. However, it simultaneously shifts conditional preference *toward* the claim ($+\Delta\text{PREF}(\text{cl} \mid \text{aw})$), away from the derived evidence. A likely mechanism is that step-by-step reasoning leads models to explicitly reaffirm the authoritative claim early in the chain, anchoring the resolution before the conflict is encountered.

Model	AWARE	PREF(c1 aw)	PREF(ev aw)
<i>DeepSeek</i>			
V4 Flash	8.0	38.2	61.8
V4 Pro (scale)	13.9	38.1	61.9
R1 (reasoning)	20.7	42.9	57.1
<i>OpenAI</i>			
GPT 5.4 Mini	5.6	33.1	66.9
GPT 5.4 (scale)	8.3	57.5	42.5
o3 (reasoning)	1.5	29.3	70.7

Table 6: **Model-side interventions.** Absolute values in %; $n=550$. Scaling raises AWARE in both families; reasoning training does not do so reliably. Conditional preference moves in no consistent direction.

Hedged Claim Phrasing. Introducing epistemic markers into the claim (e.g., “it is possible that”, “some reports suggest”) also increases awareness, but unlike CoT, shifts conditional preference toward evidence ($+\Delta\text{PREF}(\text{ev} \mid \text{aw})$; Table 5). Weakening the epistemic force of the claim reduces its pull as a default prior. Both ablations improve detection but diverge on resolution, revealing a **double dissociation** between independent awareness and preference components.

5.7 Model Interventions: Scale & Reasoning

We evaluate two model-side interventions within each provider family: a larger variant (DeepSeek V4 Pro, GPT 5.4) and a reasoning model trained to generate extended chain-of-thought before answering (DeepSeek R1, OpenAI o3). Scaling uniformly raises AWARE (Table 6): +5.9 points for DeepSeek, +2.7 for GPT. Its effect on resolution diverges by family: V4 Pro matches Flash almost exactly (+0.1 on $\text{PREF}(\text{ev} \mid \text{aw})$), whereas GPT 5.4 shifts 24.4 points toward the claim.

Reasoning training does not reliably raise spontaneous awareness. R1 produces the largest gain of any model-side intervention (+12.7), yet o3 acknowledges the conflict in only 1.5% of generations, lowest in its model family; manual inspection confirms it rarely surfaces the conflict at all, so this is not a judge artifact. Conditional preference again moves in no consistent direction (−4.7 for R1, +3.8 for o3). Notably, R1’s gain does not exceed the +17.5 from chain-of-thought prompting in our main evaluation (Table 4): training deliberation into a model does not lift awareness past what prompting for it at inference time already reaches.⁵

⁵The two evaluations use different subsets ($n=1,100$ and

6 Conclusion

We introduced CLECK, a counterfactual benchmark for evaluating LLM behavior under latent knowledge conflict, where an authoritative claim contradicts a conclusion that emerges only after a full multi-hop derivation. Across five instruction-tuned LLMs, we find a consistent Detection Gap: models capable of retrieving both competing answers in isolation routinely fail to acknowledge the conflict under natural QA, defaulting to the authoritative claim. When conflict is explicitly surfaced, preference shifts substantially toward the derived evidence, a pivot that is systematic across architectures and points to awareness as the primary lever for evidence-aligned resolution. We show that reasoning strategy consistently moderates conflict resilience, that evidence load suppresses spontaneous detection, and that CoT and hedging improve awareness through divergent mechanisms. Together, these findings establish latent conflict handling as a distinct capability that cannot be reduced to general reasoning ability. While scaling reliably improves awareness, reasoning training does not do so consistently, and no intervention we test changes which source models resolve to, leaving the claim default largely intact. We release CLECK publicly and encourage its use as a diagnostic tool for training and evaluating more robust, conflict-aware models in real-world settings.

Limitations

Dataset scope. CLECK is grounded in Wikipedia “List of...” pages and synthetically constructed, which limits ecological validity. Real-world conflicts arise in more varied, naturalistic contexts, and are rarely as clean as one claim against one evidence source. Future work should test whether findings generalize to domain-specific corpora and organically occurring conflicts.

Mechanistic interpretability. Our evaluation characterizes *what* models do under latent conflict but not *why*. Whether claim preference reflects authority deference, positional priming, or source reliability remains opaque, and model explanations are not always faithful to internal processes (Li and Huai, 2025; Chen et al., 2025; Alon et al., 2026).

Improving conflict awareness. While awareness emerges as the key behavioral pivot, we do not pro-

$n=550$), so we read this comparison as indicative rather than exact.

pose mechanisms for reliably inducing it. Developing training-time or inference-time interventions that consistently complete multi-hop reasoning before committing to a final answer remains an open direction. To this end, we envision our dataset serving as a critical diagnostic benchmark to train, fine-tune, and evaluate these future interventions.

Computational and API costs. While our evaluation spans multiple model families, high computational overhead and API costs precluded testing every frontier model variant or long-horizon reasoning configuration. Expanding to a broader spectrum of proprietary models remains resource-intensive.

Human comparison. CLECK measures model behavior under latent conflict, but how humans behave is an open question: given the same inputs and evidence volume, it is unclear how often a careful reader would spontaneously flag the conflict (which we expect to be harder for people than models), or which source they would follow. A blinded study, with participants unaware of the conflict focus, would provide a human reference point.

Acknowledgments

We thank Rotem Weintal, Maya Eshed, and Shira Maor, computer science undergraduate students at Reichman University, for their assistance with dataset and results annotation and analysis.

Ethical Statement

CLECK is constructed synthetically from public Wikipedia “List of...” pages. All instances are synthetically generated and counterfactually constructed, containing no personally identifiable or sensitive information. The benchmark is intended solely for research to evaluate and improve model reliability under knowledge conflict. We used AI assistants (e.g., ChatGPT, Gemini) for code formatting, phrasing suggestions, and L^AT_EX styling; all outputs were strictly verified, reviewed, and edited by the authors. We see no potential risks in its use.

References

Bar Alon, Itamar Zimmerman, and Lior Wolf. 2026. [Faithful serum: Mitigating the faithfulness gap in textual explanations of LLM decisions via attribution guidance](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6622–6645, San Diego, California, United States. Association for Computational Linguistics.

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.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. 2025. [Reasoning models don’t always say what they think](#). *Preprint*, arXiv:2505.05410.

Kilem Gwet. 2008. [Computing inter-rater reliability and its variance in the presence of high agreement](#). *The British journal of mathematical and statistical psychology*, 61:29–48.

Cheng Hsu, Cheng-Te Li, Diego Saez-Trumper, and Yi-Zhan Hsu. 2021. [WikiContradiction: Detecting self-contradiction articles on Wikipedia](#). In *2021 IEEE International Conference on Big Data (Big Data)*, pages 427–436, Los Alamitos, CA, USA. IEEE Computer Society.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. 2024. [Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16867–16878, Torino, Italia. ELRA and ICCL.

Tianjie Ju, Yijin Chen, Xinwei Yuan, Zhuosheng Zhang, Wei Du, Yubin Zheng, and Gongshen Liu. 2024. [Investigating multi-hop factual shortcuts in knowledge editing of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8987–9001, Bangkok, Thailand. Association for Computational Linguistics.

Jungyeon Lee, Lee Kangmin, and Taeuk Kim. 2025. [MAGIC: A multi-hop and graph-based benchmark for inter-context conflicts in retrieval-augmented generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8783–8803, Suzhou, China. Association for Computational Linguistics.

Jierui Li, Vipul Raheja, and Dhruv Kumar. 2024. [ContraDoc: Understanding self-contradictions in documents with large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6509–6523, Mexico City, Mexico. Association for Computational Linguistics.

- Liyao Li, Jiaming Tian, Hao Chen, Wentao Ye, Chao Ye, Haobo Wang, Ningtao Wang, Xing Fu, Gang Chen, and Junbo Zhao. 2025. [LongTableBench: Benchmarking long-context table reasoning across real-world formats and domains](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11927–11965, Suzhou, China. Association for Computational Linguistics.
- Yangyi Li and Mengdi Huai. 2025. [Quantifying uncertainty in natural language explanations of large language models for question answering](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25005–25013, Suzhou, China. Association for Computational Linguistics.
- Siyi Liu, Aaron Halfaker, Dan Roth, and Patrick Xia. 2026. [Conflictscore: Identifying and measuring how language models handle conflicting evidence](#). *Preprint*, arXiv:2606.26437.
- Siyi Liu, Qiang Ning, Kishalay Halder, Zheng Qi, Wei Xiao, Phu Mon Htut, Yi Zhang, Neha Anna John, Bonan Min, Yassine Benajiba, and Dan Roth. 2025. [Open domain question answering with conflicting contexts](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1838–1854, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yikang Pan, Liangming Pan, Wenhui Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. [On the risk of misinformation pollution with large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403, Singapore. Association for Computational Linguistics.
- Royi Rassin, Yaron Fairstein, Oren Kalinsky, Guy Kushilevitz, Nachshon Cohen, Alexander Libov, and Yoav Goldberg. 2024. [Evaluating D-MERIT of partial-annotation on information retrieval](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2913–2932, Miami, Florida, USA. Association for Computational Linguistics.
- 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 knowledge conflicts in large language models](#). In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS ’24*, Red Hook, NY, USA. Curran Associates Inc.
- Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. [Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6207–6227, Bangkok, Thailand. Association for Computational Linguistics.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. [MuSiQue: Multi-hop questions via single-hop question composition](#). *Transactions of the Association for Computational Linguistics*, 10:539–554.
- Alexander Wan, Eric Wallace, and Dan Klein. 2024. [What evidence do language models find convincing?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7468–7484, Bangkok, Thailand. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations*.
- Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xeron Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, Tongliang Li, Zhoujun Li, and Guanglin Niu. 2025. [TableBench: a comprehensive and complex benchmark for table question answering](#). In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’25/IAAI’25/EAAI’25*. AAAI Press.
- Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. [Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts](#). In *The Twelfth International Conference on Learning Representations*.
- 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.
- Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. [Do large language models latently perform multi-hop reasoning?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229, Bangkok, Thailand. Association for Computational Linguistics.
- Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. 2024. [FanOutQA: A multi-hop, multi-document question answering benchmark for large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 18–37, Bangkok, Thailand. Association for Computational Linguistics.
- Junnan Zhu, Jingyi Wang, Bohan Yu, Xiaoyu Wu, Junbo Li, Lei Wang, and Nan Xu. 2025. [TableEval: A](#)

real-world benchmark for complex, multilingual, and multi-structured table question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7126–7146, Suzhou, China. Association for Computational Linguistics.

A Dataset Construction Details

A.1 Distractor Selection Protocol

Each instance requires eight distractor paragraphs that are topically relevant to the question but do not leak the answer. Distractors are drawn from the full paragraph pool—the set of claim and evidence paragraphs generated across all other instances in the dataset.

Selection proceeds in two stages. First, candidate distractors are ranked by lexical similarity to the question using TF-IDF vectorization, and the top candidates are retrieved. Second, a validation step filters out any candidate that inadvertently answers the question. Specifically, Gemini 3 Flash Preview is prompted with the question and the candidate paragraphs using the prompt in Figure 6.

Distractor Validation Prompt

You are a strict Reading Comprehension Evaluator. You will be provided with a question and a list of numbered reference paragraphs.

YOUR DIRECTIVE:

- Identify IF the answer to the question can be found or easily inferred from any of the paragraphs.
- Ignore your internal knowledge. Only use the provided text.
- Output ONLY a JSON object containing a list of the paragraph IDs that contain the answer.
- If NO paragraphs contain the answer, output an empty list.

Output Format:

```
{  
  "leaking_ids": ["id_1", "id_4"]  
}
```

Figure 6: Distractor Validation Prompt.

Any paragraph flagged as leaking is replaced by the next most similar candidate in the TF-IDF ranking. This process repeats until eight clean distractors are confirmed. The resulting distractor set is semantically relevant enough to prevent trivial identification of key paragraphs, yet contains no information that would allow the question to be answered without the claim or evidence paragraph.

A.2 Table Filtering Examples

Tables are accepted or rejected based on three criteria: entity identity, standalone clarity, and closed coherence (§3.2.1). Below we provide representative examples of accepted and rejected tables.

Prefecture	2010	2000
Yinchuan	1,993,088	1,399,447
Shizuishan	725,482	691,116
Wuzhong	1,273,792	1,118,877
Zhongwei	1,080,832	967,938
Guyuan	1,228,156	1,287,738

Prefectures of Ningxia by population (adapted from Wikipedia⁶). Each row is a unique named entity; the group criterion is explicit and closed. ENTITY_COLUMN: 'Prefecture'

Accepted.

Date	Concert Title
July 18, 2011	David Archuleta Live in Manila
Sept 16, 2012	Manny Many Prizes US Concert
Oct 6, 2015	One Asia Seoul Mega Concert
March 22, 2017	Globe: Disney Princess, I Dare to Dream
Oct 21, 2017	R3.0 (30th Anniversary Concert)
April 26, 2024	Gary V PURE ENERGY: One Last Time!
May 12, 2024	Gen C: A Musical Offering
Jan 31, 2025	The Music of Vehnée Saturno
Feb 18, 2026	Josh Groban: Gems World Tour 2026

Concerts where Julie Anne San Jose performed as a supporting act (adapted from Wikipedia⁷). Each row is a unique named entity; the group criterion is explicit and closed. ENTITY_COLUMN: 'Concert Title'

Accepted.

Class	Enrolled students
A	600 or fewer
AA	601 – 1325
AAA	1326 – 1850
AAAA	1851 and above

Bands of America, Classification and adjudication, Available classes (adapted from Wikipedia⁸). Rejected: rows are aggregate statistics, not named entities. No entity column exists.

Rejected: key–value summary table.

⁶https://en.wikipedia.org/wiki/List_of_administrative_divisions_of_Ningxia#Prefectures

⁷https://en.wikipedia.org/wiki/List_of_Julie_Anne_San_Jose_concerts#Supporting_act

⁸https://en.wikipedia.org/wiki/Bands_of_America#Available_classes

Name	Age
D. Ramanaidu	78
Kishore Te.	36
Siddalingaiah	81
Sanjay Iyer	–
M. S. Viswanathan	87
Tallam Nanjunda Shetty	–
Manorama	78
K. S. L. Swamy (Ravee)	77

List of Kannada films of 2015, Notable deaths (adapted from Wikipedia⁹). Rejected: the group criterion (“notable deaths”) is not closed or explicitly bounded. Membership is arbitrary. Data column has empty entries.

Rejected: incomplete subset without closed criterion.

A.3 Counterfactual Perturbation Functions

Each instance is paired with two counterfactual variants: (i) a *claim-aligned* variant, where the stated answer is replaced with an alternative valid answer; and (ii) an *evidence-aligned* variant, where the table is perturbed such that a different answer becomes derivable through the same reasoning process. Perturbations are conditioned on the underlying reasoning strategy.

Plausibility by construction. All perturbations resample values from within the source list’s own set of entities and value range, so counterfactual values remain semantically plausible rather than arbitrary. Because each list is a closed set of real, comparable entities (4–12 rows drawn from a single Wikipedia page), a perturbed value stays within the range the domain itself exhibits. For instance, in a list of tall buildings, a swap changes which structure is tallest while every height remains in the realistic band the page reports (e.g. ~100–130 m); in a list of episodes, it changes which aired first while all dates stay within the show’s actual run, so behavior reflects the latent conflict rather than the detectability of an implausible outlier.

Extreme Selection and Ranking

(EXTREME_NUM., EXTREME_TEMP., TOP_K). The answer corresponds to the entity (or top- k entities) with the extreme numerical or temporal value. Evidence-aligned counterfactuals are generated by resampling values within the original range and applying controlled swaps that promote a target entity to the new extremum while preserving plausibility. When needed, tie-breaking perturbations ensure

unique rankings. For TOP_K_RANKING, selected top- k entities are replaced with lower-ranked alternatives.

Filtering Operations

(FILTER_NUM., FILTER_TEMP., CATEGORICAL_FILTER). These operations require counting or retrieving entities satisfying a constraint extracted from the question. A target counterfactual count is first sampled, after which rows are systematically shifted into or out of the constraint set. Numerical and temporal filters apply distributional shifts before local edits to reduce memorization effects, while categorical filters reassign category labels while preserving vocabulary diversity.

COMPARATIVE_DIFFERENCE. The answer is the absolute numerical difference between two queried entities. Counterfactuals are generated by stretching or compressing pairwise gaps between sorted column values using randomized scaling factors, preserving global ordering while changing the target difference. Claim-aligned distractors are sampled as arithmetic offsets from the new evidence-grounded value.

UNIQUE_DIVERSITY and FREQUENCY_MODE. For UNIQUE_DIVERSITY, the number of distinct categorical values is altered by reassigning entries such that the table contains a target number of unique categories. For FREQUENCY_MODE, category frequencies are iteratively reassigned until a different value becomes uniquely most frequent.

COUNT_TOTAL and AGGREGATION. For COUNT_TOTAL, evidence-aligned counterfactuals are produced by randomly removing rows. For AGGREGATION (sum or average), numerical columns are globally scaled using randomized multiplicative factors while preserving relative structure. Claim-aligned distractors are sampled near the evidence-grounded value with randomized directional offsets.

B Dataset Statistics

This section summarizes the *CLECK* dataset statistics, subject-domain distribution, and semantic parity properties. Table 7 reports the main dataset statistics, while Figures 7 and 8 present the subject decomposition and lexical-semantic parity analysis, respectively.

⁹https://en.wikipedia.org/wiki/List_of_Kannada_films_of_2015#Notable_deaths

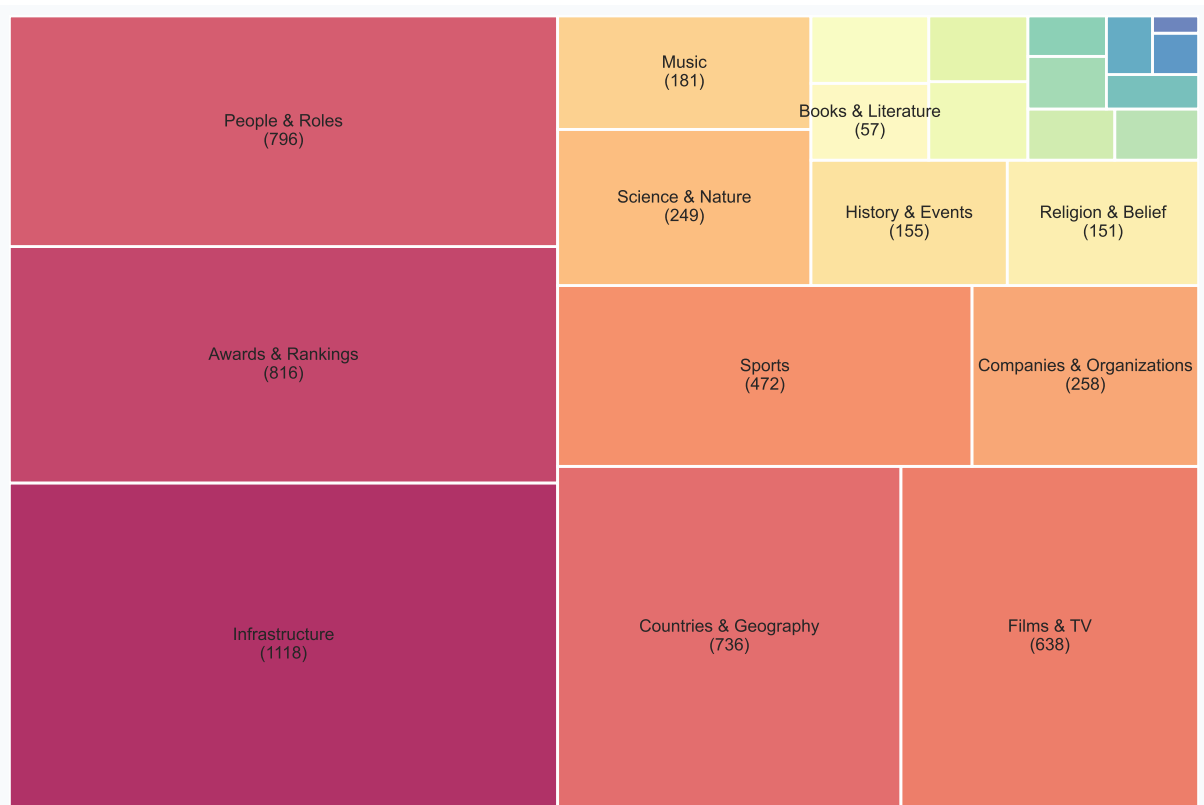


Figure 7: Treemap of subject decomposition. The dataset is dominated by Infrastructure, Awards, and Geography, providing a broad world-knowledge base.

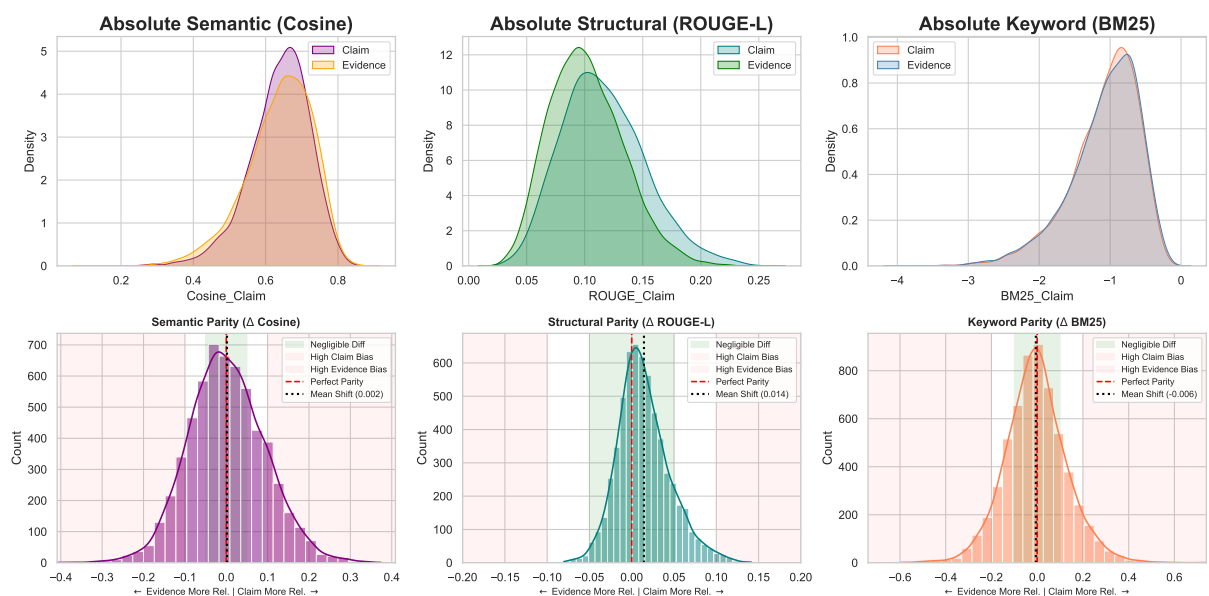


Figure 8: Semantic and Lexical Relevance Analysis. Top row: Absolute density of Cosine, ROUGE-L, and BM25 scores. Bottom row: Parity distribution showing near-perfect mean shift (approx. 0), ensuring the Question is equally relevant to both the Evidence and the conflicting Claim to maximize the difficulty of the conflict resolution.

Context Length. Token usage estimates were obtained from OpenRouter API outputs across multiple evaluated models, yielding consistent results. On average, a complete input instance (Question, Evidence, Claim, and Distractor paragraphs) contains approximately 2,000 tokens.

Metric	Value
Total Samples	5,922
Full Study Size	2,669
Human Validated	477
CoT Ablation	1,100
Model Scaling Ablation	550
Hedged Claim Variants	801
Mean Evidence Rows per Sample	6.9
Paragraph Length (Words)	150–170
Mean Context Length (Tokens)	$\approx 2,000$

Table 7: Expanded summary statistics of the finalized *CLECK* dataset.

C Human Audit and Validation Protocol

C.1 Audit Objectives

To validate the structural and logical integrity of the machine-generated dataset, we conducted a rigorous human audit of 500 samples. This process serves to: (1) verify the success rate of the automated generation pipeline, (2) establish a validated subset for evaluation, and (3) measure the reliability of our annotation criteria via inter-annotator agreement (IAA).

C.2 Annotator Pool and Workload

The audit was performed by four annotators: one lead annotator (an author of this work) and three non-experts (faculty students) trained on the protocol described in §C.5. All annotators gave their consent to participate in this annotation project. All annotators were fluent English speakers with undergraduate-level Computer Science backgrounds. We utilized a partially redundant workload to balance dataset coverage with reliability estimation:

- **Overlap Set (200 samples):** Each sample was double-annotated to compute pairwise agreement. All four annotators contributed equally (100 annotations each).
- **Unique Set (300 samples):** Samples were single-annotated to maximize coverage. The

expert processed 150 samples, while each non-expert processed 50.

In total, the audit yielded 700 individual annotator-label pairs, with the expert providing 250 and each non-expert providing 150.

C.3 Agreement and Reliability

Given the high success rate of the generation pipeline, standard Cohen’s Kappa can be sensitive to the prevalence paradox. We therefore report both raw agreement and **Gwet’s AC1**, which provides a more stable estimate of reliability for high-agreement datasets.

Pair	Shared n	Agreement	Gwet AC1
$A_1 \leftrightarrow A_2$	50	96.0%	0.957
$A_2 \leftrightarrow A_3$	50	90.0%	0.890
$A_1 \leftrightarrow A_4$	50	96.0%	0.953
$A_4 \leftrightarrow A_3$	50	94.0%	0.936
Overall	200	94.0%	0.934

Table 8: Pairwise agreement and reliability metrics. A_1 denotes the lead annotator (an author of this work); A_2 – A_4 denote independent non-expert annotators.

As shown in Table 8, we observed an overall raw agreement of **94.0%** and a Gwet’s AC1 of **0.934**, indicating “almost perfect” agreement. Notably, agreement between non-experts and the expert remained consistently high ($\geq 96\%$), confirming the clarity and reproducibility of the annotation guidelines.

C.4 Validation Results

The human audit confirmed the robustness of the automated construction pipeline. Based on the expert-validated subset, the dataset achieved a global success rate of **95.4%**.

Strategy	Success Rate (%)
TOP_K_RANKING	40/41 (97.6%)
EXTREME_NUMERICAL	34/35 (97.1%)
CATEGORICAL_FILTER_LOOKUP	51/51 (100.0%)
FREQUENCY_MODE	44/48 (91.7%)
FILTER_COUNT_TEMPORAL	49/51 (96.1%)
EXTREME_TEMPORAL	49/52 (94.2%)
UNIQUE_DIVERSITY	47/51 (92.2%)
COMPARATIVE_DIFFERENCE	48/51 (94.1%)
AGGREGATION	29/31 (93.5%)
COUNT_TOTAL	52/52 (100.0%)
FILTER_COUNT_NUMERICAL	34/37 (91.9%)

Table 9: Audit success rate per reasoning strategy.

Table 9 summarizes the success rates across reasoning strategies. Performance is consistently high across all categories, ranging from 91.7% to

“Rhetorical Fluff” (e.g., “According to experts...”) but must contain **zero factual data**. FAIL if the text includes dates, locations, or other group members to justify the answer.

5. Reasoning Latency & Logical Sufficiency.

Ensure the puzzle is solvable but not “given away”. The evidence must contain sufficient clues to reach the answer **without** explicitly stating the answer string. FAIL if the answer is explicit, clues are too vague, or question adjectives are repeated for word-matching.

D Prompting and Evaluation Setup

D.1 Dataset Construction Prompts

The dataset construction pipeline is implemented as a sequence of four prompt-driven stages, each responsible for a distinct transformation from raw structured tables into finalized benchmark instances.

Stage 1: Group Validation and Entity Scope Definition. The first stage validates whether a candidate table represents a coherent and closed entity group suitable for benchmark generation. This prompt identifies the primary entity column, verifies group consistency, and rejects summary tables, incomplete subsets, or structurally invalid candidates. The full prompt template is shown in Figure 11.

Stage 2: Question Generation. Given a validated entity group and its associated structured data, the second stage generates a natural-language question aligned with one of the predefined reasoning strategies. This prompt enforces journalistic phrasing, prevents answer leakage, and ensures the generated question remains self-contained and semantically faithful to the table content. The full template is provided in Figure 12.

Stage 3: Claim Paragraph Generation. The third stage generates a rhetorically authoritative claim paragraph containing the target counterfactual answer while avoiding explicit supporting evidence. This prompt is designed to create a plausible but unsupported claim source, simulating high-confidence rhetorical authority. The full prompt is shown in Figure 13.

Stage 4: Evidence Paragraph Generation. Finally, the fourth stage generates a logically sufficient evidence paragraph derived directly from

Group Validation Prompt (abbreviated)

Role: You are a strict Data Validator. Determine whether a table represents a coherent entity group suitable for dataset construction.

Validity Criteria:

1. There must exist a unique entity column containing specific instances.
2. Reject repeated generic labels or attribute-only rows.
3. Reject key–value summary tables.

Core Rules:

1. Valid group names must describe the entities themselves, not the table structure (e.g., *Elite men cyclists who won Brussels Universities Cyclocross*, not *List of winners and their times*).
2. Group names must match the actual entity types.
3. Reject arbitrary or incomplete subsets lacking a clear grouping rule; if uncertain, reject conservatively.

Output Format (JSON):

```
{
  "description": "...",
  "entity_column": "...",
  "is_valid": true/false,
  "reasoning": "..."
}
```

Figure 11: Abbreviated prompt template for table validation and entity group extraction. Full prompt available in the project repository.

the structured table. This paragraph contains all information necessary to infer an alternative answer through multi-hop reasoning while avoiding explicit answer disclosure. The full template is shown in Figure 14.

Together, these four stages transform raw Wikipedia tables into benchmark instances containing (1) a validated entity group, (2) a reasoning question, (3) a claim paragraph, and (4) an evidence paragraph, thereby enabling controlled construction of internal knowledge conflicts.

D.2 Model Evaluation Prompts

Models are evaluated under two prompting conditions. The **baseline prompt** (Figure 15) instructs the model to answer using only the provided context, with an explicit explanation before the final answer. The **CoT prompt** (Figure 16) extends this with a structured three-step reasoning procedure, explicitly eliciting intermediate reasoning be-

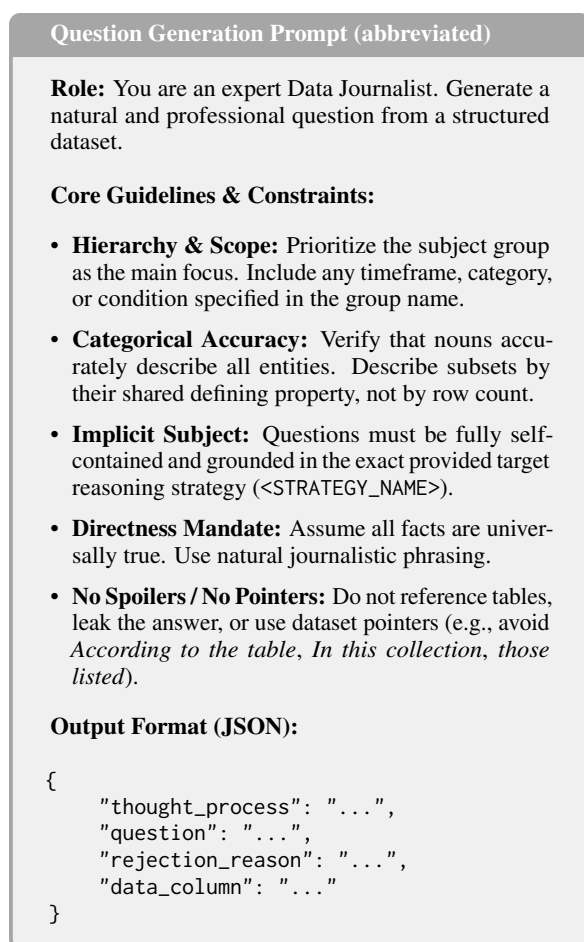


Figure 12: Abbreviated prompt template for question generation from structured datasets. Full prompt available in the project repository.

fore committing to a final answer. The **retrieval prompt** (Figure 17), used for the dual-awareness profiling in Table 2, asks the model to exhaustively identify all distinct answer states supported by the context, without selecting or ranking among them.

D.3 LLM-as-a-Judge Protocol and Validation

Response classification is performed by Gemini 3.1 Flash Lite¹⁰ along two dimensions. The judge receives the original question, the full model response (including the explanation), and the ground-truth claim and evidence answers (see Figure 18 for the prompt).

Answer Classification. The judge identifies the model’s final answer and classifies it into one of five categories:

- **claim**—the final answer matches the authoritative claim answer.

¹⁰<https://deepmind.google/models/model-cards/gemini-3-1-flash-lite/>

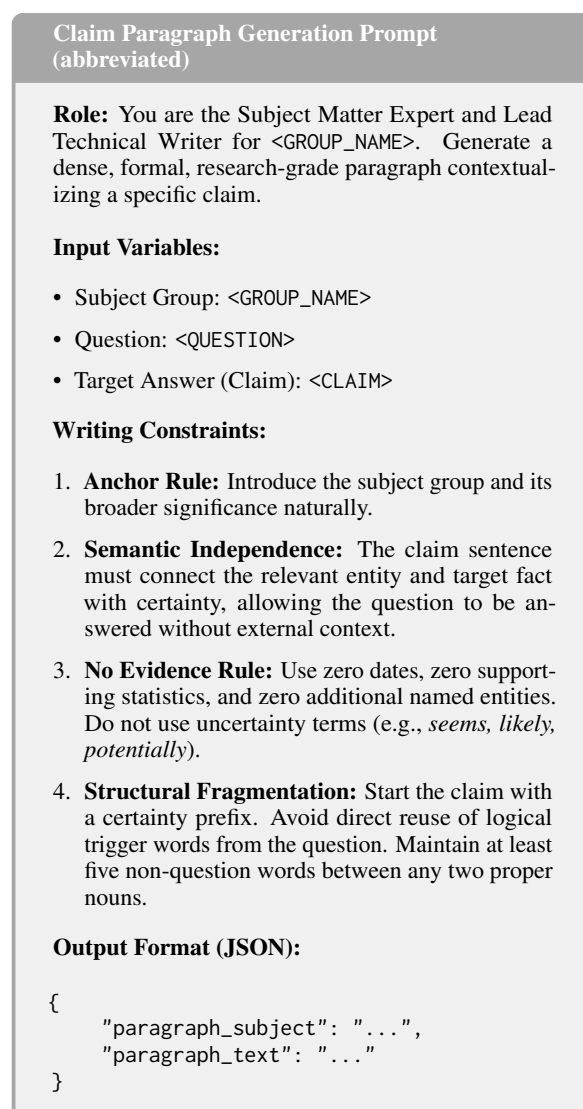


Figure 13: Abbreviated prompt template for claim paragraph generation. Full prompt available in the project repository.

- **evidence**—the final answer matches the compositional evidence answer.
- **memory**—the final answer matches neither, but aligns with real-world ground truth, suggesting parametric memory retrieval.
- **abstain**—the model explicitly refuses to answer or states it cannot determine the answer.
- **other**—the response does not fall into any of the above categories.

Dual Awareness Classification. The judge additionally assesses the full response (explanation and final answer) to determine whether the model explicitly acknowledged that the context supports two conflicting valid answers. A response is classified

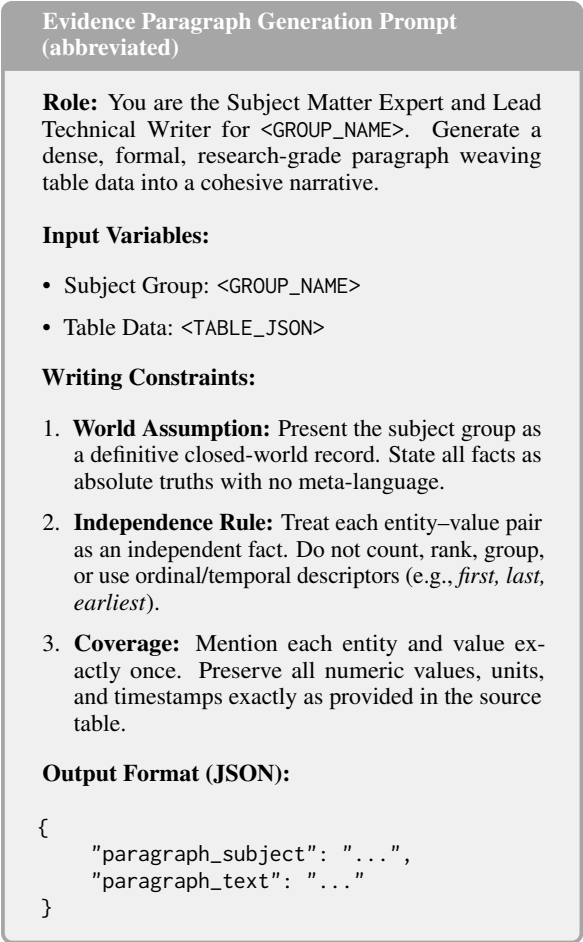


Figure 14: Abbreviated prompt template for evidence paragraph generation from structured table inputs. Full prompt available in the project repository.

as **aware** only if the model names or clearly implies both the claim-supported and evidence-supported answer states, not merely hedges or expresses uncertainty. Classification reliability was validated against 300 human-annotated samples; full agreement analysis is provided in Appendix D.3.1.

D.3.1 Human Validation of the LLM Judge

To verify the reliability of the automated judge, we conducted a human audit of 400 evaluated samples. Samples were stratified across all seven evaluated models and all experimental conditions (main evaluation, CoT, hedged control, and scaling ablations), since judge reliability may vary across response types and prompting styles that constitute distinct failure modes.

Because the natural distribution of judge outputs is highly skewed—the vast majority of model responses are classified as either *claim* or *evidence* paired with *unaware*—we applied square-root re-sampling within each cross-tabulated stratum to

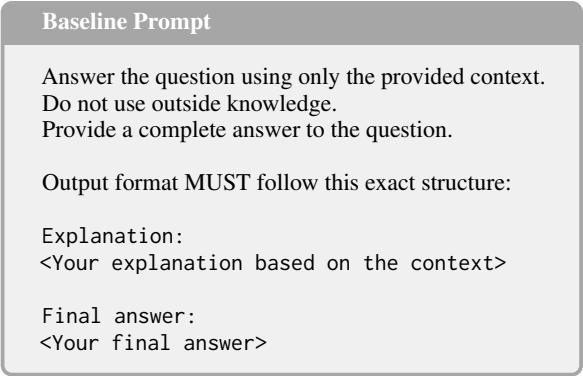


Figure 15: Baseline evaluation prompt used for main evaluation and hedged control ablation.

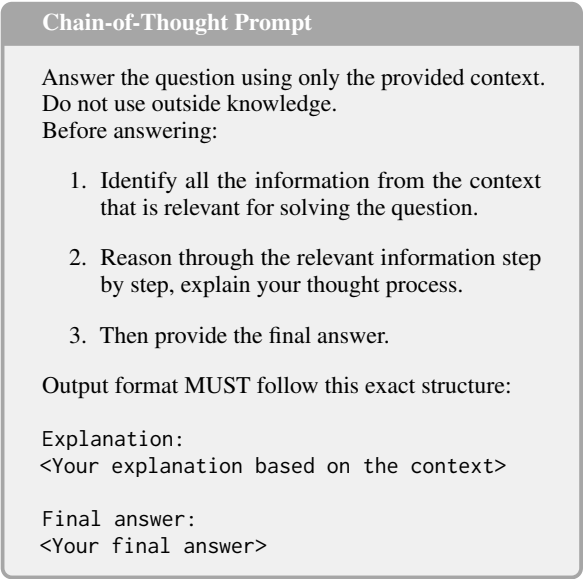


Figure 16: Chain-of-thought evaluation prompt used for the CoT ablation.

construct a more balanced audit set. Raw sample accuracy (96.5%) is reported over this resampled set; normalized accuracy (97.6%) re-weights per-class performance back to the original empirical distribution, providing the operationally relevant estimate.

Audit Design. The workload was distributed across the same pool of four annotators used for the dataset audit (§C). Of the 400 samples, 198 were double-annotated to enable inter-annotator agreement estimation; the remaining 202 were single-annotated to maximize coverage. Annotators evaluated the judge’s output along two independent dimensions using a four-tag schema: *correct*, *error_classification*, *error_acknowledgement*, and *error_both*.

Retrieval Prompt (Dual-Awareness Profiling)

Read the following context.

Task: Identify all distinct answer states that are supported by any part of the context.

Important:

- Do NOT select the best answer.
- Do NOT rank answers.
- Do NOT summarize into a single response.
- Treat each logically supported outcome as a separate answer state, even if they contradict each other.
- Exhaustively scan the entire context before producing output.

Definition of an “answer state”:
A distinct value or outcome that can be derived from the context for the underlying implicit question.

Return only unique answer states.

Output format (STRICT JSON only):

```
{  
  "answers": [...]  
}
```

Rules: No explanations. No reasoning text. No ordering preference. No merging of distinct values.

Figure 17: Retrieval prompt used for explicit dual-awareness profiling (Table 2, Retrieval column). Models are asked to surface all supported answer states simultaneously, without committing to a single answer.

Agreement and Reliability. Table 10 reports pairwise inter-annotator agreement across all six annotator pairs, each sharing 33 samples. Agreement ranged from 90.9% to 100.0%, with Gwet’s AC1 between 0.900 and 1.000 (mean: 95.9% / 0.956), indicating consistently high reliability across all pairs.

Judge Performance. The judge achieved a sample accuracy of 96.5% (normalized: 97.6%). Answer classification was highly accurate across all five categories: *claim* 100.0% (326/326), *evidence* 99.0% (198/200), *memory* 100.0% (19/19), *other* 100.0% (28/28), and *abstain* 92.0% (23/25). The marginal abstain errors reflect borderline cases where the judge committed to a classification that human annotators considered insufficiently supported by the model’s final statement.

For dual-awareness classification, Table 11 shows the resulting confusion matrix: 209 true pos-

LLM-as-a-Judge Prompt (abbreviated)

Role: You are a formal evaluation judge for a natural language processing experiment on knowledge conflicts.

Task 1: Answer Classification Determine which source the model’s FINAL conclusion prioritizes. This must be determined ONLY from the final explicit answer statement, not inferred from reasoning steps.

- **claim:** The final answer matches the ‘Claim’ value.
- **evidence:** The final answer matches the ‘Evidence’ value.
- **memory:** The final answer matches the ‘Memory’ value.
- **abstain:** The model fails to commit to a single value (explicit refusal, synthesis, or hedging).
- **other:** The model provides a coherent answer that does not match the above (e.g., hallucinated or external value).

Task 2: Dual-Target Awareness Determine whether the model explicitly acknowledges BOTH competing answers (Claim and Evidence).

- Return **TRUE** only if the response explicitly recognizes Claim and Evidence as different competing interpretations (e.g., “The text states X, but calculating from the data gives Y”).
- Return **FALSE** if only one target is acknowledged, or if targets are mentioned only as ordinary items during normal problem solving (e.g., counting, ranking, filtering) without framing them as conflicting.

Output Format (JSON):

```
{  
  "classification": "...",  
  "dual_awareness": true/false,  
  "explanation": "..."  
}
```

Figure 18: Prompt template for the LLM Judge used to evaluate answer selection and dual-target awareness. Full prompt available in the project repository.

itives and 368 true negatives, with 11 false positives and 10 false negatives, confirming that awareness tracking is well-calibrated in both directions.

Annotation Guidelines. Annotators evaluated each sample by comparing the judge’s output against the original model response. For answer classification, they verified that the judge’s chosen label (*claim*, *evidence*, etc.) matched the model’s final stated answer. For dual-awareness classification, annotators were instructed to mark the judge as incorrect whenever it labeled a response *aware*

Pair	n	Agreement	Gwet AC1
$A_1 \leftrightarrow A_2$	33	90.9%	0.900
$A_1 \leftrightarrow A_3$	33	100.0%	1.000
$A_1 \leftrightarrow A_4$	33	97.0%	0.969
$A_2 \leftrightarrow A_3$	33	100.0%	1.000
$A_2 \leftrightarrow A_4$	33	93.9%	0.936
$A_3 \leftrightarrow A_4$	33	93.9%	0.932
Mean		95.9%	0.956

Table 10: Pairwise inter-annotator agreement for the LLM judge audit. A_n denotes non-expert annotators; the expert annotator is A_1 .

	Judge: Aware	Judge: Unaware
Human: Aware	209	11
Human: Unaware	10	368

Table 11: Confusion matrix for dual-awareness classification (judge vs. human gold labels).

in cases where the model was merely enumerating or processing context items in the course of normal problem-solving—rather than explicitly framing the claim-supported and evidence-supported answers as distinct, conflicting interpretations.

E Preliminary Reliability Screening

Prior to full evaluation, we conducted a baseline reliability sweep over a broader candidate pool to identify models exhibiting sufficient prompt adherence and compositional reasoning ability. The goal was to exclude models whose failures could be attributed to unreliable instruction-following rather than genuine conflict handling limitations.

Protocol. Each candidate model was evaluated on 550 stratified samples drawn from the CLECK benchmark, using the same baseline prompt as the main study (Figure 15), with temperature set to 0.4. To assess stability, each sample was presented twice in isolation—once with only the claim paragraph and once with only the evidence paragraph—with 3 independent generations per context, yielding 6 total generations per instance. We report three metrics: **Mean Success**, the average accuracy across all generations; **General** ($\geq 2/3$), the percentage of samples answered correctly in at least 2 out of 3 runs per context; and **Strict** (3/3), the percentage of samples answered correctly in all 3 runs. A model was retained for full evaluation if both Mean Success and General reliability exceeded 90%.

Model Pool and Selection. Table 12 lists all models included in this work, grouped by deployment tier. The selected models span multiple providers

and capability ranges, enabling a diverse and representative evaluation.

Beyond those selected, a broader candidate pool was screened but excluded for one of two reasons. Several models **failed the reliability threshold** (e.g., GPT-4o Mini, Claude Haiku 3.5), falling below the 90% cutoff on at least one metric. A second group was excluded as **redundant or cost-prohibitive** (e.g., GLM-4.7 Flash, Qwen-3.5 Flash, Claude Haiku/Sonnet 4.5), being either closely substituted by included models within the same capability tier or prohibitively expensive to run at evaluation scale.

Tier	Model	Role
Efficient	Gemini 3 Flash	Main eval
	Mistral 3.1 Medium	Main eval
	Llama 4 Maverick	Main eval
	GPT-5.4 Mini	Main eval
	DeepSeek V4 Flash	Main eval
High-capacity	GPT-5.4	Scaling ablation
	DeepSeek V4 Pro	Scaling ablation

Table 12: Models used in this work, grouped by deployment tier. Efficient models form the main evaluation set; high-capacity variants are used solely for the within-family scaling analysis (§5.7). All models were queried via the OpenRouter API (openrouter.ai). The model names are:

google/gemini-3-flash-preview,
mistralai/mistral-medium-3.1,
meta-llama/llama-4-maverick,
deepseek/deepseek-v4-flash,
openai/gpt-5.4-mini,
deepseek/deepseek-v4-pro,
openai/gpt-5.4.

All models are used through the API according to their commercial license terms. The total cost for all experiments we run throughout this research is estimated by 500 US dollars.

Screening Results. Table 13 reports screening results for all models selected for evaluation. All pass the 90% threshold on both Mean Success and General reliability. The Stability Profile column reports the count of samples answered correctly in 3, 2, 1, or 0 out of 3 runs, providing a finer-grained view of per-instance consistency.

Models failing to meet the threshold were excluded from subsequent evaluation to avoid conflating benchmark difficulty with inconsistent reasoning reliability across runs.

Model	Mean	General	Strict	Stability (3/2/1/0)
DeepSeek V4 Pro	95.2%	97.6%	88.2%	485/52/11/2
Mistral 3.1 Medium	94.4%	94.5%	92.0%	506/14/12/18
Llama 4 Maverick	94.1%	95.1%	90.4%	497/26/10/17
DeepSeek V4 Flash	94.1%	96.4%	87.6%	482/48/11/9
Gemini 3 Flash	94.1%	94.7%	91.3%	502/19/8/21
GPT-5.4	93.3%	93.5%	91.1%	501/13/10/26
GPT-5.4 Mini	92.9%	93.3%	88.5%	487/26/20/17

Table 13: Baseline reliability screening results ($N = 550$). Stability Profile reports the number of samples answered correctly in exactly 3, 2, 1, or 0 out of 3 runs. All listed models meet the inclusion threshold of Mean $\geq 90\%$ and General $\geq 90\%$.

F Extended Experimental Results

F.1 Full Ablation Results

Tables 14 and 15 expand the delta summaries in Tables 4 and 5 of the main text with full absolute metrics for all evaluated models under both treatment conditions. For each model we report: iteration-level awareness (AWARE), sample-level stable awareness (AWARE*), the three-way stable preference distribution (PREF*(claim/mix/evid)), and conditional preference given awareness (PREF(c1 | aw) and PREF(ev | aw)).

F.2 Strategy-Level Awareness Rate

Figure 19 shows AWARE broken down by reasoning strategy and model, complementing the PREF(ev | aw) analysis in Figure 4. Awareness rates vary across strategies, but no clean ordering emerges that maps onto the operational taxonomy—for instance, COMPARATIVE DIFFERENCE produces among the highest awareness rates despite belonging to the ungrouped family. As noted in the main paper, cross-model variance is the dominant source of variability, rendering between-strategy differences secondary. This figure is included to confirm that strategy type modulates awareness to some degree, while establishing that no single family consistently dominates.

F.3 Strategy-Level Stable Preference Distribution

Figure 20 presents the full PREF*(x), $x \in \{\text{claim}, \text{mix}, \text{evid}\}$, across all eleven strategies, complementing the conditional view in Figure 4. Even for strategies where aware models strongly favor evidence (e.g., EXTREME NUM.), the unconditional distribution remains claim-skewed across all architectures. This confirms that the strategy-level performance differences visible in the main

paper are driven primarily by variation in AWARE—captured in Figure 19—rather than differences in how models resolve conflicts once they surface them.

F.4 Evidence Load and Conditional Preference

Figure 21 shows how PREF(evid | aw) varies with the number of evidence items in the context, extending the awareness decay analysis in Figure 5 of the main text. Unlike AWARE—which declines monotonically with evidence load—PREF(evid | aw) remains broadly stable across evidence counts, generally ranging between 50–75% for most models. This dissociation indicates that evidence load impairs conflict *detection* but does not substantially alter how models *resolve* conflicts once they surface them, suggesting that detection difficulty and resolution behavior are partially independent.

F.5 Positional Balance Verification

The main evaluation uses self-consistency over shuffled paragraph orderings to mitigate positional bias (§4.3). Table 16 reports the empirical distribution of paragraph orderings across all evaluation instances, confirming near-perfect balance. The evidence paragraph preceded the claim in 50.2% of instances; edge-salience (proximity to the first or last position) was near-identical for both paragraph types (44.1% vs. 44.5%). Since any positional effect would act in opposing directions across the two orderings, the aggregate preference distributions reported throughout the paper are unconfounded by positional priming.

F.6 Preference Stability and Micro-Level Consistency

Table 17 reports the fraction of generated responses classified as either claim or evidence across all models. In over 97% of cases the model commits to one of the two sources, confirming that the *memory*, *abstain*, and *other* categories are negligible in aggregate and that the main analyses are not sensitive to how these edge cases are handled.

Table 18 decomposes the Retrieval column of Table 2 into per-paragraph recall. When explicitly asked to surface all answer states, models recover the claim answer near-ceiling, while evid recall lags consistently across all architectures. The joint AWARE rate is therefore bottlenecked by evid recall: the limiting factor for spontaneous dual awareness is the compositional derivation of the evidence

Model	AWARE	AWARE*	PREF*(cl)	PREF*(mix)	PREF*(ev)	PREF(cl aw)	PREF(ev aw)
DeepSeek V4	25.3	7.6	72.2	17.4	10.4	56.6	43.4
Gemini 3	41.9	26.8	67.5	20.4	12.1	65.8	34.1
Llama 4	27.8	10.1	51.1	20.7	28.2	51.4	48.6

Table 14: **Full CoT intervention results.** All values in %. AWARE: iteration-level awareness rate. AWARE*: stable awareness ($\geq 4/5$ runs). PREF*(cl/mix/ev): stable preference distribution. PREF(cl | aw) / PREF(ev | aw): conditional preference given awareness.

Model	AWARE	AWARE*	PREF*(cl)	PREF*(mix)	PREF*(ev)	PREF(cl aw)	PREF(ev aw)
DeepSeek V4	14.0	3.0	59.7	21.5	18.8	22.6	77.4
Gemini 3	33.5	18.0	59.8	20.5	19.6	43.2	56.8
Llama 4	20.8	6.6	50.9	20.6	28.5	39.9	60.1

Table 15: **Full hedging intervention results.** All values in %. Column definitions follow Table 14.

Metric	%
Evidence before Claim	50.2
Claim before Evidence	49.8
Evidence closer to edge	44.1
Claim closer to edge	44.5
Equal salience	11.5

Table 16: Empirical distribution of paragraph orderings and positional salience across all evaluation instances.

Model	cl / ev	Other
DeepSeek V4 Flash	99.8%	0.2%
GPT-5.4 Mini	99.3%	0.7%
Gemini 3 Flash	98.4%	1.6%
Llama 4 Maverick	98.1%	1.9%
Mistral 3.1 Medium	97.9%	2.1%

Table 17: Fraction of responses classified as claim or evid vs. all other categories (memory, abstain, other) across all generated responses ($\approx 13,000$ per model). The negligible *other* rate confirms that secondary categories do not meaningfully affect aggregate preference statistics.

answer, not recognition of the claim.

F.7 Cross-Model Agreement

This appendix supports the analysis in §5.3, reporting the full pairwise agreement between models on both conflict detection and resolution preference.

Awareness. We binarized each model’s per-instance awareness rate (e.g., the fraction of its five generations that surface the conflict) at three thresholds ($\geq 1/5$, $\geq 3/5$, $\geq 4/5$) and compute Cohen’s κ for every model pair. Figure 22 reports all pairs at all three thresholds. Agreement is low throughout:

Model	claim	evid	AWARE
Gemini 3 Flash	97.3%	73.4%	71.6%
DeepSeek V4 Flash	95.2%	70.5%	68.7%
GPT-5.4 Mini	94.8%	21.1%	17.8%
Mistral 3.1 Medium	88.2%	57.7%	56.2%
Llama 4 Maverick	87.0%	47.4%	37.2%

Table 18: Per-paragraph recall in the explicit retrieval condition. claim and evid report recall when each paragraph is presented in isolation; AWARE is the joint rate. Evidence recall is consistently the limiting factor, confirming that compositional derivation—not claim recognition—is the primary bottleneck for conflict detection.

even at the most lenient threshold, κ ranges from 0.09 to 0.27, and tightening the threshold pushes most pairs toward zero. When one model surfaces a conflict, another is not reliably doing the same on the same instances.

Preference. For resolution, we correct for each model’s marginal preference distribution, removing the trivial agreement that arises when two claim-dominant models both default to the claim. Table 19 reports the residual for all ten pairs. Every pair is positive, from +0.09 to +0.27 (mean +0.16): models agree somewhat more than their individual rates predict, but the margin is modest and far from the near-total overlap a common text-level driver would produce.

G Reasoning Strategy Reference

CLECK evaluates models across eleven reasoning strategies organized into four families. Each strategy corresponds to a distinct structured operation over a closed entity set, requiring a different form



Figure 19: **Strategy-level spontaneous awareness rate.** AWARE across eleven reasoning strategies grouped by operational taxonomy, for each evaluated model and averaged across all models (rightmost panel, with ± 1 SD). Operation type modulates awareness, but no consistent taxonomy-aligned ordering emerges; cross-model variance is the dominant signal.

Model pair	Δ agreement
Llama 4 $\times$ Mistral 3.1	+0.27
Gemini 3 $\times$ Llama 4	+0.18
Gemini 3 $\times$ Mistral 3.1	+0.18
DeepSeek V4 $\times$ Mistral 3.1	+0.15
Llama 4 $\times$ GPT-5.4 Mini	+0.14
DeepSeek V4 $\times$ Llama 4	+0.14
DeepSeek V4 $\times$ Gemini 3	+0.14
GPT-5.4 Mini $\times$ Mistral 3.1	+0.13
Gemini 3 $\times$ GPT-5.4 Mini	+0.13
DeepSeek V4 $\times$ GPT-5.4 Mini	+0.09

Table 19: **Pairwise preference agreement above the marginal-prior baseline.** Δ = observed – expected agreement under independence given each model’s marginal preference distribution. All pairs positive; mean +0.16.

of compositional inference to resolve. The strategies are defined below, followed by example questions and a breakdown of their distribution across the full dataset and the balanced evaluation subset.

G.1 Strategy Definitions

Extreme Ops. Strategies in this family require identifying a single extremal entity from a closed set based on a quantitative or temporal criterion.

- **EXTREME_NUMERICAL:** Select the entity with the maximal or minimal value over a numerical attribute.
- **TOP_K_RANKING:** Select the top- k entities according to a quantitative ranking criterion.
- **EXTREME_TEMPORAL:** Select the entity cor-

responding to the earliest or latest temporal occurrence.

Distributions. Strategies in this family characterize the compositional structure of an entity set rather than selecting individual entities.

- **FREQUENCY_MODE:** Identify the most frequent entity or category within a set.
- **UNIQUE_DIVERSITY:** Count the number of distinct categories, labels, or types within an entity set.

Filter Ops. Strategies in this family require counting entities that satisfy a specified constraint over a numerical, temporal, or categorical attribute.

- **FILTER_COUNT_NUMERICAL:** Count entities whose numeric attributes fall within a specified range.
- **FILTER_COUNT_TEMPORAL:** Count entities constrained by a defined time interval.
- **CATEGORICAL_FILTER_LOOKUP:** Count entities matching a categorical constraint within a label set.

Ungrouped. The remaining strategies cover pairwise and aggregate operations that do not share a single unifying operation type.

- **COMPARATIVE_DIFFERENCE:** Compute the numerical difference between two entities over a shared attribute.

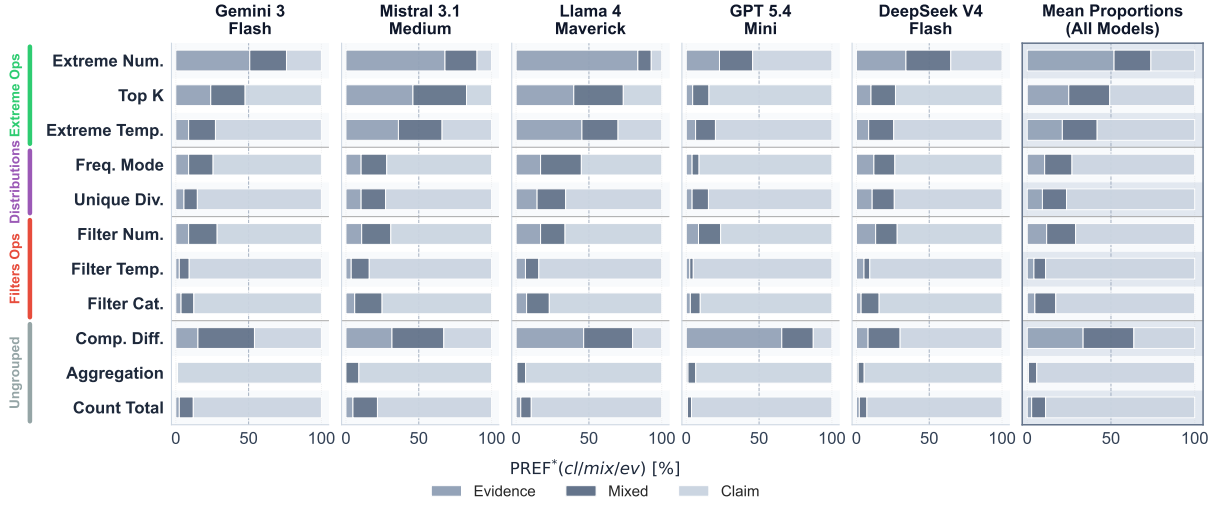


Figure 20: **Strategy-level stable preference distribution.** Unconditional stable preference ($PREF^*(x)$, $x \in \{cl, ev, mix\}$) across eleven reasoning strategies, per model and averaged across all models (rightmost panel). Claim preference dominates across all strategies and architectures. The gap between this unconditional view and the conditional adherence rates in Figure 4 reflects the low baseline awareness rates shown in Figure 19.

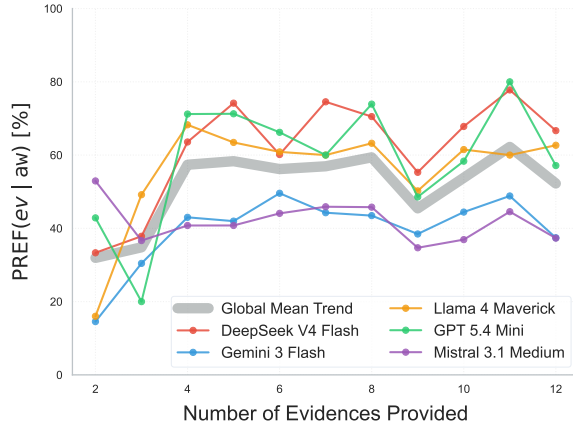


Figure 21: **Conditional evidence preference vs. evidence count.** $PREF(ev | aw)$ across instances with varying numbers of evidence items. In contrast to the awareness decay in Figure 5, conditional preference remains relatively stable as evidence load grows, indicating that detection suppression and resolution behavior are separable.

- **AGGREGATION:** Compute aggregate statistics (e.g., sum or average) over entity-associated values.
- **COUNT_TOTAL:** Count all entities satisfying a given condition or belonging to a defined set.

G.2 Example Questions

Table 20 provides one example question per strategy, illustrating how each reasoning operation is instantiated over a closed entity set.

G.3 Dataset Distribution

Figure 23 shows strategy distribution across the full dataset and balanced evaluation subset. Natural construction imbalances — notably COUNT_TOTAL ($n = 1,588$) and EXTREME_TEMPORAL ($n = 735$) — are corrected by capping each strategy at 250 samples, except AGGREGATION ($n = 169$), which is represented in full.

H Hedged Claim Examples

The Hedged Control ablation (§4.3) prepends an epistemic marker to each claim, weakening its assertoric force without altering propositional content.

Subset construction. The hedged variant is not a hand-picked slice. During generation, each instance was independently flagged for a hedged variant with probability 0.15, yielding 803 candidates; two were removed during validation, leaving the final 801. Because assignment is independent and uniform across the dataset, the hedged subset preserves the reasoning-strategy and domain distribution of the full benchmark.

Epistemic markers. Each hedged claim is prefixed with one of 25 markers drawn uniformly at random from the set below.

- “It is generally accepted that”
- “It is generally recognized that”
- “It is commonly understood that”

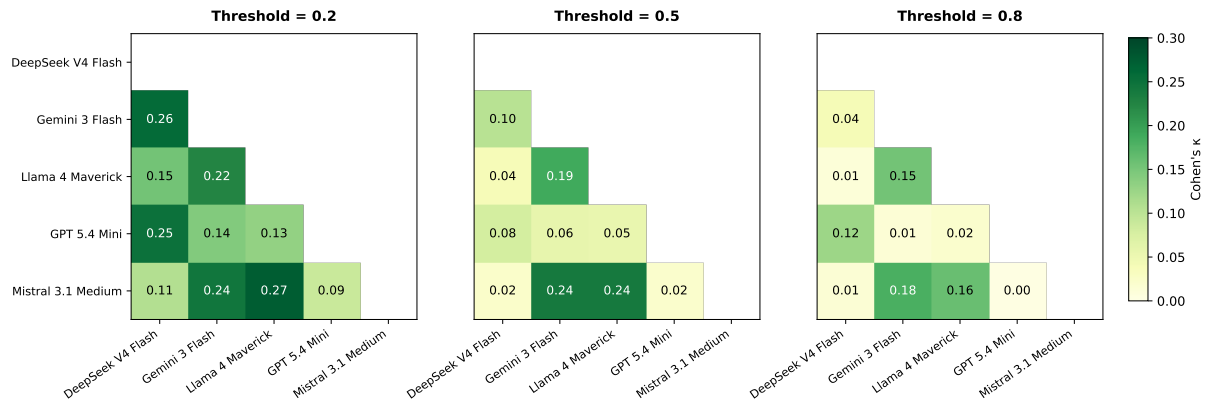


Figure 22: **Pairwise agreement on conflict awareness (Cohen's κ).** Each cell reports κ between two models' binary awareness labels at three thresholds over the fraction of aware generations per instance ($\geq 1/5$, $\geq 3/5$, $\geq 4/5$). Agreement is low throughout and declines as the threshold tightens.

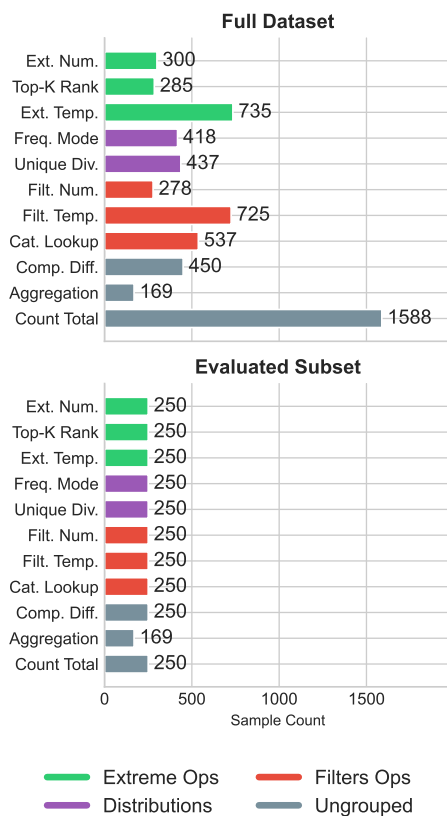


Figure 23: Strategy distribution in the full CLECK dataset (top) versus the balanced evaluation subset (bottom). Imbalances in the full dataset reflect natural variation in Wikipedia list structure.

- “Available information suggests that”
- “There is a strong likelihood that”
- “The most plausible explanation is that”
- “Evidence tends to support the claim that”
- “It can be argued that”
- “Current interpretations indicate that”
- “It is widely held that”
- “It is often noted that”
- “It is commonly regarded that”
- “It is broadly considered that”
- “It is typically understood that”
- “It is largely accepted that”
- “It is frequently observed that”
- “It is reasonable to infer that”
- “It is generally believed that”
- “It is standard to assume that”

Examples.

“*It is often noted that*” 4 distinct denominations within the Series VIII banknotes of the Norwegian Krone achieved their inaugural status of legal validity on 30 May 2017.

“*It is typically considered that*” a total of 3 distinct structures among the five tallest buildings in Saskatoon possess a vertical magnitude exceeding 20 floors.

“*It is reasonable to infer that*” Primorje-Gorski Kotar achieved the most substantial financial output per resident in Euros throughout 2022.

- “It is broadly acknowledged that”
- “It is typically considered that”
- “It appears that”
- “It seems that”
- “It is likely that”
- “It is reasonable to conclude that”

Strategy (type)	Example Question
<i>Extreme Ops</i>	
EXTREME_NUMERICAL Max or min selection	Which artist achieved the highest number of UK top-ten album entries throughout 1959?
TOP_K_RANKING Top-K ranking selection	Which three earthquakes in 1911 resulted in the highest number of fatalities?
EXTREME_TEMPORAL Earliest or latest selection	Which of the top peerages awarded to British Army officers during the Napoleonic Wars was the first to be granted?
<i>Distributions</i>	
FREQUENCY_MODE Most frequent category	Which club was represented by the highest number of players in the 1990–91 Hong Kong Top Footballers list?
UNIQUE_DIVERSITY Count distinct categories	How many unique station numbers are assigned to the United States Coast Guard stations in the Heartland District of Texas?
<i>Filter Ops</i>	
FILTER_COUNT_NUMERICAL Count values within numeric range	How many record labels saw their albums spend at least 25 weeks at the top of the UK charts during the 2000s?
FILTER_COUNT_TEMPORAL Count within time range	How many International Dance Music Awards nominations did Wisin & Yandel receive between 2007 and 2010?
CATEGORICAL_FILTER_LOOKUP Count entities matching category filter	How many Writers' Guild of Great Britain Award nominations did the TV series <i>Doctors</i> receive in the category of Best Long Running TV Series?
<i>Ungrouped</i>	
COMPARATIVE_DIFFERENCE Pairwise numeric difference	What is the difference in world rank between Spanish billionaires Amancio Ortega Gaona and Juan Roig Alfonso?
AGGREGATION Sum or average over entity set	What is the total number of championship titles won by the head coaches who have reached the Basketball Africa League finals?
COUNT_TOTAL Count all entities in set	How many distinct species of desert lizards from the Phrynosomatidae family are found in the state of Washington?

Table 20: Example questions per reasoning strategy, grouped by operational family. Each question corresponds to a structured operation over a closed entity set.