Mitigating Context-Memory Conflicts in LLMs through Dynamic Cognitive Reconciliation Decoding

Anonymous ACL submission

Abstract

Large language models accumulate extensive parametric knowledge through pre-training. However, knowledge conflicts occur when outdated or incorrect parametric knowledge conflicts with external knowledge in the context. Existing methods address knowledge conflicts through contrastive decoding, but in conflict-free scenarios, static approaches disrupt output distribution. Other dynamic decoding methods attempt to measure the degree of conflict but still struggle with complex real-world situations. In this paper, we propose a two-stage decoding method called Dynamic Cognitive Reconciliation Decoding (DCRD), to predict and mitigate context-memory conflicts. DCRD first analyzes the attention map to assess context fidelity and predict potential conflicts. Based on this prediction, the input is directed to one of two decoding paths: (1) greedy decoding, or (2) context fidelity-based dynamic decoding. This design enables DCRD to handle conflicts efficiently while maintaining high accuracy and decoding efficiency in conflict-free cases. Additionally, to simulate scenarios with frequent knowledge updates, we constructed ConflictQA, a knowledge conflict QA benchmark. Experiments on four LLMs across six QA datasets show that DCRD outperforms all baselines, achieving state-of-the-art performance.

1 Introduction

Large language models (LLMs) assimilate extensive textual knowledge during pre-training (Radford and Narasimhan, 2018; Kenton and Toutanova, 2019; Soldaini et al., 2024), demonstrating exceptional performance in knowledge-intensive tasks (Jiang et al., 2024; Zhang et al., 2023). Despite this, LLMs still face challenges, including real-time knowledge updates (Wang et al., 2024a), learning rare facts, and handling dynamic information (Chen et al., 2022; Wang et al., 2023). To address these, researchers have introduced Retrieval-Augmented Generation (RAG) techniques (Lewis et al., 2020; Sun et al., 2024; Yoran et al., 2024; Xu et al., 2024a), which combine the model's internal knowledge with external information retrieved from external sources. Although RAG methods show great promise, effectively managing conflicts when LLMs encounter contradictory information from different sources remains a challenge (Hou et al., 2024; Su et al., 2024). When the retrieved information conflicts with the model's parametric memory, a context-memory conflict occurs. In such instances, the model tends to overly rely on its internal knowledge, thereby undermining the fidelity of external information (Jin et al., 2024).

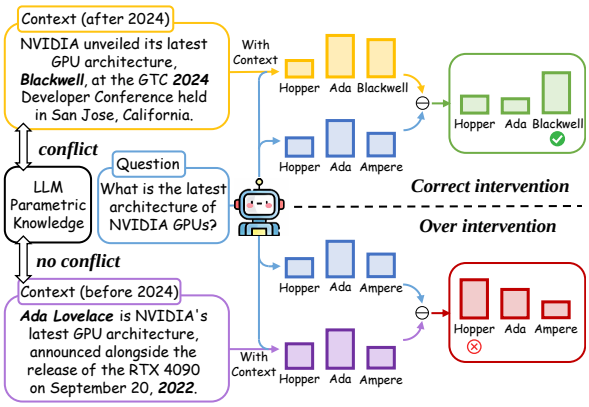

Figure 1: In conflict scenarios, context-aware decoding enhances reasoning by amplifying distribution differences; In conflict-free scenarios, it lead to incorrect answers due to excessive interference with the output distribution. ⊖ represents context-aware decoding.

Recently, various methods have been proposed to mitigate knowledge conflicts. Some focus on fine-tuning models to address these conflicts (Gekhman et al., 2023; Neeman et al., 2023), but their applicability is limited and they often compromise the model's general capabilities. Another approach involves decoding strategies, such as Context-Aware Decoding (CAD) (Shi et al., 2024), which amplifies the distinction in output probabilities between using and not using context,

thereby encouraging the LLM to focus more on the context during generation, as shown in Figure 1. However, in real-world scenarios, conflicts arise only in a subset of inputs. In most low-conflict cases, **over-intervention** of output distribution can introduce bias, resulting in performance degradation. This phenomenon is consistent with cognitive dissonance theory (Harmon-Jones and Mills, 1999; Bem, 1967; Harmon-Jones and Mills, 2019): when external information aligns with prior knowledge, the brain naturally accepts it. However, forcibly correcting non-conflicting information may lead to inconsistencies or errors. This raises a critical question: **Can we achieve cognitive reconciliation in complex context-memory conflict scenarios?**

To this end, we introduce a novel method called DCRD: **D**ynamic **C**ognitive **R**econciliation **D**ecoding. DCRD aims to mitigate cognitive dissonance in context-memory conflicts through improvements in two dimensions: (1) **Prior to decoding**, we introduce a conflict predictor that directs conflicting and non-conflicting information along separate decoding paths. By capturing the attention relationships between the newly generated tokens and the context tokens, we measure contextual fidelity, using this as the foundation for conflict classification. (2) **During decoding**, DCRD responds quickly with regular decoding for low-conflict information, while dynamically adjusting the decoding process based on contextual fidelity for high-conflict information. In this process, higher contextual fidelity signals lower conflict, warranting reduced intervention, while lower fidelity signals higher conflict, requiring increased intervention.

We extensively evaluated DCRD on knowledge conflict question-answering datasets: Counterfacts (Longpre et al., 2021) and NQ-Swap (Longpre et al., 2021), as well as general question-answering datasets: Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaQA(Joshi et al., 2017) and SQuAD (Rajpurkar et al., 2016). Additionally, we have developed a new benchmark, ConflictQA, employing a generative approach to simulate real-world knowledge conflicts. The benchmark contains 4,466 instances, both conflicting and non-conflicting, along with their respective knowledge sources, and we conducted thorough analysis and evaluation on it. We evaluated DCRD on several open-source LLMs, including Llama2-7b (Touvron et al., 2023), Llama2-13b (Touvron et al., 2023), Llama3-8b (Dubey et al., 2024), and Mistral-7b (Jiang et al., 2023). The experimental results show that DCRD outperforms other decoding methods across all datasets, achieving state-of-the-art performance.

Our work makes the following contributions:

- We propose DCRD, a method designed to alleviate context-memory conflicts. DCRD predicts conflicts and routes the information into two decoding paths: (1) regular decoding, and (2) dynamic contrastive decoding, which adaptively enhances intervention for conflicting information and reduces intervention for low-conflict information.
- We present ConflictQA, a knowledge conflict benchmark that simulates real-world scenarios, containing 4,466 conflict and non-conflict instances along with their knowledge sources.
- We conducted extensive experiments and analysis across various LLMs and multiple datasets. The results demonstrate that DCRD outperforms previous decoding approaches in both high-conflict and general scenarios, achieving state-of-the-art performance.

2 Related Work

Knowledge Conflicts. Current research classifies knowledge conflicts into three categories: intra-memory, context-memory, and inter-context (Xu et al., 2024b). Our research focuses on context-memory conflicts, particularly in the context of retrieval-augmented generation (Wu et al., 2024). We aim to ensure that the model generates responses based on the current context, rather than relying on outdated or erroneous parametric knowledge. In existing research on mitigating context-memory conflicts, prompting-based methods (Zhou et al., 2023; Peng et al., 2023) heavily depend on prompt design, whereas fine-tuning-based methods (Li et al., 2023a; Gekhman et al., 2023; Xue et al., 2023) are task-specific, restricting their generalizability and resulting in significant computational overhead. In contrast, our approach employs classification and decoding strategies to guide model generation during inference, effectively mitigating knowledge conflicts without extra fine-tuning or prompt dependency.

Contrastive Decoding. Currently, many studies focus on contrastive learning methods (Robinson et al., 2021; Khosla et al., 2020) to guide models towards specific output preferences (Li et al., 2023b; O'Brien and Lewis, 2023). Context-aware decoding (CAD) (Shi et al., 2024) leverages a con-

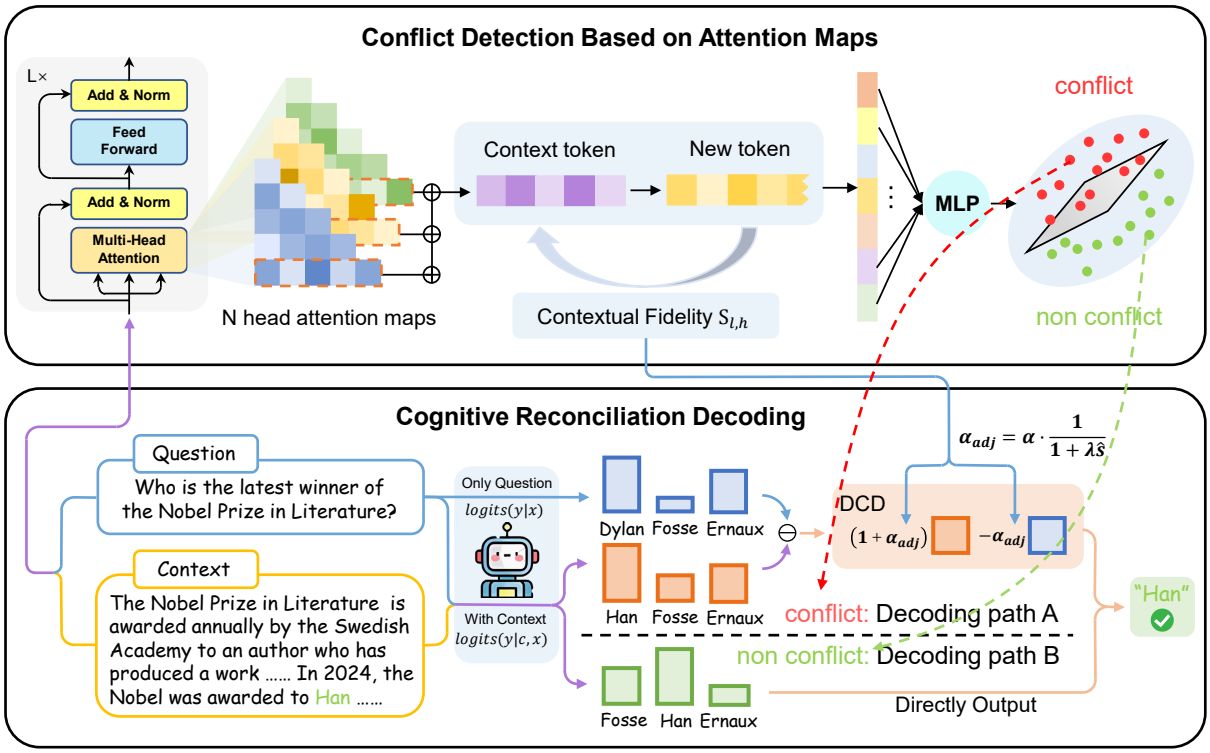

Figure 2: Overview of DCRD, a two-stage dynamic decoding method designed to mitigate context-memory conflicts, including *Conflict Prediction Based on Attention Maps* and *Cognitive Reconciliation Decoding*. We employ a dynamic routing approach, where conflict-free inputs are directed to the greedy decoding path (B), while conflicting inputs are routed to the dynamic contrastive decoding path (A) based on conflict prediction results.

trastive output distribution that amplifies the disparity in output probabilities when the model is used with and without context, thereby significantly improving the model's faithfulness to the context. COIECD (Yuan et al., 2024) identifies knowledge conflicts by measuring changes in distribution entropy at the token level and controls the decoding process based on whether the current token conflicts. ADACAD (Wang et al., 2024b) computes the Jensen-Shannon divergence between output distributions with and without context, then dynamically tunes hyperparameters. These methods introduce unintended perturbations to token distributions and rely on a simplistic conflict modeling approach, reducing their effectiveness in real-world contexts. To address this, our approach leverages attention maps to represent conflicting and non-conflicting contexts, predicts knowledge conflicts at the sentence level, and dynamically adjusts the contrastive decoding strategy based on conflict severity.

3 Methodology

Cognitive dissonance theory (Harmon-Jones and Mills, 2019) posits that **when conflict-free information undergoes unnecessary corrective processes, it may induce cognitive dissonance, leading to inconsistent or erroneous outputs.** Inspired by this, we propose DCRD, which first predicts the occurrence of context-memory conflicts. Based on this prediction, DCRD employs a dynamic routing strategy: for conflict-free inputs, it follows the default decoding path, while for inputs with higher conflict, it adjusts the output by dynamically increasing attention to the context. As shown in Figure 2, our method consists of two modules: **Conflict Prediction Based on Attention Maps** and **Cognitive Reconciliation Decoding**.

3.1 Conflict Prediction with Attention Maps

We assume that when an LLM relies more on context during generation, conflicts are less likely; if it strays from the context, conflicts are more likely. Based on this assumption, we frame the prediction of context-memory conflicts as a binary classification problem and introduce *contextual fidelity* as a key feature for classification, which is extracted from the attention maps produced by LLMs during input processing. Specifically, a GPT-based LLM consists of L Transformer layers, each with N parallel attention heads to capture the relationships between positions. During text generation,

the attention weights indicate the current token's reliance on the context. The attention weight α_i of the i-th key K_i for the query Q is defined as:

$$\alpha_i = \frac{\exp\left(\frac{QK_i^T}{\sqrt{d_k}}\right)}{\sum_{j=1}^{n}\exp\left(\frac{QK_j^T}{\sqrt{d_k}}\right)} \tag{1}$$

Given the context $C = [c_1, c_2, \ldots, c_N]$ as the model input and the output sequence $O = [o_1, o_2, \ldots, o_T]$, the attention weights for each attention head can be expressed as follows:

$$\boldsymbol{\alpha}_c^h = \frac{1}{N}\sum_{i=1}^{N}\alpha_i^h, \quad \boldsymbol{\alpha}_o^h = \frac{1}{T}\sum_{i=N+1}^{N+T}\alpha_i^h \tag{2}$$

where $\boldsymbol{\alpha}_c^h$ and $\boldsymbol{\alpha}_o^h$ denote the average attention weights for the context and output, respectively, for the h-th attention head.

To capture the model's reliance on context when generating new sequences, we define the *contextual fidelity* score $S_{l,h}$ as follows:

$$S_{l,h} = \frac{\boldsymbol{\alpha}_o^{l,h}}{\boldsymbol{\alpha}_c^{l,h} + \boldsymbol{\alpha}_o^{l,h}} \tag{3}$$

We train a single-layer MLP as classifier, using the aggregated contextual fidelity scores from L Transformer layers and H attention heads as input, to predict the conflict result $\hat{y}$:

$$\hat{y} = classifier\left(flatten(S_{l,h})\right) \tag{4}$$

In the subsequent decoding process, $\hat{y}$ will act as the routing criterion, guiding the input toward different decoding strategies according to the level of conflict, thereby alleviating the negative impact of knowledge conflicts on the output distribution. See more details in Appendix A.4

3.2 Cognitive Reconciliation Decoding

To mitigate context-memory conflicts, we employ dynamic contrastive decoding to guide the model's output. Specifically, given the context c, the question x, and the output $y_{<t}$, we define:

$$p_1(y_t) = p_\theta(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}) \tag{5}$$

$$p_2(y_t) = p_\theta(y_t|\boldsymbol{x}, \boldsymbol{c}, \boldsymbol{y}_{<t}) \tag{6}$$

where p_1 represents the output distribution based exclusively on the model's parameters, while p_2 integrates contextual information. The purpose of contrastive decoding is to amplify the difference between p_1 and p_2, thereby enhancing the influence of contextual knowledge and diminishing reliance on the model's inherent memory.

CAD (Shi et al., 2024) relies on a fixed hyperparameter α to balance the contrast between the model's parametric knowledge and contextual knowledge. However, this fixed approach struggles to dynamically handle varying levels of conflict. To address this, our method adaptively adjusts α for each token, depending on contextual fidelity:

$$\alpha_{\text{adj}} = \alpha \cdot \frac{1}{1 + \lambda\hat{s}} \tag{7}$$

where $\hat{s}$ represents the normalized context fidelity, and λ is a hyperparameter that controls the sensitivity between context fidelity and α.

Our method introduces a dynamic adaptation mechanism that facilitates fine-grained conflict mitigation at the token level, effectively mitigating conflicts of varying severity between contextual information and parametric knowledge. The dynamic contrastive decoding can be defined as:

$$\begin{aligned} p_3(y_t) = \text{softmax}\big[&\left(1 + \alpha_{\text{adj}}\right) \cdot \text{p}_\theta\left(y_t|\boldsymbol{c}, \boldsymbol{x}, \boldsymbol{y}_{<t}\right) \\ &- \alpha_{\text{adj}} \cdot \text{p}_\theta\left(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}\right)\big] \end{aligned} \tag{8}$$

Notably, we integrate a routing mechanism into our decoding strategy. Using the conflict prediction $\hat{y}$ from Section 3.1, we classify decoding paths into two types: greedy decoding (GD) for conflict-free inputs and dynamic contrastive decoding (DCD) for conflicting inputs. Given the context c, the question q, and the conflict prediction result $\hat{y}$, the routing process to generate the answer a can be defined as:

$$answer = \begin{cases} DCD(q, c), & \text{if } \hat{y} \text{ is true} \\ GD(q, c), & \text{if } \hat{y} \text{ is false} \end{cases} \tag{9}$$

Our method achieves cognitive reconciliation in two dimensions: (1) minimizing the interference of contrastive decoding on conflict-free inputs, and (2) dynamically balancing the output distributions of contextual and parametric knowledge. These strategies effectively improve both the accuracy and the stability of the decoding.

3.3 ConflictQA Benchmark

Currently, many studies on mitigating context-memory conflicts predominantly rely on summarization datasets for evaluation, such as CNN-DM (See et al., 2017), XSUM (Narayan et al.,

2018). To evaluate these methods more efficiently and comprehensively in scenarios closer to real-world scenarios, especially in question-answering scenarios where knowledge is frequently updated and model updates are delayed, we have developed a knowledge conflict question-answering dataset.

Extracting Knowledge. Our contextual knowledge is derived from Wikidata (Vrandečić and Krötzsch, 2014), a comprehensive and high-quality knowledge base. Structured knowledge in a knowledge base can be represented as triples (s, r, o), where s is the subject, r is the relation, and o is the object. Given a specific question q, we retrieve the relevant subgraph $\mathcal{G}_{\text{sub}} = \{(s_i, r_i, o_i)\}_{i=1}^{N}\}$ from the knowledge base.

Conflict Knowledge Construction. To create knowledge conflicts, we modify a triple (s, r, o) containing the answer in the subgraph $\mathcal{G}_{\text{sub}}$ to (s, r, o'), where o' is an entity that is semantically similar to o and shares the same type, resulting in a new subgraph $\mathcal{G}'_{\text{sub}}$.

Context Generation. Unlike methods based on entity replacement (Longpre et al., 2021), we leverage LLMs to generate context that is more linguistically coherent and rich in background information. Given an original sample $\{q, G_{\text{sub}}, a\}$ and its modified counterpart $\{q, G'_{\text{sub}}, a'\}$, we insert them into the prompt template and input them into LLMs to generate the conflict-free context c_{non} and the conflict context c_{conf}.

Quality Control. To ensure the quality of the data, we perform a thorough review throughout the dataset construction process. First, we filter out questions with no answer or more than three answers. In conflict knowledge construction, we further validate the generated conflicts, ensuring that the conflicting entities are of the same type as the original answer but contain contradictory content. In context generation, we match the generated results with the subgraph to ensure that each piece of knowledge is accurately integrated into the context. See more details in Appendix A.3.

4 Experiments

4.1 Experimental Setup

Datasets. We conducted extensive evaluations of DCRD using both standard question-answering datasets: Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017) and SQuAD (Rajpurkar et al., 2016), as well as knowledge conflict question-answering datasets: Counterfacts (Longpre et al., 2021) and NQ-Swap (Longpre et al., 2021). Additionally, as described in Section 3.3 we have developed a new benchmark, ConflictQA, employing a generative approach to simulate real-world knowledge conflicts. Further dataset details are provided in Appendix A.1.

LLM & Baselines. We conducted experiments on four open-source LLMs, including Llama2-7b (Touvron et al., 2023), Llama2-13b (Touvron et al., 2023), Llama3-8b (Dubey et al., 2024), and Mistral-7b (Jiang et al., 2023). And we consider four decoding strategies as baselines, including: Context-aware decoding (CAD) (Shi et al., 2024), COIECD (Yuan et al., 2024) and ADACAD (Wang et al., 2024b). Greedy decoding is the standard decoding strategy. CAD mitigates knowledge conflicts by comparing the output distributions with and without context, thereby controlling the decoding process. COIECD and ADACAD are two dynamic decoding strategies. COIECD adjusts the decoding process by imposing constraints based on the entropy of contextual information, while ADACAD regulates the process by calculating the Jensen-Shannon Divergence between the output distributions with and without context. The details of each baseline are described in Appendix A.2.

Implementation Details. To ensure a fair comparison, we standardized the sampling hyperparameters across DCRD and all baseline methods, using the simplest zero-shot prompt template. For CAD, we set $\alpha = 1$; for COIECD, we set λ and α to 0.25 and 1, respectively; for DCRD, we set both λ and α to 1. More details can be seen in Appendix A.4.

Evaluation Metrics. Traditional exact match (EM) methods are increasingly insufficient for complex context-based question answering tasks. To provide a more comprehensive evaluation, we developed an automated evaluation framework, designed a prompt-based evaluation template, and employed GPT-4o[1] to evaluate the answers. Details are provided in Appendix A.5.

4.2 Main Results

Results on Knowledge Conflict QA benchmark. We conducted experiments on Counterfacts and NQ-Swap, simulating high-conflict contextual environments, to evaluate the handling capacity of DCRD

[1]GPT-4o is from https://openai.com/

Model	Decoding	General QA			Knowledge Conflict QA		Our Benchmark	Avg.
		NQ	SQUAD	TriviaQA	NQ-Swap	Counterfacts	ConflictKG	
Llama2-7B	Greedy	51.9	71.6	81.4	36.5	33.8	66.0	56.9
	CAD	50.3$_{-1.6}$	67.7$_{-3.9}$	58.5$_{-22.9}$	47.5$_{11.0}$	47.4$_{13.6}$	77.5$_{11.5}$	58.2$_{-1.3}$
	COIECD	59.9$_{8.0}$	76.0$_{4.4}$	77.8$_{-3.6}$	48.9$_{12.4}$	48.4$_{14.6}$	75.9$_{9.9}$	64.5$_{5.6}$
	ADACAD	65.4$_{13.5}$	74.6$_{3.0}$	82.3$_{0.9}$	46.1$_{9.6}$	44.3$_{10.5}$	72.0$_{6.0}$	64.1$_{5.2}$
	DCRD (Ours)	**68.4**$_{16.5}$	**83.2**$_{11.6}$	**83.9**$_{2.5}$	**54.2**$_{17.7}$	**57.4**$_{23.6}$	**81.1**$_{15.1}$	**71.4**$_{14.5}$
Llama2-13B	Greedy	64.3	73.9	**86.1**	36.4	53.0	73.7	64.6
	CAD	51.3$_{-13.0}$	72.7$_{-1.2}$	48.9$_{-37.2}$	53.9$_{17.5}$	62.3$_{9.3}$	80.1$_{6.4}$	61.5$_{-3.1}$
	COIECD	68.6$_{4.3}$	78.9$_{5.0}$	85.8$_{-0.3}$	54.0$_{17.6}$	56.8$_{3.8}$	76.8$_{3.1}$	70.2$_{5.6}$
	ADACAD	68.2$_{3.9}$	76.5$_{2.6}$	85.7$_{-0.4}$	62.6$_{26.2}$	62.5$_{9.5}$	77.4$_{3.7}$	72.2$_{7.6}$
	DCRD (Ours)	**71.4**$_{7.1}$	**79.5** $_{5.6}$	**86.1**$_{0.0}$	**65.5**$_{29.1}$	**65.2**$_{12.2}$	**86.0**$_{12.3}$	**75.6**$_{11.0}$
Llama3-8B	Greedy	67.4	87.2	**88.6**	47.2	48.1	67.9	67.7
	CAD	51.0$_{-16.4}$	72.6$_{-14.6}$	55.8$_{-32.8}$	56.0$_{8.8}$	57.6$_{9.5}$	68.4$_{-4.7}$	60.2$_{-7.5}$
	COIECD	68.4$_{1.0}$	86.9$_{-0.3}$	87.5$_{-1.1}$	52.1$_{4.9}$	52.1$_{4.0}$	70.3$_{2.4}$	69.9$_{2.2}$
	ADACAD	65.2$_{-2.2}$	87.1$_{-0.1}$	86.3$_{-2.3}$	58.3$_{11.1}$	59.0$_{10.9}$	77.5$_{9.6}$	71.0$_{3.3}$
	DCRD (Ours)	**73.4**$_{6.0}$	**88.9**$_{1.7}$	88.0$_{-0.6}$	**65.3**$_{18.1}$	**67.3**$_{19.2}$	**79.5**$_{11.6}$	**77.1**$_{9.4}$
Mistral-7B	Greedy	58.8	67.5	75.3	45.6	41.6	66.2	59.2
	CAD	52.6$_{-6.2}$	25.2$_{-42.3}$	23.3$_{-52.0}$	48.1$_{2.5}$	45.8$_{4.2}$	36.3$_{-29.9}$	38.6$_{-20.6}$
	COIECD	59.1$_{0.3}$	48.1$_{-19.4}$	58.6$_{-16.7}$	56.0$_{10.4}$	54.0$_{12.4}$	69.1$_{2.9}$	57.5$_{-1.7}$
	ADACAD	58.1$_{-0.7}$	**75.2**$_{7.7}$	74.7$_{-0.6}$	51.7$_{6.1}$	49.8$_{8.2}$	67.4$_{1.2}$	62.8 $_{3.6}$
	DCRD (Ours)	**60.9**$_{2.1}$	69.1$_{1.6}$	**77.6**$_{2.3}$	**58.3**$_{12.7}$	**57.9**$_{16.3}$	**72.6**$_{6.4}$	**66.1**$_{6.9}$

Table 1: Conflict mitigation performance on general QA datasets, knowledge conflict QA datasets, and ConflictQA. DCRD outperforms all baselines. Greedy represents greedy decoding. The number in the subscript indicates the difference in greedy decoding compared to the baseline.

for high-conflict information. The results, shown in Table 1, demonstrate that DCRD consistently outperforms greedy decoding, CAD, COIECD, and ADACAD. For example, on NQ-Swap, DCRD outperforms the baseline greedy decoding by 17.7%, 29.1%, 12.7%, and 18.1% on Llama2-7b, Llama2-13b, Llama3-8b, and Mistral-7b, respectively. It also exceeds the baseline CAD by 6.7%, 11.6%, 10.2%, and 9.3%, respectively. In scenarios with frequent knowledge conflicts, DCRD achieves significant improvements. This highlights the effectiveness of our dynamic decoding strategy, which adjusts intervention strength based on the level of conflict, thereby enhancing the model's ability to manage complex conflict scenarios.

Results on General QA benchmark. We conducted experiments on NQ, SQuAD and TriviaQA, simulating low-conflict contexts, to evaluate the handling capacity of DCRD for low-conflict information. The results, presented in Table 1, show that DCRD consistently outperforms all baselines. For example, on the NQ dataset, DCRD surpasses the baseline greedy decoding by 16.5%, 7.1%, 2.1%, and 6.0% on Llama2-7b, Llama2-13b, Llama3-8b, and Mistral-7b, respectively. It also outperforms the baseline CAD by 18.1%, 20.1%, 8.3%, and 22.4%, respectively. It is noteworthy that CAD underperforms compared to the baseline greedy decoding in low-conflict scenarios, with an average accuracy drop of 8.1% across all models. In contrast, DCRD consistently surpasses all baselines in every low-conflict scenario. This clearly highlights the superiority of our approach, which employs a conflict prediction mechanism to route conflicting and non-conflicting information along distinct decoding paths. Meanwhile, DCRD surpasses the two dynamic contrastive decoding methods, ADACAD and COIECD, underscoring the rationale and effectiveness of dynamically adjusting the decoding process based on contextual fidelity.

Results on ConflictQA. We conducted experiments on ConflictQA, simulating complex scenarios involving context-memory conflicts, which commonly arise in situations where knowledge frequently updates and model updates lag behind. The experimental results, presented in Table 1, show that DCRD outperforms all baselines. The average results of DCRD across the four LLMs exceed baseline greedy decoding, CAD, COIECD, and ADACAD by 11.3%, 14.2%, 6.8%, and 6.2%, respec-

Decoding	Llama2-7b		Llama-13b		Llama3-8b		Mistral-7b	
	conflict	non conflict	conflict	non conflict	conflict	non conflict	conflict	non conflict
Greedy	53.8	78.1	65.7	81.5	58.3	77.5	56.4	75.9
CAD	72.9$_{19.1}$	82.0$_{3.9}$	77.9$_{12.2}$	83.1$_{1.6}$	66.1$_{7.8}$	70.7$_{-6.8}$	49.7$_{-6.7}$	22.9$_{-53.0}$
COIECD	68.8$_{15.0}$	82.9$_{4.8}$	67.7$_{2.0}$	85.8$_{4.3}$	67.2$_{8.9}$	77.6$_{0.1}$	67.9$_{11.5}$	70.3$_{-5.6}$
ADACAD	61.4$_{7.6}$	82.7$_{9.2}$	70.7$_{5.0}$	84.1$_{2.6}$	66.2$_{7.9}$	74.4$_{-3.1}$	59.4$_{3.0}$	75.4$_{-0.5}$
DCRD (Ours)	**74.8**$_{21.0}$	**87.3**$_{9.2}$	**82.5**$_{16.8}$	**89.5**$_{8.0}$	**73.8**$_{15.5}$	**85.0**$_{7.5}$	**68.0**$_{11.6}$	**77.1**$_{1.2}$

Table 2: We conduct experiments on ConflictQA to analyze the performance of five decoding methods in both conflict and non-conflict scenarios.

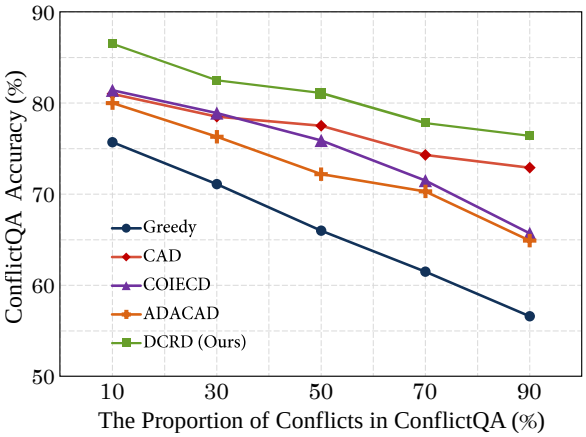

Figure 3: We experiment on ConflictQA by varying the conflict data proportion. DCRD consistently outperforms other baselines, with smaller fluctuations in performance as the conflict ratio changes.

Layer	Conflict Predictor	Results
Random		
-	50	71.5$_{-9.6}$
Hidden state		
16th Layer	76.5	77.8$_{-4.6}$
32nd Layer	73.2	77.1$_{-4.0}$
Attention maps (Ours)		
32 Layers	84.7	81.1

Table 3: The Impact of Different Classifier Settings on Llama2-7b. "Random" refers to randomly generated classification results."Result" refers to the accuracy achieved by applying the corresponding classifier settings on ConflictQA. The red subscript represents the difference from our method.

tively. This indicates that DCRD can emulate how the brain selectively focuses and dynamically adjusts its decision-making process when confronted with conflicting information, flexibly determining when correction is necessary, rather than indiscriminately processing all data. As a result, it maintains high robustness even in complex conflict scenarios. Notably, DCRD intelligently selects decoding paths based on the level of conflict. Unlike other dynamic decoding methods, it minimizes unnecessary computational overhead, thereby improving reasoning efficiency and accuracy, while showcasing its distinctive advantages in complex scenarios.

4.3 Analysis

How does DCRD perform on conflict and non-conflict samples? As shown in Table 2, Greedy decoding performs significantly worse in conflict scenarios compared to non-conflict, with an average drop of 21.5%. This highlights that decoding strategies without specific optimizations fail to effectively address knowledge conflicts. In non-conflict scenarios, CAD performs, on average, 13.6% worse than Greedy decoding. This degradation is especially pronounced in the NQ and SQuAD datasets (see Table 1), indicating that CAD can hinder the model's original decoding capability in low-conflict situations. In contrast, our method leverages the strengths of both decoding strategies, dynamically balancing the decoding process across various scenarios. Specifically, the improvement of DCRD in non-conflict scenarios highlights the effectiveness of routing low-conflict information to the Greedy decoding path, while the performance boost in conflict scenarios validates the soundness of our dynamic decoding strategy.

How robust is DCRD across different conflict proportions? As shown in Figure 3, we conducted experiments on ConflictQA with varying conflict proportions. The results show that the performance of COIECD and ADACAD fluctuates significantly as the conflict proportion increases, with drops of 15.7% and 15.1%, respectively. In contrast, DCRD's performance decreases by only 10.1%, maintaining relative stability. This advantage arises from DCRD's routing mechanism and dynamic decoding strategy, which allow it to maintain stable performance in complex scenarios and demonstrate stronger robustness and adaptability when handling varying degrees of knowledge updates.

Context: The Super Bowl 50 Halftime Show took place on February 7, 2016, at Levi's Stadium in Santa Clara, California... It was headlined by the British rock group Coldplay with special guest performers Beyoncé and Bruno Mars ...
Question: Who is playing the halftime show at super bowl 2016?
Ground Truth: Coldplay with special guest performers Beyoncé and Bruno Mars.
Greedy: Coldplay\n \n ###### Hint 1:\n \nThe Super Bowl 50 Halftime Show took place on February 7, 2.
ADACAD: The Super Bowl 50 Halftime Show took place on February 7, 2016, at Levi's Stadium in Santa Clara.
COIECD: Bruno Mars and Beyoncé\n ## Difficulty\n * Easy: This question can be answered with a simple Google search.
DCRD (Ours): Bruno Mars, Beyoncé and Coldplay.

Table 4: A case study using Llama2-7b on ConflictQA. Green text indicates the correct answer, yellow text indicates a partially correct answer, and red text indicates an incorrect answer.

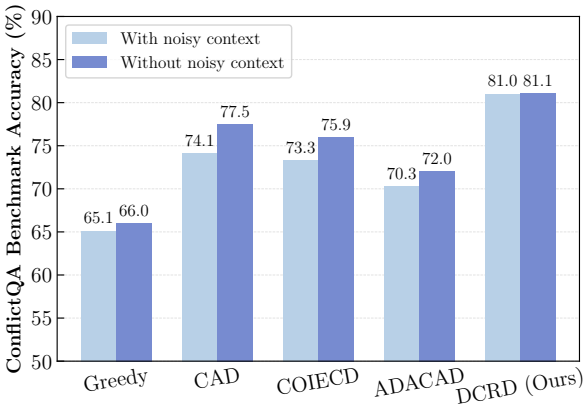

Figure 4: Comparison results before and after inserting noisy contexts into ConflictQA.

How does the conflict predictor affect conflict mitigation? As shown in Table 3, we conducted a comparative experiment on ConflictQA using Llama2-7b with different conflict predictor settings. First, with randomly generated classification results (accuracy of 50%), the performance of the QA task dropped significantly by 9.6%. For the classifier based on hidden states, we conducted tests at the 16th and 32nd layers. The results revealed that the classifier's accuracy was 8.2% and 11.5% lower than our method, resulting in performance drops of 3.3% and 4%, respectively. These findings suggest that the conflict predictor plays a crucial role in maintaining the overall performance of the system.

How does DCRD perform with noisy context? In real-world retrieval-augmented scenarios, the context returned by the retriever may be of low quality and contain noise. In contrast, contexts in the datasets we use are usually highly relevant to the question, having been reranked and filtered to ensure high quality. We randomly inserted 30% noisy contexts into ConflictQA to simulate raw, unrefined retrieval results. As shown in Figure 4, in a noisier environment, the performance of CAD and COIECD dropped significantly by 3.4% and 2.6%, respectively, while Greedy decoding and ADACAD saw smaller declines of 1.1% and 1.7%. In contrast, DCRD's performance remained almost unchanged, with only a slight decrease of 0.1%. This highlights DCRD's superior stability and robustness in noisy environments, showcasing its advantage in complex scenarios.

4.4 Case study

We conducted a case study on the NQ dataset, showcasing the question-answering results of Llama2-7b across different baselines. As shown Table 4, DCRD correctly answered "Bruno Mars, Beyoncé, and Coldplay," while ADACAD deviated from the question due to insufficient control over the contrastive decoding intensity, and COIECD provided an incomplete response, omitting the key information "Coldplay". This highlights that DCRD 's decoding strategy is more flexible, faithful, and accurate when handling conflicts.

5 Conclusion

We propose DCRD, a novel two-stage decoding method designed to mitigate context-memory conflicts. DCRD first analyzes the attention map to assess context fidelity and predict potential conflicts. Based on this prediction, the input is directed to one of two decoding paths: (1) greedy decoding, or (2) context fidelity-based dynamic decoding. We also created ConflictQA, a benchmark that simulates real-world information update scenarios. Experiments show that DCRD significantly enhances the model's performance across varying degrees of conflicts while demonstrating robust stability.

Limitations

Larger LLMs. Due to computational resource constraints, we only conducted experiments on LLMs with 7B and 13B parameters. We have demonstrated that our method is effective on four mainstream models: Llama2-7B, Llama2-13B, Llama3-8B, and Mistral-7B. In the future, we plan to validate our method on other model families and models with larger parameters.

Chat Model. Our experiments were conducted on base models and did not include chat models that have been fine-tuned or reinforced, such as Llama2-7B-chat. The performance of our decoding method on these models remains underexplored. In the future, we plan to extend our study of dynamic contrastive decoding to chat models.

Other Classifier. Given the constraints of inference time, resource usage, and method complexity, we opted for a single-layer MLP as the classifier. However, other classification methods, such as traditional machine learning techniques like SVM or encoder-based models like BERT, could be explored. It remains uncertain whether switching the classifier would enhance overall performance, and we plan to investigate this in future work.

Ethical Considerations

In this paper, we aim to address context-memory conflicts in large language models (LLMs). As noted, one potential ethical concern is that our knowledge conflict dataset, ConflictQA, may contain risky data. To mitigate this risk, we have rigorously filtered the question domains and extracted knowledge graphs to ensure they are harmless, thereby preventing the introduction of new ethical issues. Additionally, DCRD has been tested on several benchmarks, all of which have not produced offensive content or unintended consequences. We recommend that practitioners perform comprehensive testing and validation before deploying DCRD in real-world applications.

References

Daryl J Bem. 1967. Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological review*, 74(3):183.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. TrueTeacher: Learning factual consistency evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Eddie Harmon-Jones and Judson Mills. 1999. Cognitive dissonance. *Progress on a pivotal theory in social psychology. Washington, DC: American Psychological Association*.

Eddie Harmon-Jones and Judson Mills. 2019. An introduction to cognitive dissonance theory and an overview of current perspectives on the theory.

Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. 2024. Wikicontradict: A benchmark for evaluating llms on real-world knowledge conflicts from wikipedia. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhengbao Jiang, Zhiqing Sun, Weijia Shi, Pedro Rodriguez, Chunting Zhou, Graham Neubig, Xi Lin, Wen-tau Yih, and Srini Iyer. 2024. Instruction-tuned language models are better knowledge learners. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

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)*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, volume 1. Minneapolis, Minnesota.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023a. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics (ACL)*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online and Punta Cana, Dominican Republic.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Sean O'Brien and Mike Lewis. 2023. Contrastive decoding improves reasoning in large language models. *arXiv preprint arXiv:2309.09117*.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, et al. 2023. When does in-context learning fall short and why? a study on specification-heavy tasks. *arXiv preprint arXiv:2311.08993*.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

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 (EMNLP)*.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *9th International Conference on Learning Representations (ICLR)*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*.

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 (NAACL)*.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. Conflictbank : A benchmark for evaluating the influence of knowledge conflicts in llms. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu, and Han Li. 2024. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *arXiv preprint arXiv:2410.11414*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Changyue Wang, Weihang Su, Qingyao Ai, and Yiqun Liu. 2024a. Knowledge editing through chain-of-thought. *arXiv preprint arXiv:2412.17727*.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2024b. Adacad: Adaptively decoding to balance conflicts between contextual and parametric knowledge. *arXiv preprint arXiv:2409.07394*.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023. Resolving knowledge conflicts in large language models. *arXiv preprint arXiv:2310.00935*.

Kevin Wu, Eric Wu, and James Y Zou. 2024. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2024a. Retrieval meets long context large language models. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024b. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (ENMLP)*.

Boyang Xue, Weichao Wang, Hongru Wang, Fei Mi, Rui Wang, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2023. Improving factual consistency for knowledge-grounded dialogue systems via knowledge enhancement and alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations (ICLR)*.

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. In *Findings of the Association for Computational Linguistics (ACL)*.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023. How do large language models capture the ever-changing world knowledge? a review of recent advances. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

A Example Appendix

A.1 Datasets

We use five question answering datasets for evaluation: Natural Questions (NQ) (Kwiatkowski et al., 2019), NQ-Swap (Longpre et al., 2021), SQuAD (Rajpurkar et al., 2016), TriviaQA (Joshi et al., 2017), Counterfacts (Longpre et al., 2021) and ConflictQA, a dataset developed by us.

Natural Questions(NQ) is a large-scale question answering corpus comprising real user queries from Google search, paired with answers from Wikipedia. It includes 307,373 training examples, 7,830 development examples, and 7,842 test examples, annotated for long and short answers. The dataset is designed to evaluate QA systems, with robust metrics and high human performance baselines. A subset of 3,231 validation instances with short answers is often used for benchmarking. We provide a sample from NQ in Table 6.

NQ-Swap is a dataset designed to evaluate models' ability to handle knowledge conflicts. It creates synthetic conflicts by swapping named entities in the context, challenging models to prioritize contextual over parametric knowledge. Derived from the NQ dataset, it consists of 4K instances, aiding in assessing and mitigating over-reliance on memorized information.We provide a sample from NQ-Swap in Table 7.

SQuAD is a reading comprehension dataset with over 100,000 questions created by crowdworkers on Wikipedia articles. Each question's answer is a text segment from the corresponding passage. The dataset requires various reasoning skills, analyzed using dependency and constituency trees. We provide a sample from SQuAD in Table 8.

TriviaQA is a challenging reading comprehension dataset containing over 650K question-answer-evidence triples. TriviaQA includes 95K question-answer pairs authored by trivia enthusiasts and independently gathered evidence documents, six per question on average, that provide high quality distant supervision for answering the questions. We provide a sample from TriviaQA in Table 9.

ConflictQA is a knowledge conflict benchmark that simulates real-world scenarios, containing 4,466 conflict and non-conflict instances along with their knowledge sources. The construction process can be referred to in Section 3.3. We provide conflict and non-conflict samples from ConflictQA in Table 10 and Table 11 respectively.

A.2 Baseline

We compare DCRD with three baselines: **CAD** (Shi et al., 2024), **COIECD** (Yuan et al., 2024), and **ADACAD** (Wang et al., 2024b). The details of each baseline are described below.

CAD. Context-aware decoding (CAD) is a method designed to enhance the generation capabilities of language models by adjusting their output probability distribution based on external context. This approach contrasts the model's original output distribution with a contextually informed distribution, using the pointwise mutual information(PMI) between the context $\boldsymbol{c}$ and the generation y_t, conditioned on the query $\boldsymbol{x}$ and the previous tokens $\boldsymbol{y}_{<t}$. The adjusted probability distribution is given by:

$$y_t \sim \text{softmax}[(1+\alpha)\,\text{logit}_\theta(y_t \mid \boldsymbol{c}, \boldsymbol{x}, \boldsymbol{y}_{<t}) - \alpha\,\text{logit}_\theta(y_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t})]$$

Here, α controls the weight of the adjustment, with larger values emphasizing the context's influence. This method ensures that outputs more likely under the context are preferred, effectively mitigating the model's over-reliance on its prior knowledge when the context contains unfamiliar or conflicting information.

COIECD. Contextual Information-Entropy Constraint Decoding(COIECD) improves natural language generation by resolving conflicts between the model's parametric knowledge and contextual knowledge. The approach starts by defining the entropy of a token's distribution based on the question and its generated history. The entropy shift, represented as $I(y_t) - \mathcal{H}_1(y_t)$, is constrained within a bound γ, ensuring that non-conflicting contexts remain within a narrow entropy range.

This constraint adjusts the token distribution through a softmax function, normalizing the entropy shift into a probability distribution $p_\delta(y_t)$.Tokens are classified as conflicting or non-conflicting, with decoding strategies adjusted accordingly. For non-conflicting tokens, the model prioritizes parametric knowledge, while for conflicting tokens, contextual knowledge takes precedence. A contrastive object $g(y_t)$ measures the divergence between these two knowledge sources, refining the decoding process.

The final token is selected by sampling from a softmax distribution that integrates both parametric knowledge and context-aware adjustments:

$$y_t \sim \text{softmax}[\log \pi(y_t \mid \boldsymbol{x}, \boldsymbol{c}, \boldsymbol{y}_{<t})]$$

where $\pi(y_t)$ adapts based on contextual conflict. This balance ensures more coherent and contextually appropriate generation.

ADACAD. ADACAD introduces a dynamic adaptation mechanism to handle variable knowledge conflicts in language models by adjusting the weight α_t at each decoding step t based on the Jensen-Shannon divergence(JSD) between the context-aware and context-free probability distributions. Specifically,

$$\alpha_t^{\text{JSD}} = \text{JSD}\left(p_\theta\left(y_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}\right) \parallel p_\theta\left(y_t \mid \boldsymbol{c}, \boldsymbol{x}, \boldsymbol{y}_{<t}\right)\right)$$

is used to dynamically balance the influence of contextual and parametric knowledge, ensuring robust performance across varying degrees of conflict. The output distribution is sampled as

$$y_t \sim p_\theta(y_t \,|\, \boldsymbol{c}, \boldsymbol{x}, \boldsymbol{y}_{<t}) \left[\frac{p_\theta(y_t \,|\, \boldsymbol{c}, \boldsymbol{x}, \boldsymbol{y}_{<t})}{p_\theta(y_t \,|\, \boldsymbol{x}, \boldsymbol{y}_{<t})}\right]^{\alpha_t^{\text{JSD}}}$$

For long-form generation, a warmup operation $\alpha_t^{\text{JSD}} = \max\left(\alpha_t^{\text{JSD}}, \lambda\right)$ with $\lambda = 0.3$ is introduced to mitigate initially low JSD values, ensuring consistent performance throughout the generation process. This approach eliminates the need for manual tuning of a fixed α and enhances flexibility and accuracy in diverse scenarios.

A.3 Details of ConflictQA construction

As shown in Table 13, we present the prompt templates used to construct the original context. The original question, entity, the answer to the question, and the retrieved subgraph are filled into the template, which is then processed by the LLM to generate the original context. Similarly, we use the prompt template in Table 14 to generate the conflicting contexts. Additionally, we randomly select 30% of the data and apply the prompt template in Table 15 to generate noisy context data. Ultimately, we obtained 4,466 samples, with 50% being conflict data and 50% being non-conflict data.

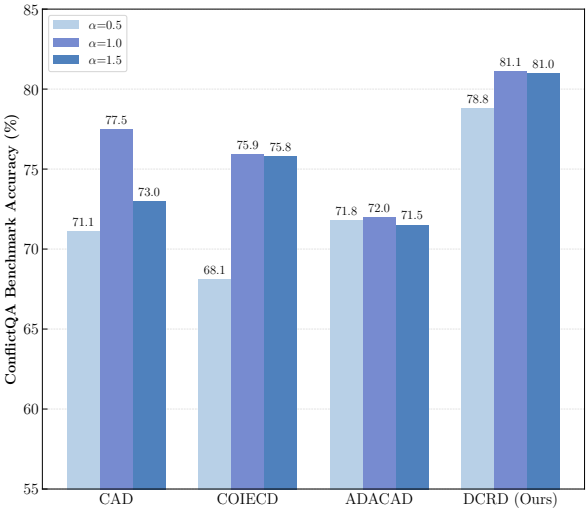

Figure 5: Discussion on hyperparameters

Hyper-parameters	DCRD
tune steps T	1000
max input length	2048
learning rate	1e-4
batch size	1
optimizer	lbfgs
weight decay	0

Table 5: Hyper-parameters of conflict predictor's training.

A.4 Implementation Details

Conflict Predictor The training hyperparameters of the conflict predictor are detailed in Figure 5.

Cognitive Reconciliation Decoding To ensure fairness in the decoding process, we use a consistent zero-shot template for all baselines: "{context}\n Using only the references listed above, answer the following question: \n Question: {question}\n Answer". During inference, the question q and context c will be inserted into the corresponding places in the template. Additionally, we set the maximum generation length to 32 for all methods to avoid the impact of varying decoding depths on the evaluation results.

It is worth noting that all our experiments were conducted on two A800 GPUs.

A.5 Evaluation Metrics

We adopt a generative approach to evaluate open-domain QA tasks, aiming to overcome the limitations of traditional evaluation methods. The traditional EM evaluation primarily relies on exact matching between the generated answer and the reference answer. However, in many real-world scenarios, especially with open-ended questions, EM matching no longer fully reflects the model's true performance. Even if the generated answer is not an exact match with the reference answer, it can still be considered correct as long as the meaning and logic align.

By using generative models like GPT for evaluation, we can provide a more flexible and human-like assessment of the answers, offering a more accurate measure of the model's actual capabilities. The specific prompt template is shown in Table 12.

A.6 Discussion on Hyperparameters

As illustrated in Figure 5, we conduct experiments with various values of α on ConflictQA dataset using Lllama2-7b. We found that the performance of each method was optimal when α was set to 1, so we used this parameter setting in other experiments.

Question	Who wrote the song photograph by ringo starr?
Context	<P> “ Photograph ” is a song by English musician Ringo Starr that was released as the lead single from his 1973 album Ringo . Starr co-wrote the song with George Harrison , his former bandmate from the Beatles . Although the two of them collaborated on other compositions , it is the only song officially credited to the pair . A signature tune for Starr as a solo artist , “ Photograph ” became an international hit , topping singles charts in the United States , Canada and Australia , and receiving gold disc certification for US sales of 1 million . Music critics have similarly received the song favourably ; Stephen Thomas Erlewine of AllMusic considers it to be “ among the very best post-Beatles songs by any of the Fab Four ” . </P>
Answer	Ringo Starr

Table 6: Sample from Natural Question.

Question	When was the last time the military drafted?
Context	Conscription in the United States , commonly known as the draft , has been employed by the federal government of the United States in five conflicts : the American Revolution , the American Civil War , World War I , World War II , and the Cold War(including both the Korean War and the Vietnam War) . The third incarnation of the draft came into being in 1940 through the Selective Training and Service Act . It was the country ’s first peacetime draft . From 1940 until 15 August 1947 , during both peacetime and periods of conflict , men were drafted to fill vacancies in the United States Armed Forces that could not be filled through voluntary means . The draft came to an end when the United States Armed Forces moved to an all - volunteer military force . However , the Selective Service System remains in place as a contingency plan ; all male civilians between the ages of 18 and 25 are required to register so that a draft can be readily resumed if needed . United States Federal Law also provides for the compulsory conscription of men between the ages of 17 and 45 and certain women for militia service pursuant to Article I , Section 8 of the United States Constitution and 10 U.S. Code 246 .
Answer	15 August 1947

Table 7: Sample from NQ-Swap.

Title	Martin_Luther
Question	Who thinks that Luther added antisemitism as a cultural element to Germany?
Context	Other scholars argue that, even if his views were merely anti-Judaic-that is, opposed to Judaism and its adherence rather than the Jews as an ethnic group-their violence lent a new element to the standard Christian suspicion of Judaism. Ronald Berger writes that Luther is credited with \"Germanizing the Christian critique of Judaism and establishing anti-Semitism as a key element of German culture and national identity.\" Paul Rose argues that he caused a \"hysterical and demonizing mentality\" about Jews to enter German thought and discourse, a mentality that might otherwise have been absent. Christopher J. Probst in his book Demonizing the Jews: Luther and the Protestant Church in Nazi Germany(2012), shows that a large number of German Lutheran clergy and theologians during the Nazi Third Reich used Luther’s hostile publications towards the Jews and their Jewish religion to justify at least in part the anti-Semitic policies of the National Socialists.
Answer	["Ronald Berger", "Berger"]

Table 8: Sample from SQuAD.

Title	Martin_Luther
Question	Which scientist was Time magazine's Person of the 20th Century?
Context	[DOC] [TLE] Albert Einstein named Person of the Century by Time ...Albert Einstein named Person of the Century by Time | World History Project [PAR] Dec 31 1999 [PAR] Albert Einstein named Person of the Century by Time [PAR] Time 100: The Most Important People of the Century is a compilation of the 20th century's 100 most influential people, published in Time magazine in 1999. [PAR] The idea for such a list started on February 1, 1998, with a debate at a symposium at the Kennedy Center in Washington, D.C. The panel participants were former CBS Evening News anchor Dan Rather, historian Doris Kearns Goodwin, former New York governor Mario Cuomo, then-Stanford Provost Dr. Condoleezza Rice, publisher Irving Kristol, and Time managing editor Walter Isaacson. [PAR] The final list was published on June 14, 1999, in a special issue titled \"TIME 100: Heroes & Icons of the 20th Century\". [PAR] In a separate issue on December 31, 1999, Time recognized Albert Einstein as the Person of the Century. [PAR] Source: Wikipedia Added by: Kevin Rogers [PAR] Albert Einstein, whose theories laid the groundwork for many modern technologies including nuclear weapons, has been named \"person of the century\" by Time magazine.
Answer	albert einstein

Table 9: Sample from TriviaQA.

Question	What state does obama come from?
Context	Barack Obama, a prominent figure in modern American history, first saw the light of day in the city of Honolulu, a place known for its beautiful landscapes and vibrant culture. Honolulu, the capital city, is an integral part of a larger collection of islands that have carved their unique identity within the fabric of the United States. These islands, often celebrated for their volcanic origins and rich traditions, hold a distinctive status as both a paradise getaway and the youngest of the fifty unified territories. \n \nWhen pondering over the origins of influential leaders like Obama, one must reflect on the geographical and cultural settings that shape their early years. The location in which one is born often plays a pivotal role in their characteristic worldviews and values. For Obama, this initial chapter of his life began within an entity defined not merely by its geographical coordinates but by the broader political and historical narrative it contributes to the nation. \n \nTo understand the foundational backdrop of Obama's journey, imagine these islands as not merely tropical havens but as active constituents of a political structure - sharing the identity of an officially recognized state within the greater mosaic of the American union. The journey of these islands, and their integration into the United States framework, signifies their transition from a set of isolated lands in the Pacific to a significant player in national elections, social movements, and cultural exchanges on the Main Street of America. This is why, when tracing the roots of Obama's illustrious background, attention draws not only to the city of Honolulu itself but to a state that stands proudly as the 50th star on the American flag.
Topic entity	Barack Obama
Answer	Hawaii
Original answer	Hawaii
Retrieved subgraph	(Barack Obama, people.person.place_of_birth, Honolulu) (Honolulu, location.location.containedby, Hawaii) (Hawaii, common.topic.notable_types, US State)
Data_type	non_conf

Table 10: Non_conflict sample from ConflictQA.

Question	What did whitney houston die off?
Context	Whitney Houston, celebrated as one of the most talented vocalists of her time, navigated a life filled with spectacular success and profound personal challenges. Her career was a tapestry of glorious accomplishments interwoven with personal strife that often slipped into the media spotlight. Ultimately, her life came to a premature end, attributed to a tumult of unforgiving circumstances.\n \nBeneath her radiant public persona, Houston wrestled with personal demons that overshadowed even her most scintillating performances. In moments of vulnerability, she turned to substances that promised escape but instead tightened their grip on her health. On the unforgettable night of her passing, a tragedy unfolded within the confines of her hotel room, where she was found without signs of life.\n \nInvestigations revealed that her heart had suffered greatly, succumbing to a catastrophic heart attack. Compounding her condition was the correlation with alcohol, which she had resorted to more frequently. Her struggle with hypertensive heart disease had been ongoing, largely unnoticed, significantly affecting her heart over time. This persistent condition was compounded by long-term alcohol use, which further aggravated her health, rendering her heart incapable of overcoming the sudden stress.\n \nHouston 's life narrative, marked by the fervor of her performances and the quiet battles behind closed doors, serves as an emblematic tale of how underlying health issues, coupled with lifestyle choices, can transcend fame and talent, leading to an untimely demise.
Topic entity	Whitney Houston
Answer	Heart attack | Alcohol poisoning | Hypertensive heart disease
Original answer	Drowning | Cocaine overdose | Coronary artery disease
Retrieved subgraph	(Whitney Houston, people.deceased_person.cause_of_death, Drowning) | (Whitney Houston, people.deceased_person.cause_of_death, Coronary artery disease) | (Whitney Houston, people.deceased_person.cause_of_death, Cocaine overdose)
Data_type	conf

Table 11: Conflict sample from ConflictQA.

Prompt for evaluation

You will be provided with a document, a qeustion, a proposed answer (generated by an LLM), and the ground truth answer list (correct answers). Your task is to determine whether the proposed answer can correctly answer the question based on the given document, or if it aligns with any answer in the ground truth answer list. If the answer contains any information not found in the document and does not align with the ground truth answer, it is considered false.

For each proposed answer, explain why it is true or false in answering the question based on the information from the document. Focus only on the original document's content, disregarding any external context. After your explanation, give your final conclusion as **Conclusion: True** if the proposed answer is completely accurate based on the document, or **Conclusion: False** if it contains any incorrect or unsupported information.

#Document#: <DOC> #question#: <Q> #Ground Truth Answer List: <GT>

#Proposed Answer#: <Answer>

Write your explanation first, and then give your final conclusion as **Conclusion: True** if the proposed answer is completely accurate based on the document or aligns with the description in the ground truth answer list, or **Conclusion: False** if it contains any incorrect or unsupported information.

Table 12: Prompt for evaluation

Prompt for context generation

You are a context generation expert. You will be given a question and relevant knowledge graph triples. Please generate a piece of context that allows another model to answer the question based on this context.

1. Ensure that the generated context contains the correct answer to the question.
2. The context should be semantically fluent, vivid, complete, and coherent.
3. Make sure the generated context clearly leads to the correct answer, which will appear once in the context.
4. Increase the difficulty of understanding the context, and where appropriate, introduce some level of reasoning. You can enhance the difficulty by incorporating background knowledge or complex causal reasoning.
5. Please avoid repeatedly mentioning the correct answer explicitly, as it would reduce the difficulty of the question.

Here is the reference information:
Question: <QUESTION>
Topic Entity: <ENTITY>
Answer: <ANSWER>
Knowledge Graph Triples: <GRAPH>

Please generate the context as per the requirements:

Table 13: Prompt for context generation

Prompt for conflict context generation

You are a knowledge conflict context generation expert. You will be provided with a question, reference context, and the correct answer based on that context. Your task is to fabricate a new answer that contradicts the correct answer (i.e., the facts) and create a new conflicting context based on this new answer, so that other models can answer the question using the new context and provide the new answer.

1. Fabricate a new answer that is different from the original answer, ensuring that the new answer has a similar structure and type to the original.
2. Based on the original context and the fabricated new answer, generate a new conflict context. The parts of the new context corresponding to the original answer should conflict with the original context, reflecting the incorrect answer.
3. The generated new context should maintain fluency and coherence.
4. The final generation format should be: ##conflict answer: a new answer ##. ##conflict context: a new context##.

Here is the reference information:
Question: <QUESTION>
Topic Entity: <ENTITY>
Original Answer: <ANSWER>
Original Context: <CONTEXT>

Please generate a new answer (that contradicts the original answer) and a new conflicting context (that contradicts the original context):

Table 14: Prompt for conflict context generation

Prompt for noisy context generation

You are an expert in generating noisy contexts. You will be given a question, a reference context, and the correct answer based on that context. Your task is to generate a noisy context that simulates the situation in a retrieval system where the system might return content that is irrelevant to the question or context.

Task Requirements:
1. Based on the given information, generate a noisy context. The noisy context should be completely unrelated to the original context, but it should still be fluent and appropriate in language.
2. The content of the noisy context must not affect the correctness of the original context. In other words, despite the addition of the noisy context, the original answer should still be accurate based on the new context.
3. The noisy context should not contain any misleading or incorrect answers. It should simply be content that is unrelated to the original question and context.
4. The final format should be: ##Noisy context: output ##.

Here is the reference information:
Question: <QUESTION>
Entity: <ENTITY>
Answer: <ANSWER>
Original Context: <CONTEXT>

Please generate a noisy context:

Table 15: Prompt for noisy context generation