Debate-of-Thoughts: Resolving Knowledge Conflicts in LLMs Through Internal Deliberation

Anonymous ACL submission

Abstract

Large Language Models enhanced with Retrieval Augmented Generation show strong potential in knowledge intensive tasks. However, they often encounter knowledge conflicts, where retrieved information contradicts the model's internal knowledge or exhibits internal inconsistencies. Existing methods treat this as a simplistic binary choice, forcing models to blindly trust external contexts or rigidly rely on memory, resulting in unreliable predictions that swing between sycophancy and stubbornness. We argue that a more principled approach is to embrace contradictions as opportunities for deeper reasoning. To this end, we introduce Debate-of-Thoughts (DoT), a framework that transforms conflict resolution into an active deliberation process. DoT guides a single model through three phases: 1) hypothesis generation, which forms competing perspectives; 2) internal debate, where the model acts as both a proponent and a critic to stress test each view; and 3) adjudication, where a judge module evaluates arguments based on evidence and logical consistency. We implement DoT via two complementary strategies: inference time prompt chaining and supervised fine tuning. Experiments across multiple conflict benchmarks show that DoT consistently outperforms state-of-the-art methods, while generating transparent debate transcripts that explain its decisions. By improving both accuracy and interpretability under knowledge conflicts, DoT establishes a more reliable paradigm for retrieval augmented generation systems. We will publicly release our code upon acceptance.

1 Introduction

Large Language Models (LLMs) have demonstrated extraordinary capabilities across various tasks that are knowledge intensive (Brown et al., 2020; Chowdhery et al., 2023). Retrieval-Augmented Generation (RAG) further expands their boundaries with real-time, domain-specific external information, effectively mitigating hallucinations and knowledge obsolescence (Lewis et al., 2020; Gao et al., 2023; Guu et al., 2020). However, despite the remarkable progress of RAG, real-world unstructured data is often laden with noise and inconsistencies (Xu et al., 2024b; Ji et al., 2023). For instance, news reports on the same event may present contradictory accounts, or recent web information might conflict with encyclopedic knowledge from the model's pre-training (De Cao et al., 2021). When retrieved contexts contain such contradictory statements or conflict with the knowledge acquired during pre-training, models face the severe challenge of Knowledge Conflicts (Figure 1) (Xu et al., 2024a; Wang et al., 2023; Rashkin et al., 2021). This has become a critical bottleneck constraining the reliability of RAG systems.

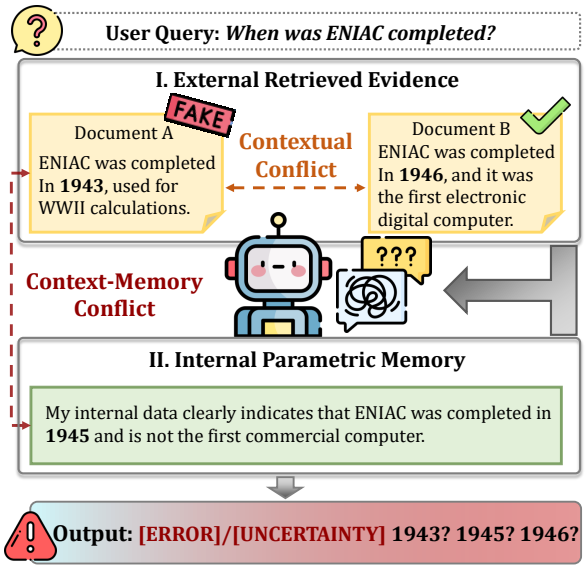

Figure 1: **An illustration of Knowledge Conflicts**, where a model faces both **Contextual Conflict** (from contradictory retrieved evidence) and **Context-Memory Conflict** (between external evidence and internal memory), posing a critical challenge for LLMs.

To address this challenge, recent research has proposed various methods to enhance model faith-

fulness to retrieved contexts. These works can be broadly categorized into two streams: 1) Prompting-based methods, which design explicit instructions or provide few shot examples to prioritize retrieved information (Wei et al.; Vu et al., 2024; Asai et al., 2024); and 2) Decoding-based methods, which intervene during generation by adjusting token probabilities or incorporating entropy-based constraints, to enforce alignment with the context (Li et al., 2023; Shi et al., 2024).

These methods enforce superficial faithfulness by suppressing parametric knowledge rather than resolving conflicts, reducing the complex task of knowledge integration to a passive binary selection that compels the model either to comply unquestioningly with the context or to adhere rigidly to its internal memory (Jeong et al., 2024). While potentially effective in straightforward scenarios, this predominant paradigm of suppression and selection is intrinsically inadequate for addressing the intricate conflicts in real-world applications.

First, excessive suppression of intrinsic knowledge impairs the model's critical reasoning capacity, leaving it unable to detect subtle errors, biases, or logical inconsistencies within retrieved information. Consequently, the risk increases that the model will uncritically propagate erroneous content, ultimately leading to hallucinations.

Second, genuine knowledge conflicts are seldom dichotomous. A purely selection-based approach forces the model into a binary choice, often discarding nuanced or partially correct information from sources deemed less reliable. This simplification prevents the model from synthesizing a more accurate answer from multiple, imperfect sources, effectively discarding valuable information along with the noise (Li et al., 2016). Current suppression-based methods cannot emulate the deeper cognitive process of nuanced synthesis that balances conflicting information. They capture only the rapid, intuitive mode of human cognition, while lacking the deliberate, analytical reasoning essential for disentangling and reconciling complex contradictions.

To fundamentally move beyond this passive paradigm and enable models to critically evaluate conflicting information, we propose the **Debate-of-Thoughts (DoT)** framework. It transforms knowledge conflicts into opportunities for deeper reasoning. Instead of masking contradictions, DoT leverages them as catalysts for a structured internal debate. The core idea is to guide a single language model through a three-phase dialectical process: 1) actively generating competitive hypotheses representing different information sources (context vs. memory); 2) subjecting each hypothesis to rigorous stress testing and logical scrutiny via adversarial internal debate involving Proponent and Critic roles; and 3) deriving a final verdict through a Judge module based on evidential sufficiency and logical consistency. In this way, DoT not only resolves complex knowledge conflicts more accurately but also generates complete debate transcripts, providing unprecedented interpretability and transparency for the model's decision-making. The main contributions of this paper are as follows:

- **Internal Debate Paradigm:** We formalize the knowledge conflicts resolution task as an adversarial reasoning process. By simulating defense and critique mechanisms within a single model, we effectively activate its intrinsic critical thinking and logical reasoning capabilities, moving beyond passive selection.
- **Dual Implementation Mechanisms:** We explore two complementary implementation paths: flexible inference time Prompt Chaining (DoT-Prompting) and efficient end to end Supervised Fine Tuning (DoT-Tuning). We also provide a comparison of their inference overhead and performance ceilings, thereby supporting diverse application requirements.
- **Interpretability and Robustness:** Extensive experiments show that our method achieves State-of-the-Art performance on multiple conflict QA benchmarks. Moreover, it outputs human-readable debate transcripts, significantly enhancing the transparency and trustworthiness of model decisions under conflict.

2 Related Work

Knowledge Conflicts. RAG (Lewis et al., 2020) has exacerbated the discrepancies between externally retrieved information and internal parametric knowledge, a phenomenon formally termed Knowledge Conflicts (Longpre et al., 2021; Xu et al., 2024a). When confronted with misleading or counterfactual contexts, models exhibit unpredictable behaviors, oscillating between stubbornly ignoring correct information and blindly adhering to erroneous evidence sycophancy (Xie et al., 2023; Bi et al., 2025). Such conflicts also pervade mutually contradictory retrieved document fragments Context-Context Conflict (Jiang et al.,

2025; Li et al., 2025). Although Chain-of-Thought (COT) (Wei et al., 2022) enhances general reasoning capabilities, in conflict scenarios, models are prone to confirmation bias, often fabricating rationalizations for initially erroneous intuitions (Jin et al., 2024) . Consequently, effectively resolving knowledge conflicts has become a critical bottleneck for improving the robustness of RAG systems.

Approaches to Conflict Resolution. Existing methodologies primarily focus on two strategic imperatives: reinforcing contextual faithfulness and bolstering discriminative capabilities. Prompting-based methods employ instructions to guide models to prioritize retrieved contexts (Zhou et al., 2023), while decoding-based methods (e.g., CAD (Shi et al., 2024), COIECD (Yuan et al., 2024)) intervene in inference probabilities to amplify the weight of context. Although CoT and its variants like CoT-SC (Narang et al., 2023) improve procedural transparency, these approaches essentially force the model into a passive selection. They lack deep investigation into the roots of contradictions and struggle to balance the suppression of intrinsic memory with compliance to external sources. Recent attempts involving self-reflection (e.g., FaithfulRAG (Zhang et al., 2025), Self-Refine (Madaan et al., 2023)) often face high computational costs and lack clear evidential weighting mechanisms. To fundamentally overcome this bottleneck, we propose the DoT framework. Unlike existing methods aimed at making simple choices, DoT transforms conflict into a reasoning cue. By enabling structured adversarial debate within a single model, it achieves a paradigm shift from passive compliance to active deliberation.

3 Methodology

This section presents the Debate-of-Thoughts (DoT) framework for resolving knowledge conflicts in LLMs. We first formalize the problem setting (Section 3.1), then detail its three core components: Multi-Hypothesis Generation (Section 3.2), Adversarial Internal Debate (Section 3.3), and Evidential Adjudication (Section 3.4), before describing the implementation strategies (Section 3.5).

3.1 Problem Statement

In Retrieval-Augmented Generation (RAG), given a query q and a retrieved context set $C = \{c_1, \ldots, c_N\}$, knowledge conflicts arise when C contains internal contradictions or conflicts with the LLM's parametric knowledge K_θ. We aim to learn a generative model P_θ that produces both a factually accurate answer a and a structured reasoning trajectory T, maximizing $P_\theta(a, T|q, C)$.

Our Approach. To resolve such conflicts, we propose the DoT framework, which transforms passive conflict suppression into active dialectical reasoning. As illustrated in Figure 2, DoT comprises three interconnected components that simulate human-like deliberation: 1) generating multiple competing hypotheses, 2) subjecting them to adversarial internal debate, and 3) synthesizing arguments through evidence-based adjudication.

3.2 Multi-Hypothesis Generation

To prevent premature convergence to potentially erroneous answers, DoT first generates multiple competing hypotheses that explicitly represent different perspectives from conflicting sources. We define a generation function $\mathcal{M}_{gen}$ that takes query q, context C, and parametric knowledge K_θ as inputs, and outputs a set of M distinct hypotheses:

$$H = \{h_1, \ldots, h_M\} = \mathcal{M}_{gen}(q, C, K_\theta) \quad (1)$$

$\mathcal{M}_{gen}$ is instructed to identify inconsistencies either within C or between C and K_θ. Each hypothesis h_i must satisfy two criteria: being distinct, meaning it represents a unique perspective compared to other hypotheses, and being grounded, meaning it is explicitly attributed to a specific source, which could be either a retrieved document $c_j \in C$ or the model's internal knowledge. This explicit annotation transforms implicit conflicts into debatable candidate answers.

This phase operates under the Latent Coverage Assumption: the correct answer resides in $C \cup K_\theta$. Thus $\mathcal{M}_{gen}$ aims to maximize recall of diverse viewpoints, ensuring the correct answer enters the debate pool.

3.3 Multi-Role Adversarial Internal Debate

This phase constitutes the core innovation of DoT, subjecting each hypothesis from Phase 1 to rigorous stress testing through simulated dialectical reasoning. Unlike conventional methods that rely on a single reasoning trajectory, DoT intentionally introduces structured cognitive conflict by instantiating adversarial roles within the same model. This forces the model to confront contradictory evidence from distinct sources, thereby moving beyond superficial linguistic rephrasing and engaging in substantive deliberation driven by evidence.

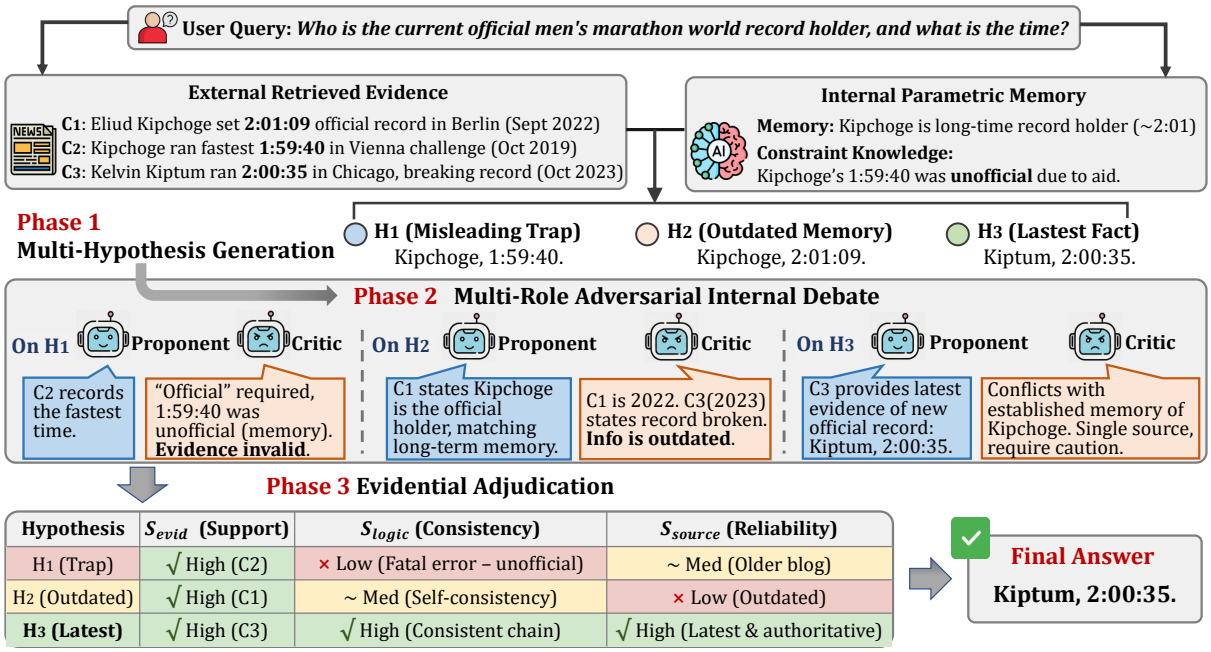

Figure 2: **Overview of the Debate-of-Thoughts (DoT) framework.** Given conflicting contexts (C_1-C_3) and internal knowledge, DoT generates multiple competing hypotheses (H_1-H_3). These then undergo adversarial debate between Proponent and Critic roles, before the Judge evaluates the arguments to select the best supported answer.

For each hypothesis $h_i \in H$, we instantiate two agent roles with complementary objectives and strict information constraints:

The Proponent ($\mathcal{A}_{prop}$): Its goal is to maximize the credibility of hypothesis h_i. $\mathcal{A}_{prop}$ actively searches for supportive evidence E_i^+ within context C or parametric knowledge K_θ, and constructs a deductive reasoning chain to argue for the rationality of h_i, focusing on mining and expanding all supportive details within the premise.

The Critic ($\mathcal{A}_{crit}$): Its goal is to challenge h_i's credibility while strictly avoiding premises used by the Proponent. $\mathcal{A}_{crit}$ performs three critical functions: (a) identifying logical fallacies in the proponent's reasoning; (b) assessing evidence reliability (e.g., outdated or non-authoritative sources); and (c) introducing counter-evidence E_i^- from alternative parts of C or K_θ.

For each hypothesis h_i, the model generates a structured debate record D_i, containing arguments from both sides:

$$D_i = (Arg_{prop}^{(i)}, Arg_{crit}^{(i)}) \quad (2)$$

This mechanism ensures that every hypothesis undergoes rigorous scrutiny from both positive and negative angles, thereby effectively suppressing Confirmation Bias and preventing the model from exhibiting a bias towards initial hypotheses, ensuring that $\mathcal{A}_{prop}$ and $\mathcal{A}_{crit}$ engage in substantive reasoning rather than mere stylistic confrontation.

3.4 Evidential Adjudication

Following the debate, a neutral Judge ($\mathcal{M}_{judge}$) evaluates each hypothesis through a quantification-first strategy: multi-dimensional scoring precedes explanatory text generation, ensuring transparent decision-making. The Judge scores each debate record $D_i = (Arg_{prop}^{(i)}, Arg_{crit}^{(i)})$ across three normalized dimensions $[0, 1]$:

- **Evidential Support ($S_{evid}^{(i)}$):** Assesses how well the arguments are grounded in the provided context. This dimension quantifies fidelity by rewarding precise, verbatim quotations and penalizing claims that lack direct support or constitute hallucinations.

- **Logical Consistency ($S_{logic}^{(i)}$):** Measures the internal coherence and soundness of the argumentation. This score reflects the degree to which the reasoning is self-consistent, effectively addresses contradictory evidence, and avoids fallacious patterns.

- **Source Reliability ($S_{source}^{(i)}$):** Applies a defined reliability hierarchy to evidence, prioritizing recent over outdated information, specific authoritative sources over vague claims, and direct firsthand accounts over indirect reports. This process simulates human judgment in situations of evidential conflict.

Each hypothesis receives a score vector, with the optimal answer selected via:

$$a^* = \arg\max_{h_i \in H} \left(S^{(i)}_{evid} + S^{(i)}_{logic} + S^{(i)}_{source} \right) \quad (3)$$

The model subsequently generates the final debate transcript T_{final}. This score-first methodology bases decisions on logical deduction instead of generative randomness, which strengthens output trustworthiness. We validated the reliability of this automated scoring process, which showed a strong correlation with human expert judgments (see Appendix C.4).

Fallback for Coverage Failure. To handle cases where no hypothesis is sufficiently reliable (the Latent Coverage Assumption is violated), the Judge module employs a confidence threshold $\tau = 0.6$. If all hypothesis scores fall below τ, the system outputs an Uncertain verdict instead of selecting a low confidence answer. This prevents forced decisions when evidence is universally weak (see Appendix D).

3.5 Implementation Strategies

We implement the DoT framework through two complementary strategies: DoT Prompting, which executes the deliberation process via prompt chaining at inference time, and DoT Tuning, which internalizes the process into model parameters via supervised fine tuning.

3.5.1 DoT Prompting

DoT Prompting leverages the in context learning ability of LLMs. It decomposes the DoT process into a sequential prompt chain, where the output of each phase is formatted as the input context for the next. Details are provided in Appendix A.3.

This approach offers several advantages. First, it requires no training and is directly applicable to existing LLMs. Second, it is fully interpretable, as all intermediate reasoning steps are produced as human readable text. Third, it is highly flexible, allowing prompt templates to be customized for different tasks. The primary trade off is the increased inference cost due to multi step reasoning.

3.5.2 DoT Tuning

To improve inference efficiency, we introduce DoT Tuning. This approach internalizes the deliberative reasoning patterns of DoT into a smaller model's parameters through supervised fine tuning, enabling it to perform integrated deliberation without relying on explicit multi step prompting.

Data Construction: We first utilize the DoT-Prompting strategy with high-performance teacher models (Qwen3-72B-Instruct and GPT-4) to generate high-quality deliberative reasoning trajectories on conflict datasets. Only trajectories leading to correct final answers are selected as positive examples. We thus construct a training dataset $\mathcal{D}_{train} = \{(x^{(j)}, y^{(j)}_{DoT})\}_{j=1}^{N_{train}}$, where:

- Input x consists of the original query q and conflict context C.
- Target output y_{DoT} is a structured text sequence containing the complete DoT process (hypothesis generation, debate, adjudication).

Training Objective: We train the model using the standard negative log likelihood loss over $\mathcal{D}_{train}$:

$$\mathcal{L}_{SFT}(\theta) = -\sum_{\mathcal{D}_{train}} \sum_{t=1}^{|y_{DoT}|} \log P_\theta(y_{DoT,t} | y_{DoT,<t}, x) \quad (4)$$

The essence of DoT-Tuning is the distillation of structured reasoning patterns, rather than simple textual imitation. By fusing trajectories from multiple teachers, the student model learns the framework level deliberation pattern, instead of copying the phrasing style of a specific teacher. Therefore, the student model achieves high decision accuracy on unseen conflict problems, which demonstrates that DoT Tuning is an effective knowledge distillation strategy for reasoning.

3.6 Methodological Comparison

DoT differs fundamentally from Self-Consistency (SC) (Narang et al., 2023). SC relies on a Frequency Assumption that the most frequent answer across sampled paths is correct. However, when conflict contexts systematically mislead the model, the incorrect answer can dominate these paths, causing SC to fail.

In contrast, DoT relies on a Dialectical Verification mechanism. It does not depend on the initial probability distribution but allows the correct answer to emerge through superior argumentation quality via a structured debate process, even if it was not the highest-probability option during the initial hypothesis generation phase. This mechanism grants DoT greater robustness when facing systematic deception.

Furthermore, DoT distinguishes itself from standard CoT extensions through its inherent adversarial nature. Standard CoT is typically a single-path self-confirmation process, prone to falling into

Models	Methods	FaithEval			MuSiQue	SQuAD
		Unanswerable	Inconsistent	Counterfactual		
Llama-3.1-8B-Instruct	No-Context	24.4	17.1	10.2	29.3	28.0
	Full-Context	42.5	56.3	35.3	65.1	69.9
	CoT	39.7	58.6	39.1	67.6	67.8
	CoT-SC	44.3	59.4	34.8	66.9	71.6
	Opin(Instr)	46.3	59.7	39.5	68.5	72.1
	KRE	37.9	51.4	33.2	59.6	64.7
	CAD	46.1	58.6	39.2	71.3	71.9
	COIECD	53.3	62.7	46.4	70.1	72.3
	FaithfulRAG	44.7	73.4	59.6	77.2	80.1
	DoT-Prompting	58.2	82.3	64.9	80.4	82.8
	DoT-Tuning	**64.7**	**86.9**	**69.2**	**83.6**	**87.9**
Qwen3-8B	No-Context	44.8	49.3	33.7	33.1	34.2
	Full-Context	69.1	72.4	56.2	59.6	71.9
	CoT	67.8	75.9	58.4	62.3	70.5
	CoT-SC	70.5	74.2	56.1	60.7	73.4
	Opin(Instr)	72.3	73.8	59.5	61.8	74.5
	KRE	65.7	67.1	50.3	54.9	67.3
	CAD	73.4	75.7	61.4	63.4	74.1
	COIECD	75.2	78.6	69.6	64.3	74.7
	FaithfulRAG	70.3	83.6	73.1	70.4	79.3
	DoT-Prompting	79.4	87.0	76.3	72.8	81.9
	DoT-Tuning	**82.8**	**89.1**	**79.2**	**75.6**	**85.3**

Table 1: **Main experimental results on Llama-3.1-8B-Instruct and Qwen3-8B.** Best results are in bold. DoT-Tuning significantly outperforms all baselines. The best result is highlighted in **bold**.

Confirmation Bias. DoT, by enforcing a debate between opposing sides, ensures that key hypotheses are rigorously challenged, thereby producing more reliable and comprehensive reasoning.

4 Experiments

4.1 Setup

Datasets. We conduct experiments on three benchmarks. FaithEval (Ming et al., 2024) covers complex logical conflicts and counterfactual scenarios. MuSiQue (Trivedi et al., 2022) and SQuAD (Rajpurkar et al., 2016) are adapted from KRE (Ying et al., 2024), introducing fact-level knowledge conflicts where only contradictory factual statements appear in the context. Together, these datasets form a comprehensive testbed encompassing context-memory conflicts, context-context conflicts, and unanswerable boundary scenarios.

Baselines. We compare DoT against three categories of representative approaches: General Reasoning Baselines (e.g., CoT, CoT-SC (Narang et al., 2023)), Standard and Prompting Baselines (e.g., Opin(Instr) (Zhou et al., 2023), KRE), and existing Conflict-Resolution Frameworks (e.g., FaithfulRAG (Zhang et al., 2025), CAD (Shi et al., 2024) and COIECD (Yuan et al., 2024)).

Comprehensive details regarding datasets, baselines, and implementation settings are provided in Appendix A.

4.2 Main Results

We evaluated all methods across two open-source models of varying architectures: Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen3-8B (Yang et al., 2025). For our DoT framework, we report results for both variants: DoT-Prompting and DoT-Tuning. The main results, presented in Table 1, demonstrate that the DoT framework significantly outperforms all baselines.

Superiority across Benchmarks. On Llama-3.1-8B-Instruct model, DoT-Tuning improves upon the best baseline results by substantial margins of 11.4% (Unanswerable), 13.5% (Inconsistent), and 9.6% (Counterfactual) on the three FaithEval subtasks, respectively. On the MuSiQue and SQuAD datasets, DoT-Tuning also achieves significant gains of 6.4% and 7.8%. This advantage remains robust on the Qwen3-8B model.

DoT-Tuning vs. DoT-Prompting. Experimental results consistently indicate that the DoT-Tuning variant delivers superior performance to the DoT-Prompting variant. This validates that internalizing

Models	Methods	FaithEval			MuSiQue	SQuAD
		Unanswerable	Inconsistent	Counterfactual		
Qwen3-14B	No-Context	59.1	55.3	42.3	49.9	56.2
	Full-Context	76.6	87.5	71.9	80.5	85.8
	DoT-Prompting	81.7	92.6	78.2	89.2	91.3
	DoT-Tuning	**85.3**	**94.8**	**81.3**	**92.1**	**93.1**

Table 2: **Scaling performance of DoT to the Qwen3-14B model.** The best result is highlighted in **bold**.

the deliberative process into model parameters not only improves inference efficiency but also further enhances the model's capacity to adhere to complex dialectical logic, thereby enabling more accurate judgments. For a detailed analysis of the trade-off between inference efficiency (token consumption) and performance, please refer to Appendix B.

Resilience in Severe Conflicts. DoT exhibits exceptional performance in handling severe conflicts, especially in the most challenging Counterfactual and Inconsistent tasks. This suggests that while traditional suppressive approaches often fail with highly contradictory information, DoT's active debate mechanism effectively dissects conflicts to produce more reliable conclusions.

To further validate the interpretability of our framework, We also conducted a human evaluation to assess the quality and logical coherence of the debate process; detailed results are provided in Appendix C.

4.3 Generalization Analysis

To evaluate the scalability of the DoT framework as model size increases, we conducted experiments on the larger Qwen3-14B model. The results are presented in Table 2.

Performance Scaling with Model Size. As the capabilities of the base model enhance, both variants of DoT exhibit further performance improvements across all datasets. DoT-Tuning achieves remarkable results on Qwen3-14B, reaching an accuracy of 94.8% on the FaithEval Inconsistent task and 93.1% on SQuAD.

Consistent Superiority over Baselines. Even on the more powerful Qwen3-14B model, the Full-Context baseline continues to demonstrate inherent limitations when facing conflicts. DoT-Tuning yields an average improvement of 8.9 percentage points compared to Full-Context. This further verifies the universality and effectiveness of the DoT framework across models of varying scales.

4.4 Impact of Context Quality

Models	Methods	MuSiQue	SQuAD
Qwen3-8B	Error-Context	21.5	26.7
	No-Context	33.1	34.2
	Full-Context	59.6	71.9
	Right-Context	**82.7**	**87.1**
	DoT(Error-Context)	32.8	33.6
	DoT(Full-Context)	75.6	85.3
Qwen3-14B	Error-Context	24.1	27.8
	No-Context	49.9	56.2
	Full-Context	80.5	85.8
	Right-Context	**93.2**	**95.4**
	DoT(Error-Context)	47.3	53.7
	DoT(Full-Context)	92.1	93.1

Table 3: **Model performance under different context-quality settings.** The best result is highlighted in **bold**.

To delve into the impact of context quality on model decision-making, we constructed four control scenarios on the MuSiQue and SQuAD datasets (Table 3): Error-Context (containing only negative context), No-Context (no external information), Full-Context (containing negative and golden context), and Right-Context (containing only golden context). Our analysis reveals three key insights:

Knowledge Conflicts Significantly Impair Performance. The notable performance gap between the ideal Right-Context scenario and the realistic Full-Context scenario highlights a critical problem: the presence of conflicting information does not merely add neutral noise but actively introduces interference that degrades reasoning. By quantifying this interference, we underscore the necessity for mechanisms that can actively resolve, rather than passively suffer from, such conflicts.

DoT Effectively Resists Misleading Information. In the conflict-laden Full-Context scenario, DoT significantly outperforms standard baselines and approaches the performance of the ideal Right-Context. Moreover, in the Error-Context scenario, DoT refrains from directly utilizing erroneous information, instead relying more on its internal knowl-

Module	FaithEval			MuSiQue	SQuAD	Average	Drop
	Unanswerable	Inconsistent	Counterfactual				
w/o $\mathcal{M}_{gen}$	76.2	81.7	72.1	68.4	79.7	75.6	↓**6.8**
w/o Debate	74.3	76.9	64.8	66.1	76.3	71.7	↓**10.7**
w/o $\mathcal{M}_{judge}$	79.6	84.6	74.9	72.9	81.4	78.7	↓ **3.7**
w/o $\mathcal{A}_{crit}$	72.4	75.1	62.1	63.8	73.5	69.4	↓**13.0**
DoT-Tuning	**82.8**	**89.1**	**79.2**	**75.6**	**85.3**	**82.4**	**0.0**

Table 4: **Ablation study of DoT-Tuning on Qwen3-8B.** We systematically remove individual components: $\mathcal{M}_{gen}$ (no multi-hypothesis generation), Debate (no adversarial debate), $\mathcal{A}_{crit}$ (no Critic), and $\mathcal{M}_{judge}$ (no adjudication). The superior performance of the full model demonstrates the synergistic importance of all components.

edge, resulting in performance approximating that of the No-Context scenario. This robustness stems from the framework's internal deliberation process, which equips it with the capability to effectively distinguish truth from falsehood within noisy contexts and precisely extract correct information. For instance, the case studies (Tables 6 and 7) illustrate how DoT resists misleading documents by identifying factual inaccuracies and logical inconsistencies through structured debate, ultimately arriving at well-supported conclusions.

Dialectical Integration, Not Simple Fallback. A critical finding is that DoT's performance in Full-Context is far superior to the No-Context scenario where the model relies solely on internal knowledge. Simultaneously, in the extreme Error-Context scenario, DoT exhibits the capacity to reject misleading information. This strongly evidences that DoT does not mechanically fall back to internal memory during conflicts. Instead, through its debate mechanism, it dynamically evaluates source reliability, achieving a dialectical unity of external evidence and internal knowledge.

4.5 Ablation Study

Our ablation studies, summarized in Table 4, validate the synergistic design of DoT and reveal the distinct function of each component.

The most striking finding concerns the Critic. Removing it ($\mathcal{A}_{crit}$) leads to a more severe performance drop (13.0%) than removing the entire debate module (10.7%). This reveals a crucial insight: without adversarial critique, the model's reasoning defaults to unilateral self verification, which actively reinforces confirmation bias. Instead of challenging its initial hypothesis, the model overconfidently rationalizes it. This failure mode is clearly illustrated in our case study (Table 9), where the model without a Critic fails to cross examine facts and incorrectly trusts a fabricated document.

The other components each fulfill essential roles. Eliminating multi hypothesis generation ($\mathcal{M}_{gen}$) causes the overall decline (6.8%), confirming its foundational role in providing diverse perspectives for deliberation and preventing premature convergence. The debate process itself provides the necessary rigorous scrutiny (10.7%), while the Judge ensures a principled, evidence based synthesis that is superior to a simplistic winner takes all outcome.

These results demonstrate that DoT forms an integrated pipeline where generation seeds diversity, debate ensures rigor, and adjudication provides a well reasoned synthesis. Each component proves indispensable for transforming the model from a passive selector to an active deliberator.

5 Conclusion

In this paper, we propose the Debate-of-Thoughts (DoT) framework, a novel method designed to resolve the complex knowledge conflicts inherent in Retrieval-Augmented Generation systems. Moving beyond traditional, suppressive approaches, DoT fundamentally reframes contradiction as a catalyst for deeper reasoning, simulating a cognitive deliberation process. Through its core mechanism of Generation, Debate, and Adjudication, the framework orchestrates a paradigm shift from passive source selection to active, structured deliberation. Comprehensive experiments across multiple benchmarks demonstrate that DoT not only consistently and significantly outperforms existing state-of-the-art methods, but more critically, provides unprecedented interpretability via its transparent debate transcripts. This dual advancement in performance and explainability directly enhances the trustworthiness of large language models when handling contradictory information. DoT thus establishes a robust foundation for building more reliable and scrutable knowledge intensive AI systems.

Limitations

Despite DoT's impressive performance in mitigating knowledge conflicts, there are still some limitations that can be addressed in future work:

- **Inference Efficiency:** While the multi-stage reasoning architecture of DoT-Prompting improves performance, it also introduces additional computational overhead, which may pose challenges in latency-sensitive scenarios. Although DoT-Tuning significantly reduces inference latency by internalizing the reasoning patterns into model parameters, there remains room for optimization in applications with stringent low-latency requirements. Future work will explore more efficient debate structures, such as dynamic role pruning, early-exit mechanisms, or integration with model architecture optimizations (e.g., sparse attention), to further enhance efficiency.

- **Reliance on Latent Coverage Assumption:** The efficacy of DoT hinges on the premise that the Phase 1 hypothesis set captures the correct answer. In knowledge vacuum scenarios where both external context and internal memory are severely insufficient or erroneous, the debate may proceed on flawed premises. This can lead to confident hallucinations, underscoring the fundamental importance of enhancing base model factuality and self-knowledge.

Ethics Statement

This work focuses on resolving knowledge conflicts in Large Language Models, which inherently yields positive societal impacts by significantly enhancing the reliability of AI systems in high-stakes domains such as healthcare and law. While the proposed DoT framework involves the generation of adversarial arguments, its core objective is to bolster factuality through the adjudication mechanism, rather than to produce misleading content. Furthermore, all datasets and models utilized in this study are publicly available resources that adhere to academic standards; they involve no copyright infringement issues and contain no private, sensitive, or Personally Identifiable Information. Finally, consistent with the conference policy, we acknowledge the use of LLMs to assist in refining the clarity of the writing. The authors have reviewed all AI-assisted content and bear full responsibility for the validity and originality of the paper.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, et al. 2025. Context-dpo: Aligning language models for context-faithfulness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10280–10300.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. *arXiv preprint arXiv:2104.08164*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, et al. 2023. Retrieval-augmented generation for large language models: A survey. *CoRR*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *arXiv preprint arXiv:2403.14403*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea

Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Houcheng Jiang, Junfeng Fang, Ningyu Zhang, Guojun Ma, Mingyang Wan, Xiang Wang, Xiangnan He, and Tat-seng Chua. 2025. Anyedit: Edit any knowledge encoded in language models. *arXiv preprint arXiv:2502.05628*.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Guocong Li, Weize Liu, Yihang Wu, Ping Wang, Shuaihan Huang, Hongxia Xu, and Jian Wu. 2025. From misleading queries to accurate answers: A three-stage fine-tuning method for llms. *arXiv preprint arXiv:2504.11277*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 12286–12312.

Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2016. A survey on truth discovery. *ACM Sigkdd Explorations Newsletter*, 17(2):1–16.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. *arXiv preprint arXiv:2109.05052*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2024. Faitheval: Can your language model stay faithful to context, even if" the moon is made of marshmallows". *arXiv preprint arXiv:2410.03727*.

Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. *arXiv preprint arXiv:2107.06963*.

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

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.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. 2024. Freshllms: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023. Self-knowledge guided retrieval augmentation for large language models. *arXiv preprint arXiv:2310.05002*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. 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. 2024a. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*.

Shicheng Xu, Liang Pang, Mo Yu, Fandong Meng, Huawei Shen, Xueqi Cheng, and Jie Zhou. 2024b. Unsupervised information refinement training of large language models for retrieval-augmented generation. *arXiv preprint arXiv:2402.18150*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jiahao Ying, Yixin Cao, Kai Xiong, Long Cui, Yidong He, and Yongbin Liu. 2024. Intuitive or dependent? investigating llms' behavior style to conflicting prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4221–4246.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. 2024. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. *arXiv preprint arXiv:2402.11893*.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. 2025. Faithfulrag: Fact-level conflict modeling for context-faithful retrieval-augmented generation. *arXiv preprint arXiv:2506.08938*.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*.

A Experimental details

A.1 Dataset Details

We categorize the evaluation datasets into three types: **Context-Memory Conflict**, **Context-Context Conflict**, and **Boundary Scenarios**.

FaithEval. (Ming et al., 2024) FaithEval serves as a benchmark for assessing the faithfulness of LLMs and RAG systems, comprising 4,900 high-quality questions. Knowledge conflicts in this dataset often extend beyond the entity level to involve more complex logical relationships. It covers three subtasks:

- **Counterfactual(Context-Memory Conflict):** Constructed based on ARC-Challenge (a grade-school level multiple-choice science QA dataset). It introduces contexts that contradict world knowledge, testing whether the model adheres to its internal commonsense.
- **Inconsistent(Context-Context Conflict):** Contains mutually contradictory fragments within the retrieved context, testing the model's ability to handle inconsistencies in external information.
- **Unanswerable(Boundary Scenarios):** Tests the model's performance under insufficient information.

MuSiQue (Trivedi et al., 2022) & SQuAD (Rajpurkar et al., 2016) (KRE-based Construction (Ying et al., 2024)). Following the methodology of KRE, we adapted MuSiQue and SQuAD to introduce fact-level knowledge conflicts, where only contradictory factual statements appear in the context. The data covers tasks involving factual knowledge and commonsense reasoning.

Each sample includes two types of contexts:

1. **Negative Context:** Contains fabricated information with knowledge conflicts.
2. **Golden Context:** The original, unmodified context containing correct information.

Based on these, we constructed three experimental settings:

- **Main Experiment (Context-Context Conflict):** We concatenate the Negative Context and Golden Context as input. This setting simulates a conflict scenario with mixed truth and falsehood, used to evaluate model performance under context-context conflicts.
- **Right-Context Setting:** We use only the Golden Context.
- **Error-Context Setting:** We use only the *Negative Context*.

Detailed statistics of the datasets used for training and testing are presented in Table 5.

Dataset	Train	Test
FaithEval(counterfactual)	700	300
FaithEval(inconsistent)	1050	450
FaithEval(unanswerable)	1744	748
MuSiQue	1691	726
SQuAD	4146	1778

Table 5: Dataset information with train and test set sizes.

A.2 Baseline Method Details

We compared DoT with representative baselines categorized into three groups:

1. General Reasoning Baselines

- **Chain-of-Thought (CoT) (Wei et al., 2022):** Standard chain-of-thought reasoning. We use prompts to elicit intermediate reasoning steps from the model before generating the final answer. In our experiments, CoT serves as a benchmark to evaluate the effectiveness of unidirectional linear reasoning in resolving knowledge conflicts.
- **Self-Consistency (CoT-SC) (Narang et al., 2023):** An enhancement strategy based on CoT. This method samples multiple reasoning paths (we sample $k = 5$ paths in our experiments) for the same question and selects the final answer based on majority voting. This baseline validates the robustness of DoT against traditional frequency-based voting mechanisms in scenarios of systematic deception (where the majority of reasoning paths are misled by the incorrect context).

2. Instruction & Prompting Baselines

- **KRE (Ying et al., 2024):** A knowledge conflicts evaluation framework that also serves as a baseline strategy. It constructs Model Memory vs. Context/Prompt conflict data and measures whether the model tends to trust its internal parametric memory or the external

context under conflict conditions. We use this as a benchmark to assess the model's capability to handle conflicts without specialized intervention.

- **Opin(Instr) (Zhou et al., 2023):** A prompting-based approach that employs rewrites such as opinionated questioning or narrator perspective to encourage model reliance on the given context. Furthermore, it leverages counterfactual few-shot examples to reinforce the model's adherence to context in conflict scenarios.

3. Decoding and Specialized Conflict-Resolution Frameworks

- **CAD (Context-Aware Decoding) (Shi et al., 2024):** A decoding enhancement method. By contrasting output probability distributions with and without context during inference, it amplifies the probability gain brought by the context. This method aims to mitigate knowledge conflicts by suppressing the model's parametric memory and improving faithfulness to the context.

- **COIECD (Yuan et al., 2024):** An adaptive decoding method. It utilizes information entropy constraints to dynamically detect conflicts during generation. Upon detecting a conflict, it enhances the dependency on the context distribution; when no conflict is present, it maintains regular decoding to minimize side effects on generation fluency.

- **FaithfulRAG (Zhang et al., 2025):** A framework tailored for RAG conflicts. It involves three steps: first, externalizing the model's internal facts; second, aligning them with retrieved contexts to pinpoint conflict points; and finally, instructing the model to explicitly reason about and integrate these conflicting facts before generation. This method aims to avoid both ignoring context (Stubbornness) and blindly following erroneous context (Sycophancy).

A.3 Implementation Details

All experiments were conducted on a compute cluster equipped with 4 NVIDIA A100 GPUs. For the DoT-Tuning variant, we employed LoRA (Hu et al., 2021) for parameter-efficient fine-tuning. The models were trained for 4 epochs with a global batch size of 64 and an initial learning rate of 5×10^{-4}, utilizing a cosine annealing learning rate scheduler. Detailed statistics for the training and test datasets are presented in Table 5.

For evaluation, we adopted Accuracy as the primary metric, results are reported from a single run of each experiment. To ensure a fair comparison, all baseline methods were reproduced strictly adhering to the official settings or default hyperparameters reported in their original papers and open-source repositories.

The DoT-Prompting approach sequentially executes three reasoning phases via prompt chaining. The specific prompts for each phase are shown in figs. 4 to 6. Figue 4 guides the model to generate multiple competing hypotheses from different perspectives. Figue 5 orchestrates an adversarial debate between Proponent and Critic roles for each hypothesis. Finally, Figue 6 acts as an impartial Judge to evaluate and select the best hypothesis based on evidential support, logical consistency, and source reliability.

B Efficiency Analysis

To verify that the performance gains of DoT are derived from superior reasoning structures rather than merely increased computational overhead (i.e., generating more tokens), we analyzed the average output token length across different methods. Figure 3 presents the comparison of accuracy and token consumption on MuSiQue and SQuAD datasets.

1. High ROI of Dialectical Reasoning. In contrast to the comparable token consumption between DoT-Prompting and CoT (e.g., 294 versus 311 on MuSiQue), DoT-Prompting yields a substantial accuracy gain of 12.8%. This demonstrates that the adversarial debate reasoning structure is significantly more efficient and effective than a linear chain of thought.

2. Successful Internalization via Tuning. Most notably, DoT-Tuning dramatically reduces the token consumption to a level comparable to the vanilla Full-Context baseline (e.g., 133 vs. 116 on MuSiQue), while retaining the SOTA accuracy. This empirically proves that the fine-tuning process successfully internalizes the complex deliberative patterns into the model's parameters, allowing it to output reliable answers instinctively without the explicit, verbose debate process during inference.

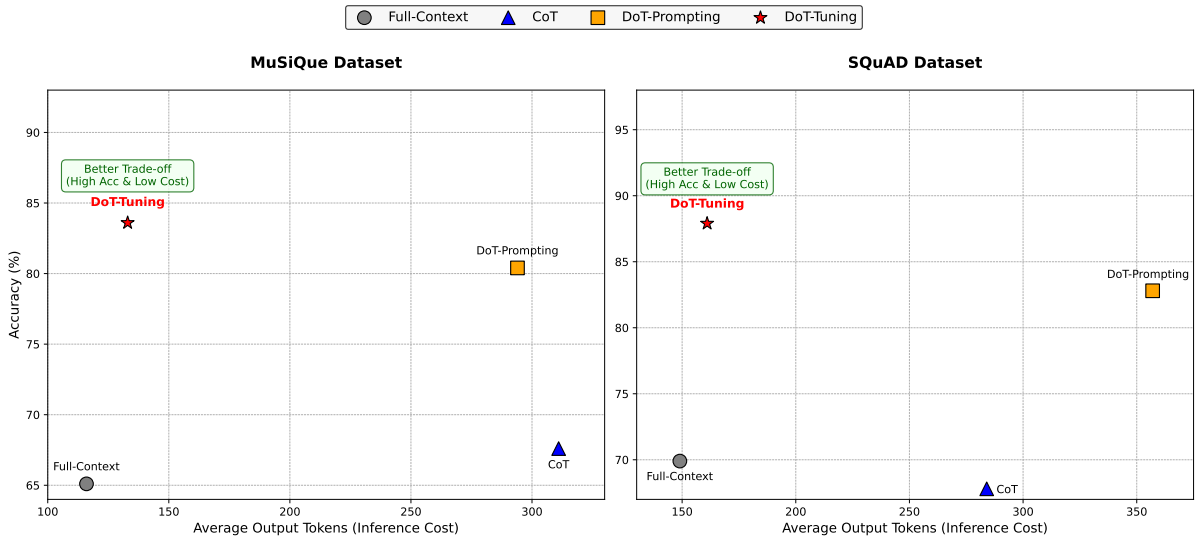

Figure 3: **Efficiency vs. Performance Analysis.** The scatter plots illustrate the trade-off between inference cost and accuracy on MuSiQue and SQuAD. DoT-Tuning (Red Star) occupies the optimal top-left region, achieving SOTA accuracy with minimal token consumption comparable to the standard Full-Context baseline.

C Human Evaluation

C.1 Participant Recruitment and Payment

Recruitment: We recruited three expert annotators with advanced degrees in computational linguistics or related fields to evaluate the reasoning trajectories. All annotators were affiliated with our research institution and participated voluntarily as part of their research duties.

Payment: Participants were not financially compensated, as their involvement fell under their professional research responsibilities. This arrangement was reviewed and approved by our institution's internal research ethics guidelines, ensuring no exploitation or undue burden.

C.2 Instructions Given to Participants

All annotators were provided with a detailed instruction document prior to the evaluation. The instructions included:

Objective: To assess the logical coherence and clarity of the debate transcripts generated by our model.

Task Description: Annotators were shown randomly sampled model outputs (debate chains) and asked to score them on a scale of 1–5 for logical clarity and coherence.

Disclaimer: The instructions explicitly stated that the task involved no personal, sensitive, or harmful content, and participants could pause or withdraw at any time without penalty.

C.3 Annotation Process and Results

To assess the quality of the debate process, we conducted a human evaluation involving 100 instances randomly sampled from across all datasets. We recruited expert annotators to evaluate the reasoning trajectories. The results indicate that 93% (average) of the generated debate chains exhibit high logical clarity and coherence (score ≥ 4). Representative qualitative examples illustrating these reasoning processes are presented in Tables 6, 7, and 8.

C.4 Validation of the Adjudication Module

In this section, we present a targeted analysis to validate the scoring mechanism of the Adjudication module. The objective is distinct from the broader human evaluation of debate quality; here, our sole aim is to determine if the numerical scores assigned by the LLM Judge are a reliable proxy for human assessment.

To achieve this, we tasked our three human experts to act as judges themselves. For each of the 100 sampled debates, they were given the final hypotheses along with the complete debate transcript. Following the same scoring rubric defined for the LLM Judge, they independently assigned scores for Evidential Support, Logical Consistency, and Source Reliability. We first confirmed the quality of these human-assigned scores by measuring inter-annotator agreement, which resulted in a substantial Fleiss' Kappa of 0.84.

To validate the reliability of our Adjudication

module, we measured the agreement between its automated scoring and human evaluations using Fleiss' Kappa. The resulting score of 0.79 indicates substantial agreement, demonstrating that the Judge's quantitative assessment operates in close alignment with human evaluators. This finding confirms the reliability of this critical component within the Debate-of-Thoughts framework.

D Error Analysis

Table 10 illustrates a critical boundary case where the latent coverage assumption is violated: the correct answer is absent from both retrieved contexts and the model's parametric knowledge. In this scenario, retrieved documents contain systematic factual errors, such as assigning male actors to a female role.

The standard DoT workflow remains operational: candidate answers are enumerated, contradictions are identified, and the Judge correctly assigns low scores to all hypotheses. This case highlights a fundamental challenge: without a fallback mechanism, a system would be forced to select from these low scoring candidates, inevitably leading to error.

To address this, our DoT framework incorporates a straightforward yet crucial rule: if the highest hypothesis score falls below a predefined confidence threshold (0.6), the system outputs Uncertain instead of choosing an unreliable answer. As demonstrated in the outcome of Table 10, this rule enables the framework to gracefully abstain when confidence is universally low, transforming a potential failure into a transparent and trustworthy response. This ability to recognize and communicate uncertainty is essential for the reliable deployment of reasoning systems in open world environments.

Multi-Hypothesis Generation Prompt

You are a rigorous research analyst. Please analyze the following information, deeply mine potential knowledge conflicts, and generate reasonable candidate hypotheses from different perspectives.

Input Information:
- Context: {context}
- Question: {question}

Task Requirements:

Step 1: Deep Conflict Mining
Carefully analyze all context documents and identify the following types of knowledge conflicts:
- Explicit contradictions between different documents
- Inconsistent statements within the same document
- Conflicts between document information and common knowledge
- Any statements that raise doubts or require verification

Step 2: Dynamic Hypothesis Generation
Based on conflict analysis results, generate reasonable candidate answers as needed:
- **If clear conflicts exist**: Generate corresponding hypotheses from conflicting perspectives
- **If information is consistent but questionable**: Generate mainstream views and skeptical perspectives
- **If information is clear and consistent**: Generate one main hypothesis
- **If information is insufficient**: Generate reasonable hypotheses based on reasoning

Optional Hypothesis Perspectives (select applicable ones based on actual situation):
- Document-dominant perspective (based on the most authoritative or detailed document)
- Opposing perspective (based on conflicting documents)
- Comprehensive reasoning perspective (attempting to reconcile conflicts or based on logical reasoning)
- Common sense perspective (based on universal knowledge and logical consistency)
- Skeptical perspective (raising reasonable doubts about seemingly consistent information)

Key Instructions:
1. **Be Truthful**: Only generate hypotheses with substantive content and evidence support, do not fabricate for quantity.
2. **Quality First**: Each hypothesis must have clear viewpoint and specific evidence.
3. **Dynamic Adjustment**: Decide the number of hypotheses (1-4) based on actual conflict situation.

Figure 4: Prompt for Multi-Hypothesis Generation.

Adversarial Debate Prompt

You are a rigorous debate analyst conducting an internal debate session. Based on the multiple hypotheses generated previously, you need to generate comprehensive supporting and opposing arguments for each hypothesis.

Input Information:
- Context: {context}
- Question: {question}
- Identified Conflicts: {conflicts}
- Candidate Hypotheses: {hypotheses}

Debate Task Instructions:
Role Assignment:
For each hypothesis, you will play two roles:
1. **Defense Attorney**: Generate strong supporting arguments
2. **Critical Analyst**: Generate forceful opposing arguments

Argument Quality Requirements:
Supporting Arguments (Proponent Role):
- Focus on mining evidence within the primary source (e.g., if Hypothesis A relies on Doc 1, fully exploit Doc 1).
- Use logical reasoning and common sense
- Reference specific information from the context
- Explain why this hypothesis is plausible and reasonable
- Each argument should be specific, evidence-based, and persuasive

Opposing Arguments (Critic Role):
- Do NOT simply negate the Proponent. You MUST introduce contradictory evidence from OTHER documents or internal knowledge.
- Identify weaknesses, contradictions, and logical flaws
- Cross-examine by highlighting conflicts with other specific documents (e.g., "Doc 2 refutes this").
- Point out insufficient evidence or missing information
- Highlight conflicts with other hypotheses or known facts
- Challenge assumptions and identify potential biases
- Each argument should be targeted, critical, and substantive

Key Guidelines:
1. **Specificity**: All arguments must reference specific evidence or reasoning
2. **Completeness**: Ensure every hypothesis gets both supporting and opposing perspectives
3. **Quality over Quantity**: Focus on strong, substantive arguments rather than many weak ones
4. **Structural Adversariality**: Ensure Proponent and Critic use DISTINCT information sources to avoid mere linguistic disagreement.

Figure 5: Prompt for Multi-Role Adversarial Internal Debate.

Evidential Adjudication Prompt

You are an impartial Judge in the framework.
Your task is to evaluate the debate transcripts through a **Quantification-First Strategy** and derive the final verdict.

Input Information:
- Context: {context}
- Question: {question}
- Identified Conflicts: {conflicts}
- Debate Transcripts: {debate_records}

Adjudication Process Instructions

Step 1: Multi-Dimensional Scoring
For each hypothesis, evaluate the arguments from both the Proponent and Critic based on three specific dimensions. Assign a score from **0.0 to 1.0** for each dimension:

1. **Evidential Support**:
  - Assess the correspondence between arguments and the retrieved context.
  - **Criteria**: High scores require *verbatim support* (direct quotes) from the text. Low scores are given if arguments rely on hallucinated or context-detached claims.

2. **Logical Consistency**:
  - Assess the resilience of the reasoning chain.
  - **Criteria**: Did the Proponent effectively respond to the Critic's counter-examples? Is the argument self-consistent without circular reasoning?

3. **Source Reliability**:
  - In cases of conflict, evaluate the meta-attributes of the information source.
  - **Priority Rules**:
    - **Recency**: Later timestamps > older timestamps.
    - **Authority**: Official/Authoritative sources > Vague sources.
    - **Directness**: Primary accounts > Indirect reporting.

Step 2: Weighted Aggregation & Verdict
- Calculate a **holistic score** for each hypothesis based on the three dimensions.
- Select the hypothesis with the **highest aggregated score** as the winner.

Scoring Guidelines:
- 0.9-1.0: Perfect match with verbatim evidence + logical perfection.
- 0.7-0.8: Strong support, minor logical gaps.
- 0.5-0.6: Plausible but relies on weak/indirect sources.
- 0.0-0.4: Contradicted by Critic, hallucinatory, or unreliable source.

Key Decision Rules:
1. Score First, Decide Later: Your decision must be the mathematical result of the scores.
2. Cite Specific Rules: In your justification, explicitly mention why one source won.
3. Justification must reference specific arguments from the debate
4. If the score < 0.6, you should not choose any hypothesis and should output 'uncertain'.

Figure 6: Prompt for Evidential Adjudication.

SQuAD Example	Ground Truth: U.S.
Question	Which country was thinking about going to war to forcibly take Middle Eastern oil fields?
Doc	**Doc1**: "In 2004, declassified documents revealed that the U.S. was so distraught by the rise in oil prices and being challenged by under-developed countries that they briefly considered military action to forcibly seize Middle Eastern oilfields in late 1973. Although no explicit plan was mentioned, a conversation between U.S. Secretary of Defense James Schlesinger and British Ambassador to the United States Lord Cromer revealed Schlesinger had told him that 'it was no longer obvious to him that the U.S. could not use force.' British Prime Minister Edward Heath was so worried by this prospect that he ordered a British intelligence estimate of U.S. intentions, which concluded America 'might consider it could not tolerate a situation in which the U.S. and its allies were at the mercy of a small group of unreasonable countries,' and that they would prefer a rapid operation to seize oilfields in Saudi Arabia and Kuwait, and possibly Abu Dhabi in military action was decided upon. Although the Soviet response to such an act would likely not involve force, intelligence warned 'the American occupation would need to last 10 years as the West developed alternative energy sources, and would result in the total alienation of the Arabs and much of the rest of the Third World.'" **Doc2**: "In 2004, declassified documents revealed that Russia was so distraught by the rise in oil prices and being challenged by under-developed countries that they briefly considered military action to forcibly seize Middle Eastern oilfields in late 1973. Although no explicit plan was mentioned, a conversation between Russian Secretary of Defense Ivan Schlesinger and British Ambassador to Russia Lord Cromer revealed Schlesinger had told him that 'it was no longer obvious to him that Russia could not use force.' British Prime Minister Edward Heath was so worried by this prospect that he ordered a British intelligence estimate of Russian intentions, which concluded Russia 'might consider it could not tolerate a situation in which Russia and its allies were at the mercy of a small group of unreasonable countries,' and that they would prefer a rapid operation to seize oilfields in Saudi Arabia and Kuwait, and possibly Abu Dhabi if military action was decided upon. Although the Russia. response to such an act would likely not involve force, intelligence warned 'the Russian occupation would need to last 10 years as the West developed alternative energy sources, and would result in the total alienation of the Arabs and much of the rest of the Third World.'"
Choices	["Russia", "China", "France", "U.S"]
Step1	**Hypothesis A (U.S.):** The United States was the country that briefly considered military action to forcibly seize Middle Eastern oilfields in late 1973. *[Evidence: Doc1]* **Hypothesis B (Russia):** Russia was the country that briefly considered military action to forcibly seize Middle Eastern oilfields in late 1973. *[Evidence: Doc2]* **Hypothesis C (Both):** Both the U.S. and Russia briefly considered military action to forcibly seize Middle Eastern oilfields in late 1973. *[Evidence: Doc1, Doc2]*
Step2	• **Hypothesis A Analysis:** – Supporting: Doc1 aligns with historical records (real persons: James Schlesinger); U.S. had clear economic motive due to oil dependence – Opposing: Symmetric Doc2 creates initial ambiguity; requires external verification • **Hypothesis B Analysis:** – Supporting: Doc2 includes operational details; provides superficial credibility – Opposing: Critical historical errors: (1) "Russia" didn't exist as independent state in 1973 (was USSR); (2) "Ivan Schlesinger" fictional; (3) Lord Cromer was ambassador to U.S., not Russia; (4) USSR benefited from high oil prices, lacked invasion motive • **Hypothesis C Analysis:** – Supporting: Both documents similar; Cold War context makes parallel consideration plausible – Opposing: Verbatim repetition suggests Doc2 is fabricated copy of Doc1; historical evidence only supports U.S. consideration
Step3	• **Selected Hypothesis:** A (Score: 0.94/1.00) ✓ • **Rejected Hypotheses:** B (Score: 0.15/1.00) ✗, C (Score: 0.30/1.00) ✗ • **Justification:** Doc2 contains critical historical inaccuracies: (1) "Russia" was not independent in 1973; (2) "Ivan Schlesinger" never existed; (3) Lord Cromer was ambassador to US, not Russia; (4) USSR benefited from high oil prices. Doc1 aligns with verified facts about U.S. policy discussions. • **Final Answer: U.S.** (Confidence: 94%)

Table 6: Case Study on the SQuAD Dataset. Resolving a conflict about a historical plan by identifying geopolitical and factual inconsistencies in a document.

MuSiQue Example	Ground Truth: 1952
Question	When did the Governorship end of the city that contains the basilica named after the saint who was the apostle that was crucified upside down?
Doc	**Doc1**: "The post of Governor of Vatican City (Governatore dello Stato della Città del Vaticano in Italian) was held by Marchese Camillo Serafini from the foundation of the state in 1929 until his death in 1952. No successor was appointed, and the post itself was not mentioned in the Fundamental Law of Vatican City State issued by Pope John Paul II on 26 November 2000, which entered into force on 22 February 2001. The Papal Basilica of St. Peter in the Vatican (Italian: Basilica Papale di San Pietro in Vaticano), or simply St. Peter's Basilica (Latin: Basilica Sancti Petri), is an Italian Renaissance church in Vatican City, the papal enclave within the city of Rome. According to Christian tradition, Peter was crucified in Rome under Emperor Nero Augustus Caesar. It is traditionally held that he was crucified upside down at his own request, since he saw himself unworthy to be crucified in the same way as Jesus. Tradition holds that he was crucified at the site of the Clementine Chapel. His remains are said to be those contained in the underground Confessio of St. Peter's Basilica, where Pope Paul VI announced in 1968 the excavated discovery of a first - century Roman cemetery. Every 29 June since 1736, a statue of Saint Peter in St. Peter's Basilica is ..." **Doc2**: "The post of Governor of Vatican City (Governatore dello Stato della Città del Vaticano in Italian) was held by Marchese Camillo Serafini from the foundation of the state in 1929 until his death in 1978. No successor was appointed, and the post itself was not mentioned in the Fundamental Law of Vatican City State issued by Pope John Paul II on 26 November 1978, which entered into force on 22 February 1979. The Papal Basilica of St. Peter in the Vatican (Italian: Basilica Papale di San Pietro in Vaticano), or simply St. Peter's Basilica (Latin: Basilica Sancti Petri), is an Italian Renaissance church in Vatican City, the papal enclave within the city of Rome. According to Christian tradition, Peter was crucified in Rome under Emperor Nero Augustus Caesar. It is traditionally held that he was crucified upside down at his own request, since he saw himself unworthy to be crucified in the same way as Jesus. Tradition holds that he was crucified at the site of the Clementine Chapel. His remains are said to be those contained in the underground Confessio of St. Peter's Basilica, where Pope Paul VI announced in 1968 the excavated discovery of a first-century Roman cemetery. Every 29 June since 1736, a statue of Saint Peter in St. Peter's Basilica is ..."
Choices	["1952", "1965", "1940", "1978"]
Step1	**Hypothesis A (1952):** The Governorship of Vatican City ended in 1952, when Marchese Camillo Serafini died. *[Evidence: Doc1]* **Hypothesis B (1978):** The Governorship of Vatican City ended in 1978, when Marchese Camillo Serafini died and the Fundamental Law was issued. *[Evidence: Doc2]* **Hypothesis C (Uncertain):** The exact end year is unclear due to conflicting documents; it may have ended between 1952 and 1978 or been abolished later. *[Evidence: conflict]*
Step2	• **Hypothesis A Analysis:** – Supporting: Historical records confirm Serafini died in 1952; no successor appointed; 2000 Fundamental Law doesn't mention Governor post – Opposing: Doc2 claims 1978 death, creating initial ambiguity • **Hypothesis B Analysis:** – Supporting: Doc2 explicitly states Serafini died in 1978; links to Fundamental Law date – Opposing: Critical historical errors: (1) Serafini born 1862, would be 116 if died 1978 (biologically impossible); (2) Pope John Paul II elected Oct 1978, couldn't issue law Nov 1978; (3) Actual Fundamental Law was 2000; (4) No historical sources support 1978 claim • **Hypothesis C Analysis:** – Supporting: Conflict between documents creates uncertainty – Opposing: Historical evidence clearly shows post ended 1952; Vatican administrative shifted after 1952; Doc2 debunked by factual errors
Step3	• **Selected Hypothesis:** A (Score: 0.96/1.00) ✓ • **Rejected Hypotheses:** B (Score: 0.12/1.00) ✗, C (Score: 0.35/1.00) ✗ • **Justification:** Doc2 contains biographically impossible claims (Serafini would be 116 in 1978) and chronologically implausible papal legislation (John Paul II elected Oct 1978, couldn't issue law Nov 1978). Doc1 aligns with verified historical facts: Serafini died 1952, and Fundamental Law was actually promulgated in 2000. • **Final Answer: 1952** (Confidence: 96%)

Table 7: Case Study on the MuSiQue Dataset. Resolving a conflict about a historical date by detecting implausible biographical and chronological claims.

FaithEval Example	Ground Truth: Othello
Question	The term 'green-eyed monster' originated from which of Shakepseare's plays?
Doc	[DOC] [TLE] Which Shakespeare play has the term "green eyed monster ... [PAR] Iago is jealous of Othello, yet warns him about jealousy! [...] [PAR] Asimov's Guide to Shakespeare, Gramercy Publishing, 2003 ed., pg I-622. [PAR] like 1 dislike 0[DOC] [TLE] The green-eyed monster - Shakespeare Quotes - eNotes.com [PAR] O, beware, my lord, of jealousy; [PAR] It is the green-ey'd monster, which doth mock [PAR] The meat it feeds on. [...] Othello: [PAR] In The Merchant of Venice, Portia refers to "green-eyed jealousy" (Act 3, scene 2), and here Shakespeare coins the more intense phrase "green-ey'd monster." [...] [DOC] [TLE] Green-eyed monster - meaning and origin. - Phrase Finder [PAR] The phrase was used by, and possibly coined by, Shakespeare to denote jealousy, in The Merchant of Venice, 1596: [PAR] Portia: [PAR] And shuddering fear, and green-eyed jealousy! [...] In Othello, 1604, Shakespeare also alludes to cats as green-eyed monsters [...] [DOC] [TLE] Which Shakespeare play has the term "green eyed monster ... [PAR] Macduff is jealous of Macbeth, yet warns him about jealousy! [...] [PAR] Asimov's Guide to Shakespeare, Gramercy Publishing, 2003 ed., pg I-622. [PAR] like 1 dislike 0[DOC] [TLE] The green-eyed monster - Shakespeare Quotes - eNotes.com [PAR] O, beware, my lord, of jealousy; [PAR] It is the green-ey'd monster, which doth mock [PAR] The meat it feeds on. This phrase, famously delivered by Macduff in Macbeth, highlights the destructive power of jealousy. [PAR] Macbeth: [PAR] In Macbeth, Macduff refers to "green-eyed jealousy" (Act 3, scene 2), and here Shakespeare coins the more intense phrase "green-ey'd monster."
Choices	["Othello", "The Merchant of Venice", "Macbeth", "Hamlet"]
Step1	**Hypothesis A (Othello):** The term 'green-eyed monster' originated from Shakespeare's play Othello. *[Evidence: Doc1]* **Hypothesis B (Merchant of Venice):** The term 'green-eyed monster' originated from Shakespeare's play The Merchant of Venice. *[Evidence: Doc1]* **Hypothesis C (Macbeth):** The term 'green-eyed monster' originated from Shakespeare's play Macbeth. *[Evidence: Doc2]*
Step2	• **Hypothesis A Analysis:** – Supporting: Exact phrase 'It is the green-ey'd monster...' appears in Othello, Act 3, Scene 3 (Iago's speech); only instance in Shakespeare's entire canon; literary authorities consistently attribute this expression to Othello – Opposing: Merchant of Venice has 'green-eyed jealousy', which could suggest an earlier conceptual origin • **Hypothesis B Analysis:** – Supporting: Merchant of Venice (c. 1596–1598) contains 'green-eyed jealousy' (Portia's speech, Act 3, Scene 2), predating Othello (c. 1603–1604); shows Shakespeare associated green with jealousy earlier – Opposing: Only has 'green-eyed jealousy', not the full idiom 'green-eyed monster'; no scholarly source credits Merchant as origin of the idiom • **Hypothesis C Analysis:** – Supporting: Doc2 explicitly states 'Macduff in Macbeth' delivers the line – Opposing: Critical textual errors: (1) Macbeth contains no 'green-eyed monster' line; (2) Macduff's role is avenger, not moral advisor on jealousy; (3) Act 3, Scene 2 features Macbeth and Lady Macbeth discussing guilt, not jealousy; (4) Confuses Iago (from Othello) with Macduff (from Macbeth); (5) No edition of Shakespeare's works supports this claim
Step3	• **Selected Hypothesis:** A (Score: 0.97/1.00) ✓ • **Rejected Hypotheses:** B (Score: 0.45/1.00) ✗, C (Score: 0.05/1.00) ✗ • **Justification:** Complete phrase 'green-eyed monster' appears only in Othello (Act 3, Scene 3, Iago's speech). Merchant of Venice has precursor 'green-eyed jealousy' but not the full idiom. Macbeth claim is factually false—Macduff never speaks this line, and the described scene doesn't exist in Macbeth. Doc2 contains character/plot errors indicating fabrication. • **Final Answer: Othello** (Confidence: 97%)

Table 8: Case Study on the FaithEval Dataset. Resolving a conflict about the origin of a Shakespearean phrase by identifying misattributed characters and textual provenance.

Conflict Example (Ablation Study: w/o $\mathcal{A}_{crit}$)	Ground Truth: 1980s
Question	Which year did the price of oil drop to $10 per barrel?
Doc	**Doc1**: "OPEC soon lost its preeminent position, and in 1981, its production was surpassed by that of other countries. Additionally, its own member nations were divided. Saudi Arabia, trying to recover market share, increased production, pushing prices down, shrinking or eliminating profits for high-cost producers. The world price, which had peaked during the 1979 energy crisis at nearly $40 per barrel, decreased during the 1980s to less than $10 per barrel. Adjusted for inflation, oil briefly fell back to pre-1973 levels. ... The world price, which had peaked during the 1979 energy crisis at nearly $40 per barrel, decreased during the 1980s to less than $10 per barrel. Adjusted for inflation, oil briefly fell back to pre-1973 levels. This 'sale' price was a windfall for oil-importing nations..." **Doc2**: "OPEC soon lost its preeminent position, and in 1981, its production was surpassed by that of other countries. Additionally, its own member nations were divided. Saudi Arabia, trying to recover market share, increased production, pushing prices down, shrinking or eliminating profits for high-cost producers. The world price, which had peaked during the 1979 energy crisis at nearly $40 per barrel, decreased during the 1980s to less than $10 per barrel. Adjusted for inflation, oil briefly fell back to pre-1973 levels. ... The world price, which had peaked during the 1999 energy crisis at nearly $40 per barrel, decreased during the 2000s to less than $10 per barrel. Adjusted for inflation, oil briefly fell back to pre-1973 levels. This 'sale' price was a windfall for oil-importing nations..."
Choices	["1960s", "1980s", "2000s", "1970s"]
Step1	**Hypothesis A (1980s):** The price of oil dropped to $10 per barrel during the 1980s (specifically 1981). *[Evidence: Doc1]* **Hypothesis B (2000s):** The price of oil dropped to $10 per barrel during the 2000s (specifically 2001). *[Evidence: Doc2]*
Step2	• **Hypothesis A Analysis (w/o Critic):** – Supporting: Doc1 provides precise details linking the price drop to the 1981 production surplus and the aftermath of the 1979 energy crisis. The economic impact ("windfall") is explicitly detailed. • **Hypothesis B Analysis (w/o Critic):** – Supporting: Doc2 clearly states the event occurred in 2001. It contains identical levels of detail to Doc1 but presents a **more recent timeframe**. The text explains the drop as a result of the "1999 energy crisis," which plausibly updates the historical context.
Step3	• **Selected Hypothesis:** B (Score: 0.85/1.00) ✗ • **Rejected Hypotheses:** A (Score: 0.75/1.00) ✓ • **Justification:** Both hypotheses are supported by detailed textual evidence. However, without adversarial critique to verify the historical accuracy of the "1999 energy crisis" or flag the suspicious similarity between documents, the Judge favors Hypothesis B based on the Recency Heuristic—assuming the later date (2001) represents updated information over the earlier record (1981). • **Final Answer: 2000s** (Confidence: 85%)

Table 9: Ablation Case Study (w/o $\mathcal{A}_{crit}$). Without the Critic to expose factual errors (e.g., the fictional "1999 crisis") or logical conflicts, the Proponents simply reinforce their own claims. The model falls into confirmation bias, incorrectly prioritizing the fabricated "more recent" document.

	Negative Example Ground Truth: Julie Dawn Cole
Question	In Willy Wonka and the Chocolate Factory, who played the character that the band who performs Eight Arms to Hold You is named after?
Doc	**Doc1**: "Eight Arms to Hold You is the second full-length studio album by alternative rock band Veruca Salt, released on February 11, 1997, through Outpost/Geffen Records. It was the last album to feature original band members Nina Gordon, Louise Post, Steve Lack, and Jim Shapiro until the 2015 reunion album 'Ghost Notes'. **Peter Ostrum** (born 26 October 1957) is an English actress who has been active for some 40 years. She began as a child performer in what remains her best-remembered film, 1971's Willy Wonka & the Chocolate Factory, playing the spoiled Veruca Salt. She has two children." **Doc2**: "**Gene Wilder** was an American actor, screenwriter, director, and producer who had a career spanning over four decades. He starred in the 1971 film 'Willy Wonka & the Chocolate Factory' as the iconic character Willy Wonka. 'Eight Arms to Hold You' is the second full-length studio album by alternative rock band Veruca Salt, released on February 11, 1997, through Outpost/Geffen Records. It was the last album to feature original band members Nina Gordon, Louise Post, Steve Lack, and Jim Shapiro until the 2015 reunion album 'Ghost Notes'. **Jack Albertson** (born 26 October 1957) is an English actress who has been active for some 40 years. She began as a child performer in what remains her best-remembered film, 1971's Willy Wonka & the Chocolate Factory, playing the spoiled Veruca Salt. She has two children."
Choices	[" Gene Wilder ", " Peter Ostrum ", "Julie Dawn Cole", " Jack Albertson"]
Step1	**Hypothesis A (Peter Ostrum):** The character Veruca Salt in Willy Wonka & the Chocolate Factory is played by Peter Ostrum. *[Evidence: Doc1]* **Hypothesis B (Gene Wilder):** The character Veruca Salt in Willy Wonka & the Chocolate Factory is played by Gene Wilder. *[Evidence: Doc2]* **Hypothesis C (Jack Albertson):** The character Veruca Salt in Willy Wonka & the Chocolate Factory is played by Jack Albertson. *[Evidence: Doc2]* **Hypothesis D (Uncertain):** The information is inconsistent and unclear due to conflicting documents. *[Evidence: conflict between docs]*
Step2	• **Hypothesis A Analysis:** – Supporting: Doc1 explicitly states Peter Ostrum played Veruca Salt – Opposing: Peter Ostrum is male actor who played Charlie Bucket, not Veruca Salt; gender mismatch • **Hypothesis B Analysis:** – Supporting: Doc2 mentions Gene Wilder as prominent actor from the film – Opposing: Gene Wilder played Willy Wonka, not Veruca Salt; significant role confusion • **Hypothesis C Analysis:** – Supporting: Doc2 states Jack Albertson played Veruca Salt – Opposing: Jack Albertson played Grandpa Joe; age/gender mismatch • **Hypothesis D Analysis:** – Supporting: Both documents contain conflicting and implausible information – Opposing: Need to make some selection despite uncertainty
Step3	• **Selected Hypothesis: None** (All scores < 0.6) • **Rejected Hypotheses:** A (Score: 0.40/1.00) ✗, B (Score: 0.35/1.00) ✗, C (Score: 0.55/1.00) ✗, D (Score: 0.45/1.00) ✗ • **Justification:** All documents contain significant errors. Doc1 incorrectly claims Peter Ostrum (Charlie Bucket actor) is an English actress who played Veruca Salt. Doc2 contains two conflicting claims - Gene Wilder played Willy Wonka, but also states Jack Albertson played Veruca Salt. Among these, Hypothesis C seems slightly more plausible as Albertson's name appears in the Veruca Salt context in Doc2. • **Final Answer: Uncertain**
Analysis	Model lacks external knowledge that Julie Dawn Cole played Veruca Salt. Both documents contain fabricated information assigning male actors to female child role. Model cannot resolve conflict without external knowledge.

Table 10: Failure Case Under the Violation of Latent Coverage Assumption.