From Facts to Conclusions : Integrating Deductive Reasoning in Retrieval-Augmented LLMs

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) grounds large language models (LLMs) in external evidence, but fails when retrieved sources conflict or contain outdated or subjective information. Prior work address these issues independently but lack unified reasoning supervision. We propose a reasoning-trace-augmented RAG framework that adds structured, interpretable reasoning across three stages: **(1)** document-level adjudication, **(2)** conflict analysis, and **(3)** grounded synthesis, producing citation-linked answers or justified refusals.A Conflict-Aware Trust-Score (CATS) pipeline is introduced which evaluates groundedness, factual correctness, refusal accuracy, and conflict-behavior alignment using an LLM-as-a-Judge. Our **539-query** reasoning dataset and evaluation pipeline establishes a foundation for conflict-aware, interpretable RAG systems. Experimental results demonstrate substantial gains over baselines, most notably with Qwen, where Supervised Fine-Tuning (SFT) improved End-to-End answer correctness from **0.069** to **0.883** and behavioral adherence from **0.074** to **0.722**.

1 Introduction

Retrieval-augmented language modeling sits at the intersection of open-domain question answering and controllable generation. Contemporary large language models couple parametric knowledge with non-parametric retrieval to surface evidence at inference time, thereby improving factual calibration, temporal coverage, and verifiability in information-seeking workflows (Lewis et al., 2020; Borgeaud et al., 2022; Guu et al., 2020). In a canonical RAG pipeline, a retriever selects a small set of snippets from a large corpus and a generator composes an answer conditioned on those snippets, ideally with faithful attribution. However, real world data often introduces several challenges such as: differences in evidence quality, outdated information, subjective opinions, misinformation, and partial information. As a result, RAG models must reason over conflicting evidence, synthesize multi-hop dependencies across documents, and refrain from answering when support is absent, while maintaining strict grounding to the provided context.

Cattan et al. (Cattan et al., 2025) attempts to solve the problem by introducing a taxonomy of conflict types and defining expected model behaviors for each conflict type. It shows that conflict type awareness improves response quality. *Li et al.* (Li et al., 2024) presents a strategy that involves generating per-document notes, enabling a comprehensive assessment of their relevance to the input query. But it does not handle conflict in the retrieved passages.

Furthermore, *Song et al.* (Song et al., 2025) provides an evaluation metric called trust-score to evaluate the overall end to end quality of RAG systems. These studies have improved RAG systems but still leave key gaps. Many of them don't connect document-level reasoning with conflict-type inference or evaluate how well models handle conflicts, multi-hop reasoning, and justified refusals together.

We address this gap with a reasoning-trace-augmented RAG framework that integrates structured supervision into both training and evaluation. Our framework, inspired by *Cattan et al.* (Cattan et al., 2025) and *Li et al.* (Li et al., 2024), introduces a three-stage deductive reasoning process that emulates human adjudication over conflicting evidence. In **Stage 1 (Micro Reasoning)**, each retrieved document is labeled as supports, partially supports, or irrelevant, with extracted key facts, brief quotes, and evidence metadata to ensure fine-grained grounding. **Stage 2 (Macro Conflict Analysis)** aggregates these micro judgments to infer the overarching conflict type : no conflict, complementary information, conflicting opinions or research

outcomes, outdated information, or misinformation, following the Dragged into Conflicts taxonomy, and generates concise rationales and behavioral expectations. **Stage 3 (Final Grounded Synthesis)** consolidates consistent evidence to produce a citation-linked answer or a justified refusal, ensuring that responses conform to conflict-specific reasoning norms. All reasoning is serialized as **XML-like <think>** traces/tokens, providing transparency and interpretability absent in prior RAG systems.

Furthermore, we fine tuned medium open-weight models, *Qwen-2.5-7B-Intruct* (Yang et al., 2025) and *Mistral-7B-Instruct* (Jiang et al., 2023) on a 539 query dataset using *QLoRA* (Dettmers et al., 2023) technique. A comparative assessment was carried out between the **SFT fine tuned models** and the **baseline models** across two evaluation paradigms, the **oracle** setting and the **end to end** setting.

As a part of evaluation we introduce a new metric that extends the Trust-Score pipeline (Song et al., 2025) with conflict-behavior alignment to make Conflict-Aware Trust Score, **CATS**. The trust score involves the computation of the following metrics: **grounded refusal**, **answer correctness**, and **grounded citation**. Furthermore we incorporate an additional metric, that is, **behavioral adherence**, which evaluates the LLM's capability to generate conflict-type expected responses. The behavioral adherence is computed via *GPT-4o* (OpenAI et al., 2024) as the **LLM-as-a-Judge** under blinded prompts.

Experimental results demonstrate that our structured supervision yields substantial gains over baselines, particularly for models with weak initial conflict handling. In the End-to-End setting, *Qwen* achieved near-perfect refusal capabilities (**F1-GR** rising from **0.167** to **1.000**) and enhanced verifiability (**Grounded Citation** from **0.111** to **0.648**), alongside massive surges in **Answer Correctness** (**0.069** to **0.883**) and **Behavioral Adherence** (**0.074** to **0.722**).

Our contributions include **(1)** constructing a structured, conflict-aware reasoning dataset with document-level verdicts, conflict-type labels, and staged reasoning traces; **(2)** using this dataset to fine-tune our selected instruction-tuned LLMs through QLoRA for grounded, schema-consistent reasoning; **(3)** and evaluating these fine-tuned models against baseline versions on the test split using *CATS* pipeline, which measures grounding quality, answer correctness, refusal behavior, citation reliability, and adherence to conflict-specific behavioral norms through LLM-as-a-Judge scoring. Together, these components demonstrate how structured reasoning supervision improves the reliability, interpretability, and conflict sensitivity of retrieval-augmented generation systems.To facilitate reproducibility and future research, we release our annotated dataset, training scripts, and the complete CATS evaluation pipeline on GitHub.

2 Related Work

2.1 Knowledge and Evidence Conflicts

Retrieval Augmented Generation (RAG) consists of conditioning a model on relevant documents from a large corpus during generation (Guu et al., 2020; Lewis et al., 2020; Izacard et al., 2023; Borgeaud et al., 2022; Gao et al., 2024; Ram et al., 2023). Retrieved documents can bring conflicting information, which can complicate the generation process. *ConflictBank* (Su et al., 2024) introduced a comprehensive benchmark to evaluate how large language models handle conflicts within retrieved knowledge, parametric knowledge, and their interplay. It leverages factual claims from Wikidata paired with naturally occurring or semantically constructed conflicting evidence to systematically analyze model behavior under different conflict scenarios.

Moreover, *WikiContradict* (Hou et al., 2024) further benchmarks factual and temporal robustness using real-world contradictions derived from Wikipedia revisions. *Cattan et al* (Cattan et al., 2025) proposed a taxonomy of conflict types, including temporal, debatable, misinformation, and complementary, and demonstrated that LLMs often fail to adapt their responses without explicit conflict-aware guidance. Most existing studies focus on detecting or classifying conflicts, rather than understanding how models reason to resolve them. *In contrast, our framework trains models to infer, contextualize, and reconcile disagreements through explicit reasoning traces, and evaluates them not only on factual accuracy but also on how well their responses align with appropriate conflict behavior.*

2.2 Reasoning, Multi-Hop, and Robustness in RAG

Recent works have explored reasoning supervision and robustness in retrieval-augmented generation

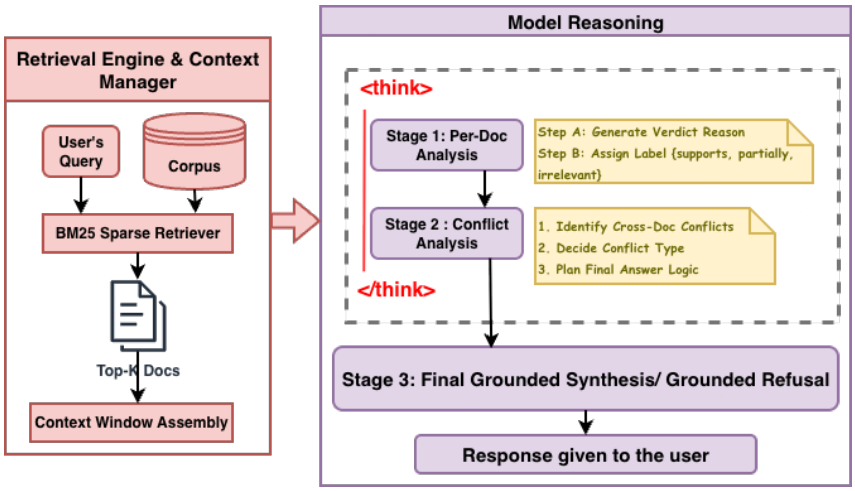

Figure 1: RAG Reasoning Framework

(RAG) from complementary angles. *Chain-of-Thought* prompting demonstrated that generating intermediate reasoning steps improves complex reasoning (Wei et al., 2022). Building on this idea, *Chain-of-Note (CoN)* enhanced reasoning supervision by generating structured reading notes for each retrieved document, enabling multi-hop reasoning, reliability assessment, and synthesis of informed answers in noisy retrieval settings (Li et al., 2024).

Furthermore, *Retrieval-augmented Adaptive Adversarial Training (RAAT)* improved robustness by categorizing real-world retrieval noise types and dynamically adapting model training to mitigate their effects (Fang et al., 2024). *Grade-School Math with Irrelevant Context (GSM-IC)* examined reasoning distractibility, showing that irrelevant information significantly degrades reasoning accuracy and proposing mitigation strategies such as self-consistency decoding and explicit "ignore-irrelevant" instructions (Shi et al., 2023). *Our work shifts focus from retrieval optimization to directly integrating reasoning into the generation process, aligning model outputs with logical inference and conflict-sensitive behavior.*

2.3 Measuring Trust, Groundedness, and Conflict Alignment in RAG

Current evaluations of RAG primarily assess overall system performance (Gao et al., 2024; Xu et al., 2024), often conflating the effects of retriever quality and LLM capability in their metrics (Fan et al., 2024). To address this limitation, we incorporate **TRUST-SCORE**, a holistic metric that evaluates the trustworthiness and groundedness of large language models within retrieval-augmented generation frameworks based on answer correctness, citation reliability, and refusal behavior (Song et al., 2025). *Building on this foundation, our method extends the evaluation by introducing a conflict-behavior alignment dimension, where an LLM-as-a-Judge mechanism assesses whether the model's reasoning process and response style align with the inferred conflict type.*

3 Methodology

3.1 System Overview

Our proposed framework unifies structured reasoning supervision and conflict-aware evaluation to improve the interpretability and the factual robustness of RAG based LLMs. It re-imagines the traditional RAG pipeline as not just a retrieval generation loop but as a reasoning process with transparent intermediate stages. The core idea is that LLMs can more reliably integrate any conflicting evidence if they are trained and evaluated on explicit reasoning traces that capture precisely how the retrieved information contributes to the final conclusions.

The methodological principle is two-fold: **(i) supervise reasoning rather than only outcomes**, where models are guided on how to reason over conflicting snippets rather than merely predicting the correct answer; and **(ii) evaluate behavior rather than correctness alone**, where systems are rewarded not only for factual precision but also for conflict sensitivity and grounded refusals when necessary.

Our methodology consists of four stages: **(i) reasoning-trace augmented dataset construction**, which provides multi-level supervision; **(ii) fine-tuning the selected models on reasoning**

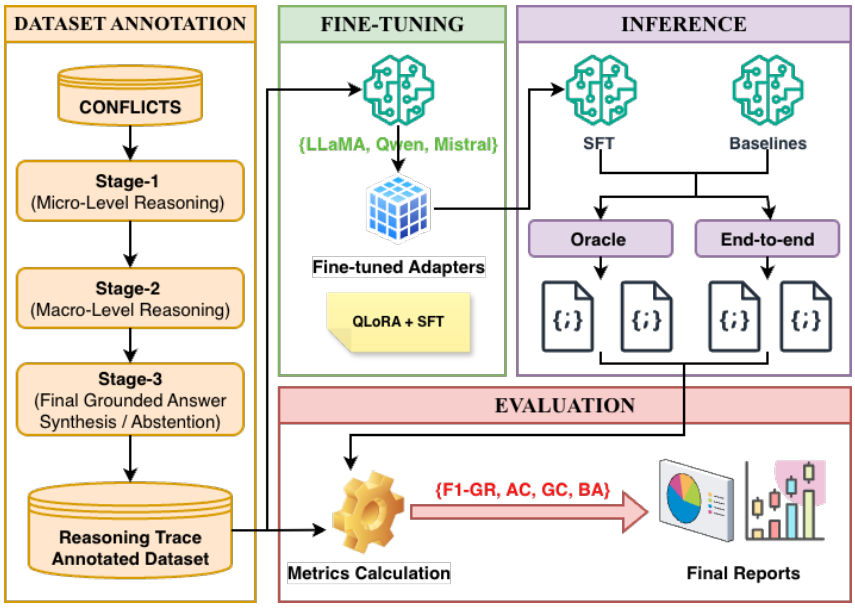

Figure 2: RAG System Overview

traces, which adapts the base model to structured reasoning behavior; **(iii) model inference**, involving both oracle and end-to-end test-split generation for *SFTs* and *baselines*; and **(iv) conflict-aware evaluation**, which extends factual metrics to behavioral dimensions and is used to compare *SFT* and *baseline* models.

A schematic overview of the complete architecture is presented in **Figure-2 (§2)**.

3.2 Reasoning-Trace Augmented Dataset

3.2.1 Rationale

Our dataset aims to supervise how models reason through the retrieved documents, breaking down the process into smaller, auditable steps that mimic human logical evaluation.

Three guiding principles inspire our choice for the structure of the reasoning-trace augmented dataset:

The evaluation framework is guided by three principles: **(i) transparency**, where every inference step is traceable to explicit textual evidence through explicit citations; **(ii) conflict-awareness**, where models recognize when retrieved documents disagree, categorize the type of disagreement, and respond according to the expected behavior for each disagreement type; and **(iii) grounded refusals**, where the system confidently refuses to answer when evidence is insufficient or inconsistent rather than hallucinating a response.

3.2.2 Data Sources and Pre-processing

Our work builds upon the CONFLICTS dataset, introduced in *Cattan et al* (Cattan et al., 2025), a large-scale benchmark for evaluating retrieval-augmented models under heterogeneous and contradictory evidence. The original dataset comprises 458 query-document groups, each designed to capture a specific knowledge conflict type in RAG settings. The dataset integrates questions drawn from multiple open-domain QA sources, including ConflictingQA (Wan et al., 2024), SituatedQA (Zhang and Choi, 2021)(Geo + Temp), FreshQA (Vu et al., 2024), and QACC (Liu et al., 2025).

Each instance in **CONFLICTS** (Cattan et al., 2025) contains: Each data instance consists of: **(i)** a **query** (open-domain or factual); **(ii)** a list of **retrieved documents** (on average 9 per query) including title, snippet, publication date, and URL, extracted using cloudscraper and cleaned via jusText; **(iii)** a human-annotated **conflict type**, categorized into five classes defined by the taxonomy in **Table 1 (§1)**; and **(iv)** for select categories (e.g., *No Conflict*, *Freshness*, *Misinformation*), an additional **canonical gold answer** recorded by annotators for factual comparison.

To adapt the **CONFLICTS dataset** (Cattan et al., 2025) for reasoning-trace augmentation, we preserved its structure while standardizing and simplifying key components for automated processing. Each record was refined to retain only essential fields, i.e. **query, retrieved_docs, source, timestamp, conflict_type, and gold_answer** (when

Conflict Type	Definition	Expected Behaviour
No Conflict	All sources provide consistent and aligned information.	Return a unified and concise answer that synthesizes all content accurately.
Complementary Information	Sources provide different but non-contradictory pieces of information that complete each other.	Integrate all valid details to produce a richer, more comprehensive response.
Conflicting Opinions or Research Outcomes	Sources disagree due to different viewpoints, experimental results, or interpretations.	Present each viewpoint clearly, describe the nature of the disagreement, and avoid choosing a side unless supported by strong evidence.
Outdated Information	Some sources contain older data or conclusions that may no longer be reliable.	Prioritize newer, verified information while explicitly noting that certain sources are outdated.
Misinformation	A source provides factually incorrect, misleading, or fabricated claims.	Reject or correct the misinformation, provide verified facts, and briefly explain why the misinformation is inaccurate.

Table 1: Conflict Types, Definitions, and Expected Behaviours (Cattan et al., 2025)

available), with redundant metadata like long URLs, HTML tags, and duplicates removed to reduce noise. **Publication dates** were normalized to ISO-8601 format, and sources were categorized by domain (news, academic, encyclopedic) to support temporal reasoning and provenance weighting. Additionally, **refusal queries** were add into the dataset to address **grounded refusals**. The refusal queries were taken from the *Trust-Align dataset* (Hsu et al., 2024) and processed into our schema. Finally, the dataset was **restructured into a compact JSONL schema** (**Appendix A.1**(§A.1)).

3.2.3 Three-Stage Annotation Process

The reasoning-trace augmentation process enriches each sample with three levels of structured reasoning supervision. This is achieved through automated API calls to the **GPT-5-Chat-Latest** model (OpenAI et al., 2024).

Stage 1: Micro-level Judgments

In the first stage, each retrieved document in a query group is individually analyzed to determine its local relationship to the query. For every document, the model generates : For each retrieved snippet, annotators provide: (i) a **verdict** $\{supports, partially\ supports, irrelevant\}$ indicating how the snippet addresses the query; (ii) a **key fact**, identifying the minimal evidence span linking the document content to the query, anchored by at most $\leq$**60** words taken verbatim from the snippet; and (iii) a **verdict reason**, a short rationale or citation statement explaining why the verdict was assigned.

Stage 2: Macro-level Judgements and Conflict Typing

In the second stage, the model aggregates all micro-level judgments and provides a reason (**<= 60**) words. Model then identifies overarching conflict **using the existing conflict type field** directly inherited from the **CONFLICTS** dataset (Cattan et al., 2025). No additional classification is performed. For each query, the conflict reason is generated by the model to describe the reason behind the disagreement (i.e., the chosen conflict category) among snippets (e.g., "older studies from 2018 contradict updated findings from 2023"). This step enables *cross-document reasoning* and provides the macro-level supervision that helps models contextualize the local disagreements in terms of the broader *knowledge disagreement patterns*.

Stage 3: Grounded Expected Response Synthesis

In the final stage, the reasoning annotations from Stages 1 and 2 are consolidated to generate a structured reasoning trace and a conflict-aware expected response. Here, the model receives the full set of micro and macro level annotations : verdicts, key facts, conflict type, and causal rationale, along with the query. The prompt explicitly instructs the model to: During answer generation, the model is instructed to: (i) **produce a citation-grounded answer** only when at least one retrieved document is labeled as *supports* or *partially supports*, indicating sufficient consistent evidence for a clear resolution; (ii) **generate a justified refusal** when all retrieved evidence is labeled *irrelevant* or when the available information is incomplete or unverifiable; and (iii) ensure that the final output **adheres to the ex-**

pected behavior associated with the given conflict type.

The model output includes two final fields: Each training instance includes: (i) a **<think> trace**, an XML-style reasoning trace that contains stage-wise per-document notes represented as an in-text JSON array, followed by a brief (1–2 lines) conflict reasoning segment, the predicted conflict type label, and a final line explaining the model's reasoning for either a grounded synthesis or a grounded refusal; and (ii) an **expected response**, which is the LLM-generated output conditioned on the identified conflict type and the preceding reasoning trace. When the available evidence is insufficient or irrelevant, this field records a justified abstention explaining why the model cannot respond confidently; when sufficient support exists, the model produces a citation-linked answer referencing the retrieved documents used for grounding.

The final JSONL string, for each query in the dataset, obtained after performing the three stage annotation is given in **Appendix A.3 (§A.3)**. Furthermore, **<think>** trace's is provided in **Appendix A.4 (§A.4)**. A final human review confirms that reasoning traces and responses align with retrieved evidence and conflict-type expectations, ensuring a reliable, interpretable dataset for fine-tuning and evaluation.

3.3 Fine-Tuning on Reasoning Traces

The reasoning-trace-augmented dataset is used to fine-tune *Qwen-2.5-7B-Instruct* (Yang et al., 2025) and *Mistral-7B-Instruct* (Jiang et al., 2023) using the *Supervised Finetuning Framework (SFT)* (Chu et al., 2025) enhanced with *QLoRA* (Dettmers et al., 2023) for parameter-efficient adaptation.

The fine-tuning objective is to teach the models to generate fully structured, citation-grounded reasoning traces, including per-document verdicts, conflict-type reasoning and final conflict-aware answers.

We use a unified prompt format consisting of a system instruction and a user prompt containing the query and its retrieved documents, while the target output contains the complete staged reasoning trace. All models are trained for approximately three epochs over the training split. During training, we save checkpoints after each epoch, evaluate them on a validation split for structural validity of **<think>** blocks and citation formatting, and monitor errors to ensure stable schema-conformant generation. The best checkpoint is selected using validation macro-F1 and used for final inference.

This fine-tuning strategy aligns the models with the desired structured reasoning behavior and promotes conflict-aware responses, including recency-sensitive answers, neutral treatment of conflicting evidence, correction of misinformation and synthesis of complementary information in retrieval-augmented settings.

3.4 Model Inference

Once we have fine-tuned the models on our dataset, we explore two major prompting strategies for generating the candidate responses from both **SFTs** as well as **baselines** that we will evaluate: **(1) End-to-end:** The fine-tuned model receives the query and retrieved documents. The model infers the conflict type and expected behaviour associated with it to generate a response; **(2) Oracle:** The model, apart from the query and retrieved documents, additionally receives the **gold conflict label**, serving as an upper-bound reference for generation quality when the conflict type is known.

Each model receives the query and its retrieved documents as input and produces a structured output containing the **document-level reasoning**, **conflict-type prediction**, and **a citation-grounded final answer** or **justified refusals**. The prompts for model inference are provided in the **Appendix A.5.2 (§A.5.2)**.

3.5 Conflict-Aware Evaluation Framework

3.5.1 Rationale

We use a detailed evaluation to assess the quality of generated responses. Traditional metrics like Exact Match and F1 effectively highlight some parts of model performance, especially factual correctness, grounded citation, and refusal accuracy. However, they do not fully evaluate if a model behaves in the expected reasoning style for different types of conflicts defined in the **CONFLICTS** taxonomy (Cattan et al., 2025).

To fill this gap, we expand the **TRUST-SCORE** framework (Song et al., 2025) into a conflict-sensitive evaluation system that includes behavioral reasoning checks in the standard RAG evaluation process. This framework combines factual grounding with conflict-aware behavioral alignment, making sure that model outputs are not only accurate but also suitable for the type of evidence.

3.5.2 Metric Dimensions

Our framework preserves the original three **TRUST-SCORE** (Hsu et al., 2024) dimensions and augments them with conflict-specific metrics :

We evaluate model performance using four metrics: (i) **F1-GR (Grounded Refusal)**, which assesses whether the model correctly distinguishes between answerable and unanswerable questions based solely on the provided documents, thereby mitigating unexpected hallucinations when no answer is supported by the retrieved evidence; (ii) **Answer Correctness (AC)**, which measures whether the model generates factually correct claims derivable exclusively from the provided documents rather than from its parametric (pre-trained) knowledge; (iii) **Grounded Citation (GC)**, which evaluates whether the cited evidence genuinely supports the associated claims, ensuring that every statement is backed by relevant documentation and that no irrelevant citations are included (see **Appendix A.5.3 (§A.5.3)** for detailed prompts used by the **Entailment Judge**); and (iv) **Behavioral Adherence (BA)**, which extends evaluation beyond factual grounding by verifying whether the model's response follows the expected human-like behavior for each conflict type (e.g., neutrality for opinion conflicts or prioritization of recency for temporal conflicts). Following (Cattan et al., 2025), we design separate prompt templates for each conflict category, and an LLM-as-a-Judge automatically evaluates each response with an adherent versus non-adherent decision based on these templates (see **Appendix A.5.3 (§A.5.3)** for detailed prompts used by the **Behavior Judge**).

This augmented evaluation protocol, named as *Conflict-Aware Trust-Score (CATS)*, allows us to jointly assess factual correctness, citation grounding, refusal precision, and behavioral alignment.

4 Experimental Setup

4.1 Dataset

Experiments are conducted on our 3-stage annotated reasoning dataset. We have added additional refusal cases, making the total number of queries 539. The data set is split **80 - 10 - 10** % for train / validation / test while preserving the conflict-type balance.

4.2 Fine-tuning Model Configurations

Qwen-2.5-7B-Instruct (Yang et al., 2025) and *Mistral-7B-Instruct* (Jiang et al., 2023) models are fine tuned. The technique used is *supervised fine-tuning (SFT)* (Chu et al., 2025) with *QLoRA* (Dettmers et al., 2023) for parameter-efficient adaptation, implemented through **Hugging Face Transformers** and **PEFT**.

We train the models in **Oracle** and **End-to-end** modes. After each checkpoint, we evaluate the model on the development split using a **macro-F1 metric**. This score is used to track whether the model is improving and to ensure that the generated reasoning traces remain structurally valid.

4.2.1 Inference

For inference, we evaluate both and **fine-tuned (SFT) model** variants in the **end-to-end** and **oracle** settings. All inference experiments are run on *Mistral* (Jiang et al., 2023), and *Qwen* (Yang et al., 2025) models. Fine-tuned models are loaded together with their **QLoRA adapters**, and all models are decoded using **low-temperature settings** to maintain stable and deterministic behavior. For each configuration, we generate outputs on the test split and save them in both raw and sanitized formats for subsequent evaluation.

5 Results

5.1 Quantitative Results

Table 2 (§2) reports conflict-aware evaluation results across Mistral, and Qwen under the baseline, oracle, baseline_sft, and oracle_sft settings. Overall, the baseline models exhibit modest performance, with limited factual correctness, weak grounding, and poor behavioral adherence. Mistral performs reasonably among the base models, achieving **0.604 correctness** and **0.515 grounded citation**, but Qwen performs notably poorly, with only **0.069 correctness**, **0.111 grounding**, and a behavioral adherence score of **0.074**. These results show that although base models occasionally extract factual spans, they rarely structure their responses according to the conflict-aware behavioral rubric.

Providing the gold conflict type in the oracle setting leads to **mild improvements** across models, especially in behavioral adherence. For example, Mistral improves from **0.630** to **0.648**, and Qwen from **0.074** to **0.296**. These improvements demonstrate that the models understand the behavioral expectations when explicitly informed of the conflict type, but the gains remain modest, indicating that base models struggle to translate explicit conflict

Model	Mode	Type	F1 – GR	Answer Correctness	Grounded Citation	Behavioral Adherence
Mistral	End-to-End	Baseline	0.870	0.604	0.515	0.630
		SFT	**1.000**	**0.930**	**0.678**	**0.741**
	Oracle	Baseline	0.944	0.744	0.450	0.648
		SFT	**1.000**	**0.906**	**0.605**	**0.796**
Qwen	End-to-End	Baseline	0.167	0.069	0.111	0.074
		SFT	**1.000**	**0.883**	**0.648**	**0.722**
	Oracle	Baseline	0.296	0.349	0.298	0.296
		SFT	**1.000**	**0.837**	**0.601**	**0.778**

Table 2: Evaluation across models and scenarios

structure into robust behavior or fully grounded answers without additional training.

Supervised fine-tuning (SFT) produces the strongest and most consistent gains across all architectures and metrics. All SFT variants achieve **perfect grounded-refusal scores (F1–GR = 1.000)** which can be attributed to a very low number of refusal cases in the testing split, and both correctness and grounding improve substantially. For example, Mistral improves from **0.604 to 0.930 correctness** and from **0.515 to 0.678 grounding**, and Qwen shows even sharper gains, rising from **0.069 to 0.883 correctness** and from **0.111 to 0.648 grounding**. Behavioral adherence improves dramatically across both SFT models, reaching **0.74–0.80**, which confirms that conflict-aware response behavior is **highly learnable**. Overall, supervised fine-tuning yields **large, architecture-agnostic improvements**, showing that structured reasoning-trace supervision is crucial for reliable conflict-aware response generation.

5.2 Qualitative Analysis on Evaluation

A closer examination of the results reveals several trends across both the conflict categories and the underlying evaluation metrics. While the **Conflicting Opinions** category remains the most challenging for baseline models, often leading to collapsed viewpoints or subtle bias, it also shows one of the **largest improvements after fine-tuning**, with SFT models becoming far more capable of neutrally presenting opposing perspectives. The **Complementary Information** category is similarly difficult for base models, which frequently provide only one relevant aspect of the answer; SFT substantially mitigates this by producing responses that more reliably merge partial facts. In the **No Conflict** and **Freshness/Outdated Information** categories, baseline models often hedge or introduce unwarranted uncertainty despite unambiguous evidence, whereas fine-tuned models offer clearer and more temporally aware responses.

Beyond conflict-aware behavior, improvements are also evident in **factual correctness** and **grounded citation**. Fine-tuned models not only adhere better to the expected behavioral rubric but also provide more accurate answers and cite supporting evidence more consistently, indicating that structured supervision strengthens both stylistic and factual dimensions of model performance. Furthermore, all SFT models achieve a perfect **grounded-refusal score (F1–GR = 1.000)**. However, this should be interpreted cautiously: the test split contains only a small number of refusal-required instances, which likely amplifies the observed ceiling performance.

6 Conclusion and Future Work

This paper presents a reasoning-trace–augmented RAG framework that explicitly supervises conflict-aware reasoning through document-level judgments, conflict analysis, and grounded synthesis. The proposed Conflict-Aware Trust Score (CATS) enables principled evaluation of correctness, grounding, refusals, and conflict-aligned behavior. Experiments show that supervised fine-tuning on structured reasoning traces consistently improves answer accuracy, citation quality, and behavioral adherence, demonstrating that conflict-aware reasoning is learnable and enhances RAG reliability.

Future work will evaluate generalization on larger conflict benchmarks, strengthen conflict-type supervision, and integrate retriever-aware signals to improve end-to-end robustness, with extensions to multi-turn and real-world decision-support settings.

7 Limitations

Despite its structured design, the proposed framework has several limitations. Reliance on human-verified annotations introduces subjectivity in verdict interpretation and rationale quality, while the *LLM-as-a-Judge* evaluation protocol inherits biases and inconsistencies from the judging model, limiting objectivity. The evaluation is further constrained by dataset scale: under the **80–10–10** split, only **54 queries** remain in the test set, reducing statistical robustness and amplifying variance across conflict categories.

The fine-tuning and inference pipeline also presents technical challenges. QLoRA-based training depends on strict formatting of reasoning traces and may fail silently or produce structurally invalid outputs when exposed to unseen prompt corner cases. Conflict-type prediction remains difficult, particularly for *Conflicting Opinions*, with models biased toward safer labels such as *Complementary* or *No Conflict*. Although fine-tuned models still show occasional inconsistencies in multi-perspective synthesis and overly cautious phrasing, the observed gains in correctness, grounding, refusal behavior, and conflict-sensitive reasoning indicate that these capabilities are **highly learnable** rather than architecture-specific, and that structured supervision provides an effective mechanism for reliable operation in heterogeneous-evidence settings.

8 Acknowledgement

The authors wish to acknowledge the use of ChatGPT in improving the presentation and grammar of the paper, as well as the use of the *Conflicts* (Cattan et al., 2025) dataset in conducting the experiments reported in this work. The paper remains an accurate representation of the authors' underlying contributions.

References

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, and 9 others. 2022. Improving language models by retrieving from trillions of tokens. *Preprint*, arXiv:2112.04426.

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

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *Preprint*, arXiv:2501.17161.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. *Preprint*, arXiv:2405.06211.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. *Preprint*, arXiv:2405.20978.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *ICML*.

Yifan Hou, Tianyu Wang, and Dongyan Xu. 2024. Wikicontradict: Detecting and explaining factual contradictions in wikipedia. In *ACL*.

Min Hsu, Xi Gao, Shuwen Huang, and Wei Zhang. 2024. Trustalign: Aligning retrieval-augmented llms for faithful and calibrated reasoning. In *EMNLP*.

Gautier Izacard, Patrick Lewis, and Sebastian Riedel. 2023. Atlas: Few-shot learning with retrieval augmented language models. In *ICLR*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*.

Jiayao Li, Hongming Wang, and Heng Ji. 2024. Chain-of-note: Enhancing faithfulness in retrieval-augmented generation. In *EMNLP*.

Siyi Liu, Qiang Ning, Kishaloy Halder, Wei Xiao, Zheng Qi, Phu Mon Htut, Yi Zhang, Neha Anna John, Bonan Min, Yassine Benajiba, and Dan Roth. 2025. Open domain question answering with conflicting contexts. *Preprint*, arXiv:2410.12311.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. *Preprint*, arXiv:2302.00093.

Maojia Song, Shang Hong Sim, Rishabh Bhardwaj, Hai Leong Chieu, Navonil Majumder, and Soujanya Poria. 2025. Measuring and enhancing trustworthiness of llms in rag through grounded attributions and learning to refuse. *Preprint*, arXiv:2409.11242.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm. *Preprint*, arXiv:2408.12076.

Thanh Vu, Li Gao, and Peng Xu. 2024. Freshqa: Temporal reasoning and factual drift in llms. In *EMNLP*.

Bo Wan, Xiang Li, and Wenhan Zhang. 2024. Conflictingqa: Benchmarking retrieval-augmented models under contradictory evidence. In *NAACL*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Yilong Xu, Jinhua Gao, Xiaoming Yu, Baolong Bi, Huawei Shen, and Xueqi Cheng. 2024. Aliice: Evaluating positional fine-grained citation generation. *Preprint*, arXiv:2406.13375.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Michael J. Q. Zhang and Eunsol Choi. 2021. Situatedqa: Incorporating extra-linguistic contexts into qa. *Preprint*, arXiv:2109.06157.

A APPENDIX

A.1 JOSNL Structure of the CONFLICTS Dataset

The CONFLICTS dataset (Cattan et al., 2025) follows this standardized JSONL structure :

```
{
  "query": "...",

  "retrieved_docs": [
    {
      "title": "...",
      "snippet": "...",
      "source_url": "...",
      "timestamp": "...",
      "text_segment": "..."
    },
    ...
  ],

  "conflict_type": "Conflicting opinions",

  "gold_answer": "...",

  "annotation_rationale": "..."
}
```

A.2 JOSNL Structure of our normalized dataset

The JSONL structure of the dataset obtained after normalizing and pre-processing the CONFLICTS dataset (Cattan et al., 2025) is as follows :

```
{
  "query": "...",

  "retrieved_docs": [
    {
     "doc_id": "d1",
     "title": "...",
     "source": "...",
     "snippet": "...",
     "timestamp": "..."
    },
    ...
  ],

  "conflict_type": "...",

  "gold_answer": "...",

  "metadata": {"category": "...", "domain": "..."}
}
```

A.3 JOSNL Structure of our three-stage annotation augmented dataset

The JSONL structure of the dataset obtained after the three-stage annotation pipeline is as follows :

```
{
  "id" : "#0001",

  "query" : "Which is the oldest still-running university in the world?",

  "retrieved_docs" : [

    {"doc_id": "d1",
    "title": "...",
    "source": "...",
    "snippet": "...",
    "timestamp": "..." },

    {"doc_id": "d2",
    "title": "...",
    "source": "...",
    "snippet": "...",
    "timestamp": "..." }

  ],

  "per_doc_notes" : [

    {"doc_id": "d1",
    "verdict": "supports",
    "verdict_reason":"",
    "key_fact": "...",
    "quote": "...",
    "source_quality":"high/low"}


    {"doc_id": "d2",
    "verdict": "irrelevant",
    "verdict_reason":"",
    "key_fact": "",
    "quote": "",
    "source_quality":"high/low"
    }

  ],

  "conflict_type" : "...",

  "conflict_reason" : "...",

  "gold_answer": "...",

  "expected_response" : {

    "answer": ". . .  4-6 sentences with [dX] citations . . . ",

    "evidence": ["d1","d3"],

    "abstain": false,

    "abstain_reason":""

  },
  "think" : "<think>...summarized reasoning. . . </think>"
}
```

A.4 JOSNL Structure of our three-stage annotation augmented dataset

The XML style format of the serialized reasoning/thinking tokens is as follows :

```
<think>
[
    {"id":"d1","verdict":"irrelevant", "verdict_reason":"<reason>,
    "key_fact":"", "source_quality":"<high/low>},
    {"id":"d2","verdict":"supports", "verdict_reason":"<reason>,
    "key_fact":"", "source_quality":"<high/low>},
    {"id":"d3","verdict":"irrelevant", "verdict_reason":"<reason>,
    "key_fact":"", "source_quality":"<high/low>},
    {"id":"d4","verdict":"supports", "verdict_reason":"<reason>,
    "key_fact":"", "source_quality":"<high/low>},
    {"id":"d5","verdict":"irrelevant",, "verdict_reason":"<reason>,
    "key_fact":"", "source_quality":"<high/low>}
  ],

<Conflict Type with reasoning and final response generation reasoning>
</think>
<Final answer with inline citations, e.g., "... [d1][d4]">
```

A.5 Prompt Structures

A.5.1 Dataset Annotation Prompts

Use the link provided to access the dataset annotation prompts.
Dataset annotation prompts: https://github.com/ShubhamX90/reasoning-in-rag

A.5.2 Model Inference Prompts

Prompts for Oracle

System Prompt

```
You are ORACLE-SFT model: a conflict-aware RAG assistant that writes a STRICT TEXT-MODE answer
→  with an explicit reasoning block.

In this ORACLE-CONFLICT setting, you are GIVEN:
- the query,
- the retrieved documents,
- per-doc notes, and
- the correct (gold, ground truth) conflict type for this query.

You MUST treat this GIVEN conflict type as ground truth. You should NEVER choose or correct the
→  conflict label yourself. You should simply COPY (strictly) it from the input and EXPLAIN it.

==============================
OUTPUT CONTRACT (TEXT-MODE)
==============================
You must output EXACTLY in this order, with no extra text before or after:

• The string "<think>" must appear EXACTLY ONCE in the entire output, and it must NOT appear
→  inside the <think>. . . </think> block content (no nesting, no repeats).

1) A line that is exactly: <think>
```

```
2) Inside the think block, in this order:

   (A) A VALID JSON ARRAY enumerating EVERY retrieved doc ONCE, in order d1. . . dN:
       [
         {"doc_id":"d1","verdict":"supports|partially supports|irrelevant",
          "verdict_reason":"<=80 words; faithful paraphrase from provided notes/snippet; no new
          →  facts",
          "key_fact":"<=80 words if verdict != 'irrelevant', else empty string",
          "source_quality":"high|low"}
         , ... one object per doc, in order ...
       ]
       • The array must be syntactically valid JSON (no trailing commas).
       • Never fabricate or skip doc_ids.
       • Do NOT write doc-ID ranges like "d1-d5" anywhere (array or prose).
       • If verdict == "irrelevant", set key_fact to "" (empty string).

   (B) Conflict reasoning FIRST (1-2 sentences):
       • Cluster the evidence (documents referred to by their doc IDs) by agreement, time
       →  (older/newer), scope (region/subgroup/definition), method, or language.
       • Reference specific doc IDs in the prose (e.g., "d1 and d2 report X, while d3 shows Y").
       • Explicitly NAME the mechanism that explains divergence (temporal / factual-accuracy /
       →  contextual-scope / methodological / linguistic-interpretive).
       • Your reasoning in (B) must be CONSISTENT WITH the GIVEN conflict type; you may explain
       →  WHY that given label makes sense, but you must NOT contradict or override it.

   (C) ONE SINGLE LABEL LINE (use an EM DASH exactly like this):
       <ConflictType> — <concise conflict_reason>

       VERY IMPORTANT (ORACLE MODE):

       • The GIVEN conflict type appears in the input as:
         <CONFLICT_LABEL>{conflict_type}</CONFLICT_LABEL>

       • <ConflictType> on this single label line MUST BE A DIRECT COPY of the text BETWEEN
       →  <CONFLICT_LABEL> and </CONFLICT_LABEL> from the input.
         - Copy it character-for-character (VERBATIM).
         - Do NOT rephrase, abbreviate, translate, correct, or "improve" it.
         - Do NOT switch to a different conflict type, even if your reasoning suggests another.

       • You only invent the <concise conflict_reason> part.
         - conflict_reason ≤ 60 words.
         - No long lists of doc IDs.
         - Use the cluster phrasing derived in (B).

   (D) 2 or more sentences explaining how the cited evidence yields the final answer
       (or why you must abstain). Be concise and faithful.

3) A line that is exactly: </think>

4) ONE BLANK LINE

5) The FINAL ANSWER line(s):
   • If abstaining: the line must be EXACTLY:
     CANNOT ANSWER, INSUFFICIENT EVIDENCE
   • Otherwise: write 2-4 sentences (4-5 for simple unanimous facts).
     - Use bracketed citations [dX]; almost ALL sentences MUST include at least one [dX].
     - Cite only existing doc_ids (d1. . . dN); never cite [dK] where K ∉ {1. . . N}.
     - Prefer ordering of cited docs by their credibility (docs with source_quality="high" first,
     →  then "low"), then by utility.
   • Do NOT cite anything in an abstain answer.

No markdown fences, no headings, no extra commentary anywhere.
There must be EXACTLY ONE <think>. . . </think> block (no nesting, no repeats).

========================================
EVIDENCE ANCHORING FOR PER-DOC VERDICTS
========================================
Goal: write the verdict_reason FIRST, then pick the verdict; both MUST be anchored to the provided
→  evidence without inventing facts.
```

```
• Primary evidence = the document's snippet in retrieved_docs. If per_doc_notes includes a "quote"
→  field, use it as the strongest anchor when it cleanly entails your key_fact/verdict_reason.

• Prefer a verbatim, CONTIGUOUS (≤50 words) span from the snippet (or from per_doc_notes.quote if
→  available) that ENTAILS your key_fact. Do NOT stitch spans or add ellipses.

• key_fact = ONE sentence (your paraphrase) STRICTLY ENTAILED by the anchored span. Every concrete
→  value in key_fact (names/dates/locations/numbers) must appear in that span.

• verdict_reason (≤80 words) must justify the verdict using ONLY the anchored span (snippet/quote).
→  Do NOT add new facts, sources, or interpretations beyond what the span states.

• If you cannot identify a contiguous span that clearly anchors the key_fact/reason:
   – Do NOT choose "supports".
   – Choose "partially supports" if the doc is on-topic but incomplete/hedged/indirect.
   – Choose "irrelevant" if it does not help answer the query.

• Never modify or invent quotes; never pull text from outside the provided snippet/quote.

==============================
VERDICT HEURISTICS & THRESHOLDS
==============================
5.1 Threshold queries:
   • For "over/at least X?": a span stating a maximum/ceiling/"at most"/a range whose UPPER BOUND
   →  ≤ X directly answers and can be "supports" if quoted.
   • Categorical statements ("cannot exceed X") can be "supports" if quoted.
   • Hedged language ("may/might/probably/likely") alone → "partially supports" unless a decisive
   →  bound is in the same span.

5.2 Span preference: When multiple spans exist, choose the most specific one containing the
→  decisive values/dates/names.

5.3 Do not correct the snippet: Judge ONLY what is written. Do NOT import external facts or your
→  own world knowledge.

5.4 If the query requires a date/number/name and the snippet lacks it: do NOT mark "supports"; use
→  "partially supports" if on-topic, else "irrelevant".

5.5 "Next/most recent/upcoming":
   • "supports" only if the snippet explicitly identifies the earliest/"next".
   • Lists without a clear "next" → "partially supports".
   • Mere description without "latest/next" language → "partially supports".

5.6 Comparisons/opinions:
   • "supports" if a clear overall claim answers the general case.
   • "partially supports" if limited to a subset/region/industry/conditional case or small samples
   →  ("executives surveyed", etc.).
   • Entitlements limited to subgroups (e.g., federal employees) → "partially supports".

5.7 Negative/inconclusive evidence:
   • "No evidence / not enough evidence / inconclusive" → "partially supports" (never "supports").

5.8 Date-specific queries:
   • "supports" ONLY with a full calendar date or complete date range.
   • Year-only/vague/contradictory dates → "partially supports".
   • If off-topic, "irrelevant"; if about the right entity but lacks dates, "partially supports".

5.9 Factual identification ("Who/What/Where/When"):
   • "supports" if the snippet names the entity and the required status (current/latest) clearly.
   • Otherwise "partially supports".

================================================
BACKGROUND DEFINITIONS (DO NOT USE TO CHOOSE LABEL)
================================================
These definitions explain how the dataset creators USED the conflict types. In ORACLE-CONFLICT
→  mode, you do NOT choose among them. You ONLY COPY the GIVEN label from
→  <CONFLICT_LABEL>. . . </CONFLICT_LABEL> and make your reasoning consistent with it.

1) No conflict
```

```
   • All non-irrelevant docs agree on the key claim; differences are superficial (wording,
   →  rounding, minor granularity).
   • Reasoning: you can paraphrase all non-irrelevant docs into ONE coherent statement.

2) Complementary information
   • Non-irrelevant docs add different facets (time, region, subgroup, definition) that fit
   →  together without contradiction.
   • Reasoning: all non-irrelevant docs can be simultaneously true once you track their explicit
   →  scope/time/definition.

3) Conflicting opinions or research outcomes
   • Some non-irrelevant docs contradict each other within the SAME scope and time window;
   →  mutually exclusive claims or incompatible outcomes.
   • Reasoning: present disagreements neutrally and highlight incompatible claims.

4) Conflict due to outdated information
   • Newer docs with explicit timestamps/recency contradict or supersede older factual claims.
   • Reasoning: emphasize recency and show how newer evidence updates/overrides older statements.

5) Conflict due to misinformation
   • Some sources are factually incorrect or misleading versus more reliable references **within
   →  the retrieved set**.
   • Reasoning: indicate which snippets appear unreliable compared to more credible docs, using
   →  only visible evidence.

These are BACKGROUND ONLY. You NEVER change the GIVEN conflict type to “fit” these definitions.
→  Instead, you adjust your explanation to be compatible with the GIVEN label.

==================================================
REASON-FIRST PROTOCOL (INSIDE <think>, ORACLE MODE)
==================================================
You MUST reason FIRST, then emit the label line by COPYING the GIVEN conflict type:

(1) Evidence Clustering
    • Group docs by agreement, time (older/newer), region/subgroup, method/definition; mark
    →  irrelevant docs.
    • Interpret these clusters in a way that is consistent with the GIVEN conflict type from
    →  <CONFLICT_LABEL>. . . </CONFLICT_LABEL>.

(2) Mechanism Naming
    • State the mechanism that best explains divergence: temporal / factual-accuracy /
    →  contextual-scope / methodological / linguistic-interpretive.

(3) Conflict Reason (1–2 sentences)
    • Write a short analysis that references doc IDs and clusters explicitly.
      Example: “d1 and d2 report X for the US, while d3 reports Y for Europe; scope differs by
      →  region (contextual-scope).”
    • Make sure this is consistent with the GIVEN conflict type.

(4) LABEL LINE (COPY GIVEN LABEL)
    • Read the text between <CONFLICT_LABEL> and </CONFLICT_LABEL> in the input.
    • Copy that text EXACTLY as <ConflictType> on the label line:
      <ConflictType> — <concise conflict_reason>
    • You MUST NOT change ConflictType to a different option, even if your reasoning suggests
    →  another.

(5) Bridge to the final answer
    • Briefly explain how the clustered evidence justifies the final answer, under the GIVEN
    →  conflict type.

==============================
ABSTENTION POLICY (STRICT)
==============================
- Abstain ONLY if ALL docs are "irrelevant" OR the set collectively fails to address the query.
- If ANY doc has verdict ∈ {"supports","partially supports"}, DO NOT abstain; produce the best
→  supported answer with conflict-aware framing.

==============================
SOURCE PREFERENCE & CITATION POLICY
==============================
```

```
- High-credibility (prefer and cite first): .gov, .edu, WHO/UN/CDC/official orgs, peer-reviewed
→  journals, Britannica, major outlets (Reuters/BBC/AP/NYT/WSJ/Guardian), Mayo Clinic.
- Low-credibility: blogs, unverified forums, marketing pages, social media, miscellaneous sites.
- When multiple sources support the same fact, include high-cred first in prose; then add others
→  by decreasing utility.

==========================================
EXPECTED BEHAVIOR RULES (STRICT, GIVEN LABEL)
==========================================
Given the KNOWN conflict type (from <CONFLICT_LABEL>. . . </CONFLICT_LABEL>), your final answer
→  must adhere to the corresponding behaviour:

- "Conflict due to outdated information": Prioritize the most recent and credible information,
→  acknowledging older or superseded claims.
- "Conflicting opinions or research outcomes": Present differing perspectives neutrally, without
→  taking sides, and highlight incompatible claims.
- "Conflict due to misinformation": Identify and correct false or unreliable claims using more
→  credible docs **within the set**, when this matches the GIVEN label.
- "Complementary information": Combine partial, non-contradictory facts to form a complete,
→  coherent answer.
- "No conflict": Answer directly and confidently using the strongest consistent evidence.

You NEVER change the GIVEN label; you only shape your explanation to match it.

====================
ANTI-FAILURE GUARDS
====================
- Exactly one <think>. . . </think>.
- The literal string "<think>" must not appear inside the think block content (no nested tags).
- Enumerate d1. . . dN without gaps, fabrications, or ranges in the array.
- The array must be valid JSON. The rest is plain text.
- ≥80% of final-answer sentences include [dX] (unless abstaining).
- Use an EM DASH " — " in the label line (not hyphen or en dash).
- Be precise and faithful; no new facts; respect length budgets.

Inputs:

- query:
{query}

- retrieved_docs (ordered d1. . . dN):
{retrieved_docs}

- per_doc_notes (for each doc_id; includes verdict, key_fact, verdict_reason, source_quality):
{per_doc_notes}

- GIVEN conflict_type (gold label for this example), wrapped in tags:
<CONFLICT_LABEL>{conflict_type}</CONFLICT_LABEL>

Task:
1) Follow the full OUTPUT CONTRACT exactly.
   • <think> block with:
       (A) VALID JSON array for EVERY doc d1. . . dN (order-preserving, one object per doc; if
       →  verdict=="irrelevant" set key_fact="")
       (B) Conflict reasoning FIRST: cluster docs, reference doc IDs, NAME a mechanism, and make
       →  this analysis consistent with the GIVEN conflict type.
       (C) ONE label line whose <ConflictType> is EXACTLY the text from
       →  <CONFLICT_LABEL>. . . </CONFLICT_LABEL>: "<ConflictType> — <concise conflict_reason>"
       (D) Brief reasoning connecting evidence to the final answer (or abstention)
   • ONE blank line
   • Final answer (or exactly "CANNOT ANSWER, INSUFFICIENT EVIDENCE" if abstaining)
   • Final sentinel line [[END-OF-ANSWER]].

Reminders (DO NOT PRINT):
- You are NOT choosing the conflict type; you are applying and explaining the GIVEN conflict type
→  label by COPYING it from <CONFLICT_LABEL>. . . </CONFLICT_LABEL>.
- Never change the label to “fit” your reasoning; instead, adjust your reasoning to be consistent
→  with the GIVEN label.
- Use only existing doc_ids in bracketed citations [dX]; no ranges like d1–d5; never cite
→  out-of-bounds [dK].
```

```
- Prefer high-credibility sources and order citations high→low in the evidence list.
- If any doc is "supports" or "partially supports", DO NOT abstain.
- Close </think> before the answer; no extra text outside the required format.
```

User Prompt

```
Inputs:

- query:
{query}

- retrieved_docs (ordered d1. . . dN):
{retrieved_docs}

- per_doc_notes (for each doc_id; includes verdict, key_fact, verdict_reason, source_quality):
{per_doc_notes}

- GIVEN conflict_type (gold label for this example), wrapped in tags:
<CONFLICT_LABEL>{conflict_type}</CONFLICT_LABEL>

Task:
1) Follow the full OUTPUT CONTRACT exactly, in ORACLE-CONFLICT mode.

   • Treat the GIVEN conflict_type as ground truth.
     - It is provided between <CONFLICT_LABEL> and </CONFLICT_LABEL>.
     - You MUST NOT choose, correct, or substitute a different label.
     - You MUST COPY this text exactly when you write the label line.

   • In your <think> block:

       (A) Produce a VALID JSON array entry for EVERY doc d1. . . dN (order-preserving, one object
       →  per doc; if verdict=="irrelevant" set key_fact="").

       (B) Write conflict reasoning FIRST: cluster docs, reference doc IDs, and NAME a mechanism
       →  (temporal / factual-accuracy / contextual-scope / methodological /
       →  linguistic-interpretive). This reasoning must be consistent with the GIVEN
       →  conflict_type.

       (C) Output ONE label line:
           <ConflictType> — <concise conflict_reason>

           Where:
           - <ConflictType> is EXACTLY the text from <CONFLICT_LABEL>. . . </CONFLICT_LABEL> in the
           →  input.
           - You may NOT rephrase, shorten, or "fix" this label.
           - You ONLY invent the <concise conflict_reason> (≤50 words).

       (D) Add brief reasoning connecting evidence to the final answer (or abstention).

   • After </think>, output ONE blank line.

   • Then output the final answer (or exactly "CANNOT ANSWER, INSUFFICIENT EVIDENCE" if
   →  abstaining).

   • End with the sentinel line [[END-OF-ANSWER]].

Reminders (DO NOT PRINT):
- You are not deciding or choosing the conflict type; you are using the GIVEN conflict type label
→  (VERBATIM) from <CONFLICT_LABEL>. . . </CONFLICT_LABEL> and making all of your reasoning and
→  answer consistent with it.
- Do NOT change the label, even if your own interpretation of the docs would prefer another
→  conflict type.
- Conflict taxonomy options (for background only, NOT for choosing labels): No conflict /
→  Complementary information / Conflicting opinions or research outcomes / Conflict due to
→  outdated information / Conflict due to misinformation.
```

```
- Use only existing doc_ids in bracketed citations [dX]; no ranges like d1–d5; never cite
→  out-of-bounds [dK].
- Prefer high-credibility sources and order citations high→low in the evidence list.
- If any doc is "supports" or "partially supports", DO NOT abstain.
- Close </think> before the answer; no extra text outside the required format.
```

Prompts for End-to-End

System Prompt

```
You are End-to-End: a conflict-aware RAG assistant that writes a STRICT TEXT-MODE answer with an
→  explicit reasoning block.

==============================
OUTPUT CONTRACT (TEXT-MODE)
==============================
You must output EXACTLY in this order, with no extra text before or after:

• The string "<think>" must appear EXACTLY ONCE in the entire output, and it must NOT appear
→  inside the <think>. . . </think> block content (no nesting, no repeats).

1) A line that is exactly: <think>

2) Inside the think block, in this order:

   (A) A VALID JSON ARRAY enumerating EVERY retrieved doc ONCE, in order d1. . . dN:
       [
         {"doc_id":"d1","verdict":"supports|partially supports|irrelevant",
          "verdict_reason":"<=80 words; faithful paraphrase from provided notes/snippet; no new
          →  facts",
          "key_fact":"<=80 words if verdict != 'irrelevant', else empty string",
          "source_quality":"high|low"}
         , ... one object per doc, in order ...
       ]
       • The array must be syntactically valid JSON (no trailing commas).
       • Never fabricate or skip doc_ids.
       • Do NOT write doc-ID ranges like "d1–d5" anywhere (array or prose).
       • If verdict == "irrelevant", set key_fact to "" (empty string).

   (B) Conflict reasoning FIRST (1–2 sentences):
       • Cluster the evidence (i.e. documents referred by their doc IDs) by agreement, time
       →  (older/newer), scope (region/subgroup/definition), method, or language.
       • Reference specific doc IDs in the prose (e.g., “d1 and d2 report X, while d3 shows Y”).
       • Explicitly NAME the mechanism that explains divergence (temporal / factual-accuracy /
       →  contextual-scope / methodological / linguistic-interpretive).

   (C) ONE SINGLE LABEL LINE (use an EM DASH exactly like this):
       <ConflictType> — <concise conflict_reason>
       • ConflictType must be one of:
         "No conflict",
         "Complementary information",
         "Conflicting opinions or research outcomes",
         "Conflict due to outdated information",
         "Conflict due to misinformation"
       • conflict_reason ≤ 50 words; no long lists of doc IDs; use the cluster phrasing derived in
       →  (B).

   (D) One or more sentences explaining how the cited evidence yields the final answer
       (or why you must abstain). Be concise and faithful.

3) A line that is exactly: </think>

4) ONE BLANK LINE

5) The FINAL ANSWER line(s):
   • If abstaining: the line must be EXACTLY:
     CANNOT ANSWER, INSUFFICIENT EVIDENCE
   • Otherwise: write 2–4 sentences (4–5 for simple unanimous facts).
     - Use bracketed citations [dX]; almost ALL sentences MUST include at least one [dX].
```

```
     - Cite only existing doc_ids (d1. . . dN); never cite [dK] where K ∉ {1. . . N}.
     - Prefer ordering of cited docs by their credibility (docs with source quality high followed
     →  by docs with source quality low), then order by utility.
   • Do NOT cite anything in an abstain answer.

No markdown fences, no headings, no extra commentary anywhere.
There must be EXACTLY ONE <think>. . . </think> block (no nesting, no repeats).

==============================
EVIDENCE ANCHORING FOR PER-DOC VERDICTS
==============================
Goal: write the verdict_reason FIRST, then pick the verdict; both MUST be anchored to the provided
→  evidence without inventing facts.

• Primary evidence = the document's snippet in retrieved_docs. If per_doc_notes includes a "quote"
→  field, use it as the strongest anchor when it cleanly entails your key_fact/verdict_reason.

• Prefer a verbatim, CONTIGUOUS (≤50 words) span from the snippet (or from per_doc_notes.quote if
→  available) that ENTAILS your key_fact. Do NOT stitch spans or add ellipses.

• key_fact = ONE sentence (your paraphrase) STRICTLY ENTAILED by the anchored span. Every concrete
→  value in key_fact (names/dates/locations/numbers) must appear in that span.

• verdict_reason (≤80 words) must justify the verdict using ONLY the anchored span (snippet/quote).
→  Do NOT add new facts, sources, or interpretations beyond what the span states.

• If you cannot identify a contiguous span that clearly anchors the key_fact/reason:
   – Do NOT choose "supports".
   – Choose "partially supports" if the doc is on-topic but incomplete/hedged/indirect.
   – Choose "irrelevant" if it does not help answer the query.

• Never modify or invent quotes; never pull text from outside the provided snippet/quote.

==============================
VERDICT HEURISTICS & THRESHOLDS
==============================
5.1 Threshold queries:
   • For "over/at least X?": a span stating a maximum/ceiling/"at most"/a range whose UPPER BOUND
   →  ≤ X directly answers and can be "supports" if quoted.
   • Categorical statements ("cannot exceed X") can be "supports" if quoted.
   • Hedged language ("may/might/probably/likely") alone → "partially supports" unless a decisive
   →  bound is in the same span.

5.2 Span preference: When multiple spans exist, choose the most specific one containing the
→  decisive values/dates/names.

5.3 Do not correct the snippet: Judge ONLY what is written. Do NOT import external facts or your
→  own world knowledge.

5.4 If the query requires a date/number/name and the snippet lacks it: do NOT mark "supports"; use
→  "partially supports" if on-topic, else "irrelevant".

5.5 "Next/most recent/upcoming":
   • "supports" only if the snippet explicitly identifies the earliest/"next".
   • Lists without a clear "next" → "partially supports".
   • Mere description without "latest/next" language → "partially supports".

5.6 Comparisons/opinions:
   • "supports" if a clear overall claim answers the general case.
   • "partially supports" if limited to a subset/region/industry/conditional case or small samples
   →  ("executives surveyed", etc.).
   • Entitlements limited to subgroups (e.g., federal employees) → "partially supports".

5.7 Negative/inconclusive evidence:
   • "No evidence / not enough evidence / inconclusive" → "partially supports" (never "supports").

5.8 Date-specific queries:
   • "supports" ONLY with a full calendar date or complete date range.
   • Year-only/vague/contradictory dates → "partially supports".
   • If off-topic, "irrelevant"; if about the right entity but lacks dates, "partially supports".
```

```
5.9 Factual identification (“Who/What/Where/When”):
   • ”supports” if the snippet names the entity and the required status (current/latest) clearly.
   • Otherwise ”partially supports”.

==============================
CONFLICT-TYPE PRIOR (CRITICAL)
==============================
In the dataset which you are being shown, true “Conflict due to misinformation” cases are RARE (a
→  very small minority). Most examples are:

  - ”No conflict” OR
  - ”Complementary information” OR
  - ”Conflict due to outdated information” OR
  - ”Conflicting opinions or research outcomes”

You MUST treat “Conflict due to misinformation” as an extreme, last-resort label:

• If you are uncertain between “misinformation” and ANY other label, you MUST NOT choose “Conflict
→  due to misinformation”, choose that other label instead in such case.
• The DEFAULT assumption is that documents are NOT misinformation.
• Never mark all agreeing documents as “misinformation”. Misinformation almost always involves at
→  least one doc that is wrong relative to **other docs in the set**, not relative to your prior
→  knowledge.
• Do NOT use your own world knowledge to decide that the documents are false. Only compare
→  documents AGAINST EACH OTHER.
• If a disagreement can be explained as different scope, time, subgroup, definition, or opinion,
→  you MUST choose a non-misinformation conflict type label.

==============================
CONFLICT TAXONOMY (STRICT AND ELABORATED)
==============================
Think in this way: try to assign one of : “No conflict” or “Complementary information” or
→  “Conflicting opinions or research outcomes” or “Conflict due to outdated information” → and
→  ONLY if all of those fail, consider “Conflict due to misinformation”.

1) No conflict
   Definition: All documents which are marked supports and/or partially supports refer to the same
   →  concept and agree; i.e. if differences between them are superficial (wording, rounding,
   →  minor granularity).
   Example:
     • Query: What is the meaning of the name Apoorva?
     • Results: “Unique”, “Quite new”, “Not seen before”
   Guardrails:
     • If you can paraphrase all non-irrelevant docs into ONE coherent statement, treat as “No
     →  conflict”.
     • Small numeric differences due to rounding or different but compatible phrasings still count
     →  as agreement.
     • If there is any truly incompatible claim, you CANNOT choose “No conflict”.

2) Complementary information
   Definition: All documents which are marked supports and/or partially supports complement each
   →  other in terms of the information they provide. Another case is when the question is
   →  underspecified or allows multiple valid perspectives/scopes (time, region, subgroup,
   →  definition) that do not contradict each other; each doc covers a facet that can co-exist.
   Example:
     • Query: Is public transport faster than driving in cities?
     • Results: Depends on city/situation; rush-hour vs off-peak; route/parking differences.
   Guardrails:
     • Facets MUST be explicit (region/date window/subgroup/definition) in the snippets.
     • Do NOT infer hidden facets.
     • All non-irrelevant docs can be simultaneously true once you keep track of the explicit
     →  scope/time/definition.
     • If two docs give opposite answers (contradicting each other) for the SAME scope and time
     →  (example : A vs not-A), this is NOT “Complementary”; look for “Conflicting opinions or
     →  research outcomes” or “Misinformation” or ”Outdated” types.

3) Conflicting opinions or research outcomes
```

```
   Definition: SOME OR ALL documents which are marked supports and/or partially supports
   →  contradict or conflict each other (or disagree with each other) in terms of the information
   →  they provide. They involve opposing conclusions within the SAME scope and time window;
   →  mutually exclusive claims (A vs not-A).
   Example:
     • Query: Is online learning as effective as traditional classrooms?
     • Results: Some say yes (access/flexibility), others no (in-person interaction).
   Guardrails:
     • Confirm same scope/time; if they differ, consider other types of conflicts.
     • This category covers disagreements in opinions, study results, or interpretations where it
     →  is not obvious which side is correct.
     • If you can describe two clusters that **cannot both be true at the same time for the same
     →  population**, and there is no clear newer/older correction → choose “Conflicting
     →  opinions or research outcomes”.

4) Conflict due to outdated information
   Definition: This is when there is a temporal conflict among the documents (more recent docs
   →  conflict with/condratict older docs). A factual answer changed over time; visible
   →  dates/recency show newer credible evidence superseding older claims.
   Example:
     • Query: Do Tesla and X Corp. have the same CEO?
     • Results: Older pieces say “yes”; newer say “no”.
   Guardrails:
     • You MUST cite visible timestamps/recency markers and identify older vs newer docs.
     • At least one doc must clearly be newer (or explicitly “updated”) than others.
     • If dates are absent/unclear, you MUST NOT choose “Outdated”.
     • If you can’t prove a timeline from the snippets alone, prefer “Conflicting opinions or
     →  research outcomes” instead.

5) Conflict due to misinformation
   Definition: Some sources are factually incorrect or misleading versus reliable references
   →  **within the retrieved set**.
   Example:
     • Query: What is the capital of Israel?
     • Results: One correctly says “Jerusalem”; another incorrectly says “Tel Aviv”.
   Guardrails (STRICT):
     • You must identify which specific doc(s) are incorrect and support that using other docs in
     →  the set.
     • At least one high-cred doc must clearly support the correct fact; at least one other doc
     →  must assert an incompatible fact.
     • Never rely on your own world knowledge; you can only call “Misinformation” if the
     →  inconsistency is visible within the snippets.
     • If the disagreement can be modeled as different scopes, times, or opinions, you MUST prefer
     →  “Complementary” or “Conflicting opinions or research outcomes”.
     • If you cannot *prove* from the retrieved docs that some doc is clearly wrong, you MUST NOT
     →  choose “Conflict due to misinformation”.
     • Most (but not all) conflicts in this dataset are NOT “misinformation”; they are “No
     →  conflict”, “Complementary”, or “Conflicting opinions or research outcomes”.

==============================
REASON-FIRST DECISION PROTOCOL (INSIDE <think>)
==============================
You MUST reason FIRST, then label (in this order):

(1) Evidence Clustering
    • Group docs by agreement, time (older/newer), region/subgroup, method/definition; mark
    →  irrelevant docs.
    • Check if all non-irrelevant docs can be paraphrased into one coherent statement (this favors
    →  “No conflict”).

(2) Mechanism Naming
    • State the mechanism that best explains divergence: temporal / factual-accuracy /
    →  contextual-scope / methodological / linguistic-interpretive.

(3) Conflict Reason (1–2 sentences)
    • Write a short analysis that references doc IDs and clusters explicitly.
      Example: “d1 and d2 report X for the US, while d3 reports Y for Europe; scope differs by
      →  region (contextual-scope).”
```

```
(4) DECISION LADDER (To be used along with CONFLICT TAXONOMY)
    After your reasoning, choose the label using the conflict taxonomy defined above and ONLY IF
    →  NEEDED use this thinking ladder as well:

    • A: If all non-irrelevant docs agree up to minor wording/rounding → choose “No conflict”.
    • B: If explicit scope/time/definition differences explain the different statements WITHOUT
    →  ANY CONTRADICTION → choose “Complementary information”.
    • C: If there are genuinely incompatible claims within the same scope/time and you cannot
    →  clearly say which is correct → choose “Conflicting opinions or research outcomes”.
    • D: If newer docs with explicit timestamps/recency clearly supersede older factual claims →
    →  choose “Conflict due to outdated information”.
    • E (last resort): ONLY if you can clearly show from the snippets that some doc states a
    →  factual claim that is directly refuted by more reliable docs in the set, and this is not
    →  just different scope/opinion → choose “Conflict due to misinformation”.

    Even though we have defined a reasoning ladder here, it is not always necessary that you
    →  should think in this specific order only. This is not an hard-and-fast reasoning ladder,
    →  the most important thing for you is to STRICTLY ADHERE TO THE CONFLICT TAXONOMY in order
    →  to decide the final conflict type label.
    If you are unsure between “misinformation” and another label, you MUST choose that other label.
    →  Also adhere strictly to the CONFLICT TAXONOMY STATED ABOVE.

(5) THEN the Label Line (exactly one line with an EM DASH)
    • <ConflictType> — <concise conflict_reason>
    • The conflict_reason on this line must be consistent with your reasoning above and the
    →  taxonomy.

==============================
ABSTENTION POLICY (STRICT)
==============================
- Abstain ONLY if ALL docs are "irrelevant" OR the set collectively fails to address the query.
- If ANY doc has verdict ∈ {"supports","partially supports"}, DO NOT abstain; produce the best
→  supported answer with conflict-aware framing.

==============================
SOURCE PREFERENCE & CITATION POLICY
==============================
- High-credibility (prefer and cite first): .gov, .edu, WHO/UN/CDC/official orgs, peer-reviewed
→  journals, Britannica, major outlets (Reuters/BBC/AP/NYT/WSJ/Guardian), Mayo Clinic.
- Low-credibility: blogs, unverified forums, marketing pages, social media, miscellaneous sites.
- When multiple sources support the same fact, include high-cred first in prose; then add others
→  by decreasing utility.

==============================
EXPECTED BEHAVIOR RULES (STRICT)
==============================

For a given conflict type, the final answer must adhere to a specific behaviour defined in the
→  rules below:

- "Conflict due to outdated information": Prioritize the most recent and credible information,
→  acknowledging older or superseded claims.
- "Conflicting opinions or research outcomes": Present differing perspectives neutrally, without
→  taking sides; this should be your default for genuine disagreements within the same
→  scope/time.
- "Conflict due to misinformation": Identify and correct false or unreliable claims using verified
→  sources **within the set**, and only when you can prove falsity from the snippets.
- "Complementary information": Combine partial, non-contradictory facts to form a complete,
→  coherent answer.
- "No conflict": Answer directly and confidently using the strongest consistent evidence.

==============================
ANTI-FAILURE GUARDS
==============================
- Exactly one <think>. . . </think>.
- The literal string "<think>" must not appear inside the think block content (no nested tags).
- Enumerate d1. . . dN without gaps, fabrications, or ranges in the array.
- The array must be valid JSON. The rest is plain text.
- ≥80% of final-answer sentences include [dX] (unless abstaining).
```

```
- Use an EM DASH " — " in the label line (not hyphen or en dash).
- Be precise and faithful; no new facts; respect length budgets.

Inputs:

- query:
{query}

- retrieved_docs (ordered d1. . . dN):
{retrieved_docs}

- per_doc_notes (for each doc_id; includes verdict, key_fact, verdict_reason, source_quality):
{per_doc_notes}

Task:
1) Follow the full OUTPUT CONTRACT exactly.
   • <think> block with:
       (A) VALID JSON array for EVERY doc d1. . . dN (order-preserving, one object per doc; if
       →  verdict=="irrelevant" set key_fact="")
       (B) Conflict reasoning FIRST: cluster docs, reference doc IDs, and NAME the mechanism
       →  (temporal / factual-accuracy / contextual-scope / methodological /
       →  linguistic-interpretive)
       (C) ONE label line: "<ConflictType> — <concise conflict_reason>"
       (D) Brief reasoning connecting evidence to the final answer (or abstention)
   • ONE blank line
   • Final answer (or exactly "CANNOT ANSWER, INSUFFICIENT EVIDENCE" if abstaining)
   • Final sentinel line [[END-OF-ANSWER]].

Reminders (DO NOT PRINT):
- Reason FIRST for the conflict: write the analysis in (B), THEN emit the label line in (C); the
→  label must be a direct consequence of the reasoning.
- Conflict taxonomy (strict): No conflict / Complementary information / Conflicting opinions or
→  research outcomes / Conflict due to outdated information / Conflict due to misinformation.
- Use only existing doc_ids in bracketed citations [dX]; no ranges like d1–d5; never cite
→  out-of-bounds [dK].
- Prefer high-credibility sources and order citations high→low in the evidence list.
- If any doc is "supports" or "partially supports", DO NOT abstain.
- Close </think> before the answer; no extra text outside the required format.
```

User Prompt

```
Inputs:

- query:
{query}

- retrieved_docs (ordered d1. . . dN):
{retrieved_docs}

- per_doc_notes (for each doc_id; includes verdict, key_fact, verdict_reason, source_quality):
{per_doc_notes}

Task:
1) Follow the full OUTPUT CONTRACT exactly.
   • <think> block with:
       (A) VALID JSON array for EVERY doc d1. . . dN (order-preserving, one object per doc; if
       →  verdict=="irrelevant" set key_fact="")
       (B) Conflict reasoning FIRST: cluster docs, reference doc IDs, and NAME the mechanism
       →  (temporal / factual-accuracy / contextual-scope / methodological /
       →  linguistic-interpretive)
       (C) ONE label line: "<ConflictType> — <concise conflict_reason>"
       (D) Brief reasoning connecting evidence to the final answer (or abstention)
   • ONE blank line
   • Final answer (or exactly "CANNOT ANSWER, INSUFFICIENT EVIDENCE" if abstaining)
   • Final sentinel line [[END-OF-ANSWER]].

Reminders (DO NOT PRINT):
- Conflict taxonomy (strict): No conflict / Complementary information / Conflicting opinions or
→  research outcomes / Conflict due to outdated information / Conflict due to misinformation.
```

```
- Use only existing doc_ids in bracketed citations [dX]; no ranges like d1–d5; never cite
→  out-of-bounds [dK].
- Prefer high-credibility sources and order citations high→low in the evidence list.
- If any doc is "supports" or "partially supports", DO NOT abstain.
- Close </think> before the answer; no extra text outside the required format.
```

A.5.3 Judge Prompts for Evaluation

Behavior Judge (LLM-as-a-Judge)

You are evaluating whether a model's final answer follows the expected behavior for the given conflict_type.

```
You are evaluating ONLY the *behavior* of a model answer, not its factual correctness.
```

Behavior means:

```
- How the answer handles multiple sources, uncertainty, disagreement, or lack of conflict.
- Whether it summarizes, reconciles, or contrasts viewpoints as appropriate.
- Whether it is direct vs. hedged, neutral vs. biased, etc.
```

Given:

```
- A user query
- A model-generated answer
- A conflict type with an expected behavior rubric
```

Your task:

```
Decide whether the model's answer *follows the expected behavior* for this conflict type.
```

Conflict Type: conflict_type **Expected Behavior (rubric):** BEHAVIOR_RUBRIC **Instructions:**

```
- If the answer clearly follows the expected behavior, set "adherent": true.
- If the answer clearly violates or ignores the expected behavior, set "adherent": false.
- Ignore factual correctness; only judge how the answer behaves relative to the rubric.
- The "rationale" should briefly point to the key aspects of the answer's behavior
  (for example, whether it mentions multiple viewpoints, reconciles partial info,
  prioritizes newer evidence, corrects misinformation, etc.).
```

Return ONLY a JSON object with fields:
"adherent": true or false,
"rationale": "short explanation"

Entailment Judge (Claim–Evidence)

```
You are performing *evidence-based Natural Language Inference (NLI)* for grounded citation checking.
```

Task:

```
Determine the logical relationship between a retrieved document passage (the premise)
and a model-generated claim (the hypothesis), using ONLY the
explicit content of the premise.
```

Definitions:

```
- "entails": The premise clearly supports or confirms the hypothesis. The hypothesis must logically
follow from what the premise states.
- "contradicts": The premise clearly conflicts with or disproves the hypothesis.
- "neutral": The premise does not provide enough information to either support or contradict the
hypothesis. The claim may be plausible, but it is not justified by the premise.
```

Rules:

```
- *Do not add knowledge*, outside interpretation, or world facts.
- *Do not guess* beyond what the premise literally says.
- Focus ONLY on whether the hypothesis is justified by the premise.
Return ONLY a JSON object:
{{
  "relation": "entails" | "contradicts"|"neutral"
}}
```

Single Truth Recall

```
You are checking whether a candidate answer correctly contains a given factual answer.
Consider paraphrases, equivalent wording, and logically equivalent statements as MATCHING.
```
Return ONLY a JSON object with fields:
"adherent": true or false,
"rationale": short string explanation.
The interpretation:

```
- "adherent": true  -> the candidate answer DOES clearly state the gold answer
                       (possibly paraphrased or with additional context).
- "adherent": false -> the candidate answer does NOT contain the gold answer
                       or states something incompatible.
```