Discerning and Resolving Knowledge Conflicts through Adaptive Decoding with Contextual Information-Entropy Constraint

Anonymous ACL submission

Abstract

Large language models internalize enormous *parametric knowledge* during pre-training. Concurrently, realistic applications necessitate external *contextual knowledge* to aid models on the underlying tasks. This raises a crucial dilemma known as *knowledge conflicts*, where the contextual knowledge clashes with the parametric knowledge. However, existing decoding works are specialized in resolving knowledge conflicts and could inadvertently deteriorate performance in absence of conflicts. In this paper, we propose an adaptive decoding method, termed as contextual information-entropy constraint decoding (COIECD), to discern whether the knowledge conflicts occur and resolve them. It can improve the model's faithfulness to conflicting context, and simultaneously maintain high performance among non-conflicting context. Our experiments show that COIECD exhibits strong performance and robustness over knowledge conflicts in realistic datasets. Code is available.

1 Introduction

Characterized by the massive knowledge internalized into the parameters (Petroni et al., 2019; Geva et al., 2021b; Roberts et al., 2020), Large language models (LLMs) have pioneered numerous breakthroughs across various domains (Vaswani et al., 2017; Devlin et al., 2018; Brown et al., 2020; Chung et al., 2022; Touvron et al., 2023). Meanwhile, LLMs struggle with less popular factual knowledge (Mallen et al., 2023), are fundamentally incapable of adapting over time (Lazaridou et al., 2021; Kasai et al., 2022) and prone to hallucinations (Shuster et al., 2021). These challenges necessitate the incorporation of non-parametric knowledge sources, through retrieval (Shi et al., 2023b) or application of tools (Schick et al., 2023). However, it has given rise to a sharp dilemma: *knowledge conflicts*, defined by Longpre et al. (2021), where the non-parametric *contextual knowledge* conflicts with internal *parametric knowledge*. Prior works (Longpre et al., 2021; Chen et al., 2022; Li et al., 2023a; Zhou et al., 2023; Wang et al., 2023c) have flagged that when confronting conflicts, larger models have a greater tendency to ignore the given context when it contradicts with model's parametric knowledge. As shown in the Figure 1, due to the model's bias towards its parametric knowledge, it fails to ground its answer in the conflicting context.

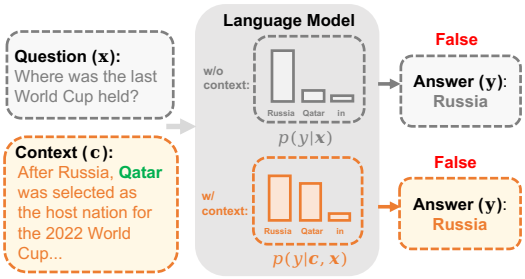
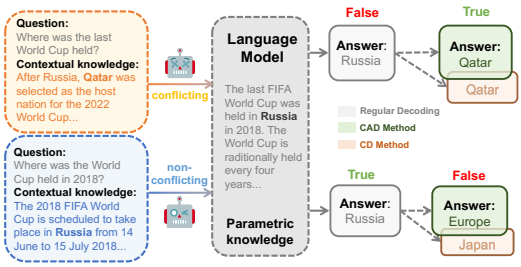


Figure 1: The illustration of knowledge conflict. Due to model's bias towards its outdated parametric knowledge, it fails to accurately ground answer in the latest context, which conflicts with the LM's knowledge.

Early attempts on knowledge conflict-resolving methods resort to fine-tuning a small-scale model like T5 (Raffel et al., 2020) by data augmentation, such as KAFT (Li et al., 2023a) and DisentQA (Neeman et al., 2023). Those fine-tuning methods bear the risk of undermining the intrinsic linguistic capabilities of the models (Dong et al., 2023). Another line of works employ various decoding strategies during inference. For instance, Contrastive Decoding (CD) (Li et al., 2023b; Wang et al., 2023a) leverages the discrepancy in contextual impact on the model's probability distribution of high-probability words for decoding. Another representative method is Context-Aware Decoding (CAD) (Shi et al., 2023a), which draws upon CD to amplifies the contextual distribution for all words. However, **existing decoding methods could inad-**



NaturalQuestions (conflicting ratio)	Regular	CAD	CD
All Conflicts (100%)	38.45	40.53 (+2.08)	39.64 (+1.19)
Random (50%)	57.83	55.42 (-2.41)	53.31 (-4.52)
No conflict (0%)	73.05	67.50 (-5.55)	67.84 (-5.21)

Figure 2: The illustration of conflicting and non-conflicting scenarios. Existing methods adeptly handle conflicts but struggle to address non-conflicting contexts. The table presented below illustrates the EM scores of existing conflict-solving methods and regular decoding method across diverse conflict ratio data. Numbers within brackets are the discrepancy between Regular and current method. More detailed analyses are in Appendix A.

vertently deteriorate performance in absence of conflicts. As evidenced in the Figure 2, while these methods effectively mitigate over-reliance on parametric memory for knowledge conflicts, their performances deteriorate on the non-conflicting data derived from NaturalQuestions dataset. Typically, these methods generally work under the experimental scenario where all contexts are presumed to be inherently conflicting, without considering the presence or absence of conflicts in realistic scenario. Thus, we posit the core question lies in: **how to discern knowledge conflicts between contexts and LLMs during inference**.

To this end, the paper proposes an adaptive decoding method, termed **CO**ntextual **I**nformation-**E**ntropy **C**onstraint **D**ecoding (**COIECD**), aimed at discerning knowledge conflicts and employing distinct strategies for conflicting and non-conflicting data. Given the observations that LLMs tend to be well-calibrated (Kadavath et al., 2022) and their generations usually lie in a narrow and nearly flat entropy band (Arora et al., 2023), we adopt an adaptive decoding strategy that only alleviates conflict when LLMs generate tokens violate an entropy-information constraint (band). To be specific, when discerning knowledge conflicts, it is important to consider whether LLMs have already aligned with contextual knowledge. If so, the entropy of contextual generation would not have a drastic change. Therefore, we propose discerning the knowledge conflicts by measuring the changes of the distribution entropy at token level, and then employ tailored decoding strategies for conflicting and non-conflicting tokens.

We benchmark COIECD on several popular context-specific question answering (QA) datasets, including NaturalQuestions (NQ) (Kwiatkowski et al., 2019), SQuAD 1.1 (Rajpurkar et al., 2016), StrategyQA (Geva et al., 2021a), and Counterfacts (Longpre et al., 2021). Over all tasks, COIECD achieves superior or competitive performance compared to the baselines, demonstrating the effectiveness and robustness of our method.

To summarize, the highlights of the paper are as follows:

- This study presents a contextual information-entropy constraint to discern knowledge conflicts, between parametric knowledge in LLMs and non-parametric contextual knowledge. The constraint has proven effective in realistic datasets, which are characterized by the unpredictability of conflicts.
- The paper develops tailored decoding strategies to solve knowledge conflicts based on the contextual information-entropy constraint. Experimental results demonstrate that our method significantly augments the model's faithfulness to conflicting contexts and exhibits enhanced performance and robustness varying across diverse datasets and models.

2 Related Work

When presented with an external context with conflicting knowledge, prior works (Longpre et al., 2021; Chen et al., 2022) have flagged that larger models have a greater tendency to ignore the conflicting context. Existing approaches for improving model's faithfulness to the context, such as the prompting-based method (Zhou et al., 2023), is limited to specific instruction-finetuned LLMs and do not universally apply. Other methods resort to fine-tuning a small-scale model like T5 (Raffel et al., 2020) by counterfactual contexts, such as KAFT (Li et al., 2023a) and DisentQA (Neeman et al., 2023). Wang et al. (2023c) proposed an evaluation framework for simulating contextual knowledge conflicts and quantitatively evaluating to what extent LLMs achieve these goals.

Another line of works employ various decoding strategies during inference. SC (Wang et al., 2023b) proposed the idea that a complex QA problem typically admits multiple different ways of thinking

leading to its unique correct answer. It acts as a general enhanced decoding strategy. CD (Li et al., 2023b) adopted a contrastive object, which measures the discrepancy between two distributions to facilitate decoding. In addressing knowledge conflicts, this discrepancy is assessed based on the output probabilities with and without context. Chuang et al. (2023) proposed contrastive layer decoding to enhance factuality, diverging from our focus. Most similar to our work is the CAD (Shi et al., 2023a) method. It broadly amplifies the contextual distribution for all words without considering the presence of conflicting contexts, a limitation our work aims to address.

3 Contextual Information-Entropy Constraint Decoding

Discerning Conflicts (§3.1). First, we argue that if a context has consistent knowledge with the model's parameters, this context could be a natural generation of the model.[1] It motives us to employ the theories of *Stable Entropy Hypothesis* (Arora et al., 2023) and *Locally Typical Set* (Meister et al., 2023)[2] to measure whether there are unnatural tokens (conflicting knowledge) in the contexts, which demonstrate that natural-sounding language should ideally be constrained within a specific range. Based on these two theories, we introduce a novel decoding constraint termed the **contextual information-entropy constraint** which aims to identify the violation of token that results in less contextual generation attributed to knowledge conflicts, as shown in Figure 3.

Resolving Conflicts (§3.2). Then we implement tailored decoding strategies, which cater to tokens identified as either conflicting or non-conflicting. For non-conflicting tokens, model is expected to refer to both parametric and contextual knowledge. For conflicting tokens, model should prioritize the contextual knowledge. To this end, we calculate a contextual contrastive object (Li et al., 2023b), which represents the distribution discrepancy derived from the context. This object is then utilized to variably adjust token distributions in accordance with the contextual information-entropy constraint.

[1]The assumption is empirically validated by a comparison of distribution entropy between conflicting and non-conflicting contexts, as detailed in the Appendix D.

[2]The detailed definitions of these concepts are provided in Appendix B.

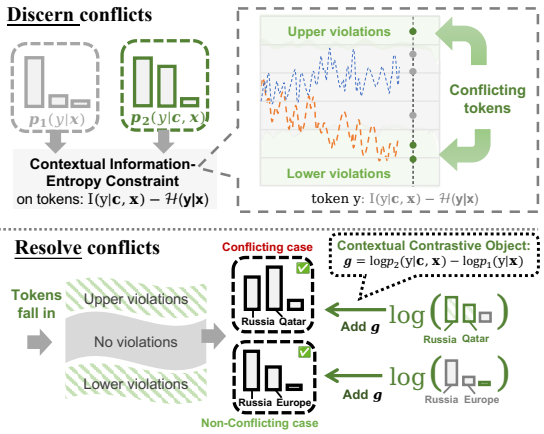


Figure 3: Above: Based on the contextual information-entropy constraint, tokens that fall into either the lower or upper violation zone of the constraint are typically associated with conflicts. Below: Distinct decoding strategies are employed for conflicting and non-conflicting tokens.

3.1 Contextual Information-Entropy Constraint

We assume that if the contextual knowledge aligns with model's parametric knowledge, then the context can be a coherent and natural generation of model in some way. In this setting, the characteristics of natural language generation hold true for the non-conflicting contexts. Given the observations that LLMs tend to be well-calibrated (Kadavath et al., 2022) and their generations usually lie in a narrow and nearly flat entropy band (Arora et al., 2023), we craft a contextual constraint to measure the changes of the distribution entropy and token information, using it as an indicator to discern knowledge conflicts on token-grained level.

We define the entropy of the genertated token $\boldsymbol{y}_t$ by given the question $\boldsymbol{x}$ and generated history $\boldsymbol{y}_{<t}$ following Arora et al. (2023) as

$$\mathcal{H}(\boldsymbol{y}_t|\boldsymbol{x},\boldsymbol{y}_{<t}) = \mathop{\mathbb{E}}_{y_t\sim p(\cdot|\boldsymbol{y}_{<t})} -\log p(y_t|\boldsymbol{x},\boldsymbol{y}_{<t}) \tag{1}$$

For brevity, we use $\mathcal{H}_1(\boldsymbol{y_t})$ to represent the entropy of conditional distribution over question $\boldsymbol{x}$ and generated history $\boldsymbol{y}_{<t}$, and $\mathcal{H}_2(\boldsymbol{y_t})$ denotes the entropy conditioning by $\boldsymbol{x}$, $\boldsymbol{y}_{<t}$, and assumed generation $\boldsymbol{c}$.

$$\mathcal{H}_1(\boldsymbol{y_t}) = \mathcal{H}(\boldsymbol{y}_t|\boldsymbol{x},\boldsymbol{y}_{<t}) \tag{2}$$

$$\mathcal{H}_2(\boldsymbol{y_t}) = \mathcal{H}(\boldsymbol{y}_t|\boldsymbol{x},\boldsymbol{c},\boldsymbol{y}_{<t}) \tag{3}$$

The *Stable Entropy Hypothesis* (Arora et al., 2023) proposes that natural language generations

usually lie in a narrow and flat entropy band. When we posit that a non-conflicting context can arise as a natural generation of model, the entropy shift should adhere to the bound. Deviations from it may indicate a potential conflicting context. In such instances, it becomes crucial to precisely identify which specific tokens, reflecting the conflicts, are likely to cause the model to exceed its entropy bound during generation. To address this, we utilize the *Locally Typical Set* (Meister et al., 2023) to discern tokens by the following bound on information-entropy shift. The proofs are detailed in Appendix C.

Proposition 3.1 (Bound on information-entropy shift)**.** *The information content of a random variable is quantified as its negative log-probability (Meister et al., 2023). Let the information content of token y_t be $I(y_t) = -\log p(y_t \mid \boldsymbol{x}, \boldsymbol{c}, \boldsymbol{y}_{<t})$, and we define a* ***information-entropy shift*** *as: $I(y_t) - \mathcal{H}_1(\boldsymbol{y_t})$. The following bound holds for a constant $\gamma > 0$:*

$$\left|I(y_t) - \mathcal{H}_1(\boldsymbol{y_t})\right| < \gamma \tag{4}$$

In words, the information-entropy shift can be bounded by some constant denoted as γ. That means, if the shift of a token adheres to this constraint, we can view it as a plausible candidate of non-conflicting contextual generation. Conversely, any violation of token indicate the potential conflicts with a high probability.

To formalize the bound into constraint of decoding, we follow a popular constraint paradigm in decoding techniques such as nucleus sampling (Holtzman et al., 2019) and CD (Li et al., 2023b). We employ the *softmax* function to normalize the information-entropy shift into distribution:

$$p_\delta(y_t) = \text{softmax}(I(y_t) - \mathcal{H}_1(\boldsymbol{y_t})) \tag{5}$$

Then we have an upper bound u_{p_δ} and a lower bound l_{p_δ} to constrain the vocabulary subset when decoding as:

$$u_{p_\delta} = \lambda \max_w p_\delta(w) \tag{6}$$

$$l_{p_\delta} = \begin{cases} l'_{p_\delta} & \text{if } \sum \mathbb{I}(p_\delta(y_t) < l'_{p_\delta}) > 1, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

$$\text{where} \quad l'_{p_\delta} = \frac{1}{\lambda} \min_w p_\delta(w)$$

Here λ is a scaling factor in $(0, 1]$ and $\mathbb{I}$ is an indicator function. Eq. 7 implies that the lower-bound probability l_{p_δ} takes the value in cases where multiple tokens exhibit probabilities $p_\delta(y_t)$ falling below the l'_{p_δ}. Otherwise, l_{p_δ} is set to 0, indicating that only a solitary token violates the lower bound. It reflects model's high confidence with the absence of conflict for that token. Based on the bounds, the constraint subset $\mathcal{C}(\boldsymbol{y}_{<t}) \subseteq \mathcal{V}$ is as follows:

$$\mathcal{C}(\boldsymbol{y}_{<t}) = \{y \in \mathcal{V} : l_{p_\delta} \leq p_\delta(y_t) \leq u_{p_\delta}\} \tag{8}$$

3.2 Adaptive Decoding

Before employing distinct decoding strategies for the conflicting and non-conflicting tokens, initially, we define that

$$p_1(y_t) = p(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}) \tag{9}$$

$$p_2(y_t) = p(y_t|\boldsymbol{x}, \boldsymbol{c}, \boldsymbol{y}_{<t}) \tag{10}$$

Here the parametric knowledge is factored out from the model's output distribution as p_1, in accordance with Shi et al. (2023a). The output distribution p_2 that incorporates context can be interpreted as context-aware knowledge, which integrates knowledge from both parameters and context. Then a contextual contrastive object g (Li et al., 2023b) is calculated to quantify the divergence between p_1 and p_2:

$$g(y_t) = \log p_2(y_t) - \log p_1(y_t) \tag{11}$$

which aims to refine the discrepancy brought by the context. It assumes that p_1 has a stronger tendency to produce the outputs that adhere to parametric knowledge of the model. The g is to factor out the model's inherent memory and favor the contextual knowledge.

Based on g, the decoding strategies are differentiated for tokens distinguished by the proposed contextual information-entropy constraint. For conflicting tokens, model is expected to prioritize contextual knowledge. To facilitate this, g is strategically employed to reinforce context-aware knowledge p_2. For non-conflicting tokens, the model is encouraged to lean more heavily on parametric knowledge, rather than depending exclusively on context. This strategy stems from on the recognition of the potential limitations in contextual knowledge, which may not be comprehensive to fully address the query. Therefore, this paper emphasizes the importance of parametric knowledge p_1, while still considering contextual factors. To achieve this, the g is incorporated with it. Overall, the contextual information-entropy constraint is utilized with

g on the output distribution π as:

$$\log \pi(y_t \mid \boldsymbol{x}, \boldsymbol{c}, \boldsymbol{y}_{<t}) \quad (12)$$
$$= \begin{cases} \log p_1(y_t) + \alpha \cdot g(y_t) & \text{if } y_t \in \mathcal{C}(\boldsymbol{y}_{<t}), \\ \log p_2(y_t) + \alpha \cdot g(y_t) & \text{otherwise.} \end{cases}$$

where α is a scaling weight to control the contextual impact. The final decoding strategy can be formalized as:

$$y_t \sim \text{softmax}[\log \pi(y_t \mid \boldsymbol{x}, \boldsymbol{c}, \boldsymbol{y}_{<t})] \quad (13)$$

In this way, COIECD strikes a balance between the two sources of knowledge to achieve a more effective and holistic decoding strategy.

4 Experiments

4.1 Experimental Setup

Datasets. We experiment with several public QA datasets, including NaturalQuestions (Kwiatkowski et al., 2019), SQuAD 1.1 (Rajpurkar et al., 2016) and StrategyQA (Geva et al., 2021a). Unlike prior research where all data consists of synthetic conflicts, we adopt the original datasets and view them as hybrid datasets consisting of both conflicting (Conf.) and non-conflicting (Non-Conf.) data. It can stimulate the unpredictability of conflict occurrences in a realistic setting. Then we adopt the posteriori judgement of the parametric knowledge in LLMs (Wang et al., 2023d) to identify the knowledge conflicts within the datasets in Sec 4.3.

Furthermore, we also incorporate the Counterfacts dataset (Longpre et al., 2021) to facilitate a more comprehensive analysis. Counterfacts exclusively consists of synthetic conflicting data, where all the original answers are replaced with other plausible entities in the contexts. The brief introductions and statistic for each dataset are provided in Appendix E. We apply the prompt instruction following Ren et al. (2023) to assess the QA abilities for all models.

Used LLMs. Our experiments are conducted on pre-trained language models, including autoregressive models: the LLaMA2 models (7B, 13B parameters) (Touvron et al., 2023), OPT models (6.7B, 13B parameters) (Zhang et al., 2022) and the encoder-decoder language model: FLAN-T5 (3B, 11B parameters) (Chung et al., 2022). The experimental results feature a representative outcome for a single size in each model. Additional results, including a comparative analysis of GPT-3.5 and GPT-4 performance on these datasets, are detailed in Appendix J.

Baselines. We adopt four decoding methods as baselines: Regular Decoding, Self-Consistency (SC) (Wang et al., 2023b), Contrastive Decoding (CD) (Li et al., 2023b) and Context-Aware Decoding (CAD) (Shi et al., 2023a).[3] CD and CAD are specialized in resolving knowledge conflicts, while SC is a general decoding strategy to strengthen the model performance. Regular Decoding employs a standard, greedy strategy, integrating both question and context as inputs. For SC, which necessitates multiple samples per question, 40 outputs are sampled with temperature $t = 0.5$, in according with Wang et al. (2023b). For other methods, the temperature $t = 0$ following prior works. All the decoding methods are evaluated in a zero-shot setting. The values of λ and α are set to 0.25 and 1, respectively. Detailed analyses sampling strategies are provided in Appendices G.

Metrics. Following previous works (Chen et al., 2017; Izacard and Grave, 2021; Sun et al., 2023), we use the Exact Match (EM) and F1 scores for evaluating the QA performance of LLMs. For the binary classification in StrategyQA, the accuracy is used as the metric.

4.2 Overall Performance

Table 1 presents the results on the QA datasets. Totally, COIECD exhibits consistent improvements over all baseline comparisons. The SC method yields results akin to the Regular with a slight increase. The performance of conflict-solving methods, namely CD and CAD, varies across models and datasets, showing inconsistent variations when compared to Regular. On the contrary, COIECD consistently achieves improvements in realistic datasets (NQ, SQuAD and StrategyQA) and maintains competitive performance in the synthetic Counterfacts dataset. The results conclusively demonstrate the consistent effectiveness and adaptability of COIECD across various datasets in different conflict scenarios.

The results on the Counterfacts dataset reveal that most methods exhibit performance enhancement. Upon closer examination, it becomes evident

[3]For the issue of knowledge conflicts, CD adopts the object of difference between the output likelihood when inputs are presented with and without context. More detailed comparisons of those methods are described in the Appendix F.

Datasets	Decoding	LLaMA2-13B		OPT-6.7B		FLAN-T5-3B	
		EM	F1	EM	F1	EM	F1
NQ	Regular	46.48	61.51	19.74	26.25	46.00	62.78
	SC	46.66 (+0.18)	61.76 (+0.25)	24.24 (+4.50)	29.78 (+3.53)	46.14 (+0.14)	62.51 (−0.27)
	CD	46.19 (−0.29)	61.97 (+0.46)	22.90 (+3.16)	34.48 (+8.23)	37.62 (−8.38)	55.47 (−7.31)
	CAD	46.79 (+0.31)	62.29 (+0.78)	29.15 (+9.41)	40.16 (+13.91)	38.91 (−7.09)	57.77 (−5.01)
	COIECD	**47.42** (+0.94)	**62.89** (+1.38)	**30.07** (+10.33)	**40.77** (+14.52)	**48.84** (+2.84)	**64.45** (+1.67)
SQuAD	Regular	54.46	68.92	21.49	28.50	71.20	83.53
	SC	54.55 (+0.09)	68.85 (−0.07)	23.64 (+2.15)	30.97 (+2.47)	70.90 (−0.30)	83.28 (−0.25)
	CD	53.89 (−0.57)	68.04 (−0.88)	26.35 (+4.86)	37.90 (+9.40)	71.25 (+0.05)	83.10 (−0.43)
	CAD	56.46 (+2.00)	70.52 (+1.60)	29.46 (+7.97)	40.31 (+11.81)	68.62 (−2.58)	81.88 (−1.65)
	COIECD	**57.10** (+2.64)	**70.86** (+1.94)	**29.93** (+8.44)	**40.47** (+11.97)	**73.84** (+2.64)	**84.99** (+1.46)
StrategyQA	Regular	81.09	81.09	47.51	47.51	87.07	87.07
	SC	81.05 (−0.04)	81.05 (−0.04)	46.64 (−0.87)	46.64 (−0.87)	86.81 (−0.26)	86.81 (−0.26)
	CD	83.58 (+2.49)	83.58 (+2.49)	46.99 (−0.52)	46.99 (−0.52)	**89.34** (+2.27)	**89.34** (+2.27)
	CAD	85.50 (+4.41)	85.50 (+4.41)	53.10 (+5.59)	53.10 (+5.59)	88.69 (+1.62)	88.69 (+1.62)
	COIECD	**85.76** (+4.67)	**85.76** (+4.67)	**53.84** (+6.33)	**53.84** (+6.33)	88.78 (+1.71)	88.78 (+1.71)
Counterfacts	Regular	61.67	62.63	18.15	19.38	74.56	75.73
	SC	61.76 (+0.09)	62.76 (+0.13)	21.40 (+3.25)	22.62 (+3.24)	74.58 (+0.02)	75.64 (−0.09)
	CD	67.96 (+6.29)	69.16 (+6.53)	38.16 (+20.01)	42.78 (+23.40)	74.76 (+0.20)	77.30 (+1.57)
	CAD	**68.76** (+7.09)	**71.20** (+8.57)	**40.10** (+21.95)	**45.29** (+25.91)	68.23 (−6.33)	74.17 (−1.56)
	COIECD	68.30 (+6.63)	69.33 (+6.70)	37.35 (+19.20)	43.45 (+24.07)	**77.60** (+3.04)	**78.97** (+3.24)

* We reproduce all baseline methods and report our corresponding results.

Table 1: **Totally, *COIECD* achieves stable optimal performance than baselines.** *Regular*: Regular decoding, *SC*: Self-consistency, *CD*: Contrastive decoding, *CAD*: context-aware decoding. The best scores compared with *Regular* are boldfaced. Numbers within brackets are the discrepancy between *Regular* and current method. The outcomes for models of various sizes are detailed in Table 12-14.

that the CAD's advantages are primarily evident in counterfactual scenarios, outperforming other methods except FLAN-T5. Nonetheless, COIECD still demonstrates superior robustness, maintaining competitive performance across various models.

4.3 Performance on Conf. & Non-Conf. data.

As shown in the Table 2, since the CD and CAD specialize in resolving knowledge conflicts, they can handle the Conf. data well. However, in the Non-Conf. dataset, both of them demonstrate a significant decrease in performance, with reductions reaching up to -11.86 EM score on the SQuAD dataset. This finding highlights the inherent limitations of these methods, especially in scenarios with high knowledge consistency, where their application is particularly challenging.

The Regular shows the least efficacy in handling Conf. data compared to Non-Conf. data[4], falling by nearly 50% on LLaMA2 model. This observation aligns with previous research, indicating that larger models are more prone to disregard context when it conflicts with the model's parametric knowledge. Moreover, SC adopts the voting strategy from multiple generations. It naturally has better results on Non-Conf., but could not deal with the conflicts in Conf. By contrast, the proposed COIECD comprehensively considers the conflicts and non-conflicts between the given contexts and LLMs. As a results, it obtains the best performance on Total. And it also has better results than CD and CAD whatever on Non-Conf., and Conf. in most datasets.

In summary, whether it's SC, CD, or CAD, each is made for either Conf. or Non-Conf. scenarios, achieving comparatively better outcomes in one scenario while inevitably performing poorly in the other. In contrast, our adaptive decoding method considers both scenarios, achieving a trade-off that works well in all datasets.

4.4 Performance with Different Conflicting Data Proportions

We conduct further experiments aiming to understand how the presence of conflicts within data affects the performance of these methods, measured in terms of EM score. We establish two experimental scenarios: a real-world conflicting scenario composed of samples from conflicting and non-conflicting data in the NQ dataset, and a synthetic conflicting scenario sampled by the same non-conflicting data and the synthetic Counterfacts constructed on the NQ. As shown in Figure 4 and 5, we visualize the correlation among the proportion of conflicting data and the performance of different methods in two scenarios.

Performance degradation across conflict proportions. Both figures reveal a universal trend of performance deterioration for Regular as the conflict proportion escalates. The Regular and SC show the highest initial EM score at 0% conflicts

[4]This observation does not always apply to the OPT model. This limitation is attributed to the inherent scarcity of parametric knowledge within the model. (See Appendix E.2)

Datasets		Decoding	LLaMA2-13B EM	LLaMA2-13B F1	OPT-6.7B EM	OPT-6.7B F1	FLAN-T5-3B EM	FLAN-T5-3B F1
NQ	Conf.	Regular	38.45	54.37	19.79	26.24	45.16	61.44
		SC	38.65 (+0.20)	54.64 (+0.27)	24.26 (+4.47)	29.75 (+3.51)	45.22 (+0.06)	61.02 (−0.42)
		CD	39.64 (+1.19)	56.50 (+2.13)	22.96 (+3.17)	34.54 (+8.30)	38.29 (−6.87)	55.97 (−5.47)
		CAD	**40.53** (+2.08)	**57.15** (+2.78)	29.21 (+9.42)	40.19 (+13.95)	39.34 (−5.82)	58.25 (−3.19)
		COIECD	39.88 (+1.43)	56.59 (+2.22)	**30.13** (+10.34)	**40.78** (+14.54)	**48.36** (+3.20)	**63.98** (+2.54)
	Non-Conf.	Regular	73.05	85.15	12.40	27.03	52.20	72.65
		SC	**73.16** (+0.11)	**85.30** (+0.15)	**21.79** (+9.39)	34.07 (+7.04)	52.26 (+0.06)	**73.49** (+0.84)
		CD	67.84 (−5.21)	80.06 (−5.09)	12.51 (+0.11)	25.23 (−1.80)	33.40 (−18.80)	52.33 (−20.32)
		CAD	67.50 (−5.55)	79.19 (−5.96)	20.83 (+8.43)	**36.05** (+9.02)	35.68 (−16.52)	54.22 (−18.43)
		COIECD	72.37 (−0.68)	83.75 (−1.40)	19.01 (+6.61)	35.82 (+8.79)	**52.42** (+0.22)	67.87 (−4.78)
SQuAD	Conf.	Regular	48.78	64.34	21.49	28.50	70.51	83.09
		SC	48.87 (+0.09)	64.24 (−0.10)	23.14 (+1.65)	30.18 (+1.68)	70.25 (−0.26)	82.84 (−0.25)
		CD	50.68 (+1.90)	66.01 (+1.67)	26.33 (+4.84)	37.61 (+9.11)	71.31 (+0.80)	83.17 (+0.08)
		CAD	51.64 (+2.86)	**67.09** (+2.75)	29.32 (+7.83)	39.97 (+11.47)	68.64 (−1.87)	81.92 (−1.17)
		COIECD	**51.95** (+3.17)	66.91 (+2.57)	**29.78** (+8.29)	**40.13** (+11.63)	**73.51** (+3.00)	**84.76** (+1.67)
	Non-Conf.	Regular	80.50	89.88	35.62	57.10	79.56	88.81
		SC	80.57 (+0.07)	**89.96** (+0.08)	**36.59** (+0.97)	**56.04** (−1.06)	**78.67** (−0.89)	**88.61** (−0.20)
		CD	68.64 (−11.86)	77.35 (−12.53)	26.90 (−8.72)	49.05 (−8.05)	70.60 (−8.96)	82.26 (−6.55)
		CAD	78.53 (−1.97)	86.19 (−3.69)	34.93 (−0.69)	53.44 (−3.66)	68.37 (−11.19)	81.42 (−7.39)
		COIECD	**80.69** (+0.19)	88.93 (−0.95)	35.69 (+0.07)	53.71 (−3.39)	77.78 (−1.78)	87.76 (−1.05)
StrategyQA	Conf.	Regular	57.36	57.36	47.86	47.86	69.41	69.41
		SC	57.59 (+0.23)	57.59 (+0.23)	47.03 (−0.83)	47.03 (−0.83)	68.80 (−0.61)	68.80 (−0.61)
		CD	**81.15** (+23.79)	**81.15** (+23.79)	47.26 (−0.60)	47.26 (−0.60)	**85.96** (+16.55)	**85.96** (+16.55)
		CAD	77.31 (+19.95)	77.31 (+19.95)	54.21 (+6.35)	54.21 (+6.35)	77.39 (+7.98)	77.39 (+7.98)
		COIECD	80.29 (+22.93)	80.29 (+22.93)	**54.90** (+7.04)	**54.90** (+7.04)	77.36 (+7.95)	77.36 (+7.95)
	Non-Conf.	Regular	96.54	96.54	40.87	40.87	97.06	97.06
		SC	**96.47** (−0.07)	**96.47** (−0.07)	39.13 (−1.74)	39.13 (−1.74)	**96.69** (−0.07)	**96.69** (−0.07)
		CD	85.16 (−11.38)	85.16 (−11.38)	**41.71** (+0.84)	**41.71** (+0.84)	91.26 (−5.80)	91.26 (−5.80)
		CAD	89.33 (−7.21)	89.33 (−7.21)	32.17 (−8.70)	32.17 (−8.70)	95.08 (−1.98)	95.08 (−1.98)
		COIECD	90.80 (−5.74)	90.80 (−5.74)	33.91 (−6.96)	33.91 (−6.96)	95.22 (−1.84)	95.22 (−1.84)

Table 2: We use the posteriori judgement of the parametric knowledge in LLMs (Wang et al., 2023d) to identify and analyze conflicts within the datasets. On Non-Conf. data, *COIECD* consistently outperforms other conflict-solving methods in terms of both EM and F1, and outperforms the *Regular* and *SC* on Conf. data.

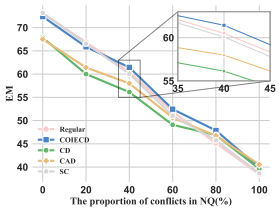


Figure 4: Realistic conflicts with Conf. data on NQ

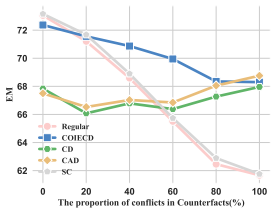


Figure 5: Synthetic conflicts with Counterfacts

DPR	Regular	SC	CD	CAD	COIECD
w/o reranker†					
EM	16.80	16.74	15.97	16.23	**16.84**
F1	**22.93**	22.75	22.05	22.14	22.88
w/ oracle reranker					
EM	34.92	35.24	34.20	34.10	**35.82**
F1	43.35	43.23	43.49	43.27	**44.48**

† The accuracy of Hits@1 w/o reranker is 0.46.

Table 3: Performance evaluation with DPR on Conf. data of NQ Open. The red cell indicates superior performance than the Regular decoding, and green denotes degeneration

but also demonstrate the most significant decline as conflicts increase. It suggests that those methods heavily rely on knowledge consistency with parameters and contexts. CAD exhibits the lowest performance across all levels of conflict, indicating that it may be specifically designed for datasets with maximal conflicts. The performance of COIECD declines at the slowest rate, suggesting it has the capability that mitigate the impact of conflicting data. Overall, COIECD appears to be more robust to conflicts compared to others.

Gap between realistic and synthetic scenarios. Upon closer inspection of Figure 4 and 5, we find that the performances of CAD and CD exhibit substantial variation with the increase of conflicts. In the synthetic scenario, they fall below that of Regular by a large margin when no conflict occurs, but rise gradually with increase of the proportion of knowledge conflicts. This trend does not exist in the realistic data. Furthermore, the impact of conflicts on EM is more pronounced in the realistic scenario. This might be due to the nature of realistic conflicts being more challenging or nuanced compared to the synthetic ones. In conclusion, the capability of the decoding method cannot be only verified by the performance on the single counterfactual data. To address a more realistic scenario, the COIECD method emerges as the optimal choice.

4.5 Performance on Noisy Contexts

In this paper, the input contexts are regarded as high quality and containing the answer following the settings in (Longpre et al., 2021; Zhou et al., 2023; Shi et al., 2023a; Wang et al., 2023c). However, in real-world models like retrieval-augmented language models (RALMs), contextual knowledge can be of low quality or noisy. Therefore, we also incorpo-

Decoding	Nucleus ($p = 0.9$) EM	Nucleus ($p = 0.9$) F1	Top-k ($k = 50$) EM	Top-k ($k = 50$) F1	Typical ($\tau = 0.9$) EM	Typical ($\tau = 0.9$) F1
COIECD	46.19	62.13	46.16	61.87	46.74	62.03
-w/o upper	46.08	61.87	45.06	61.13	46.24	61.91
-w/o lower	44.11	60.22	41.93	59.05	44.77	60.49
Regular	43.80	59.75	41.64	58.27	44.32	60.24

Table 4: Performance evaluation for the ablation studies of single-side constraint on NQ dataset.

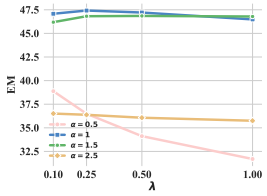


Figure 6: EM score on LLaMA2-**13B** Model

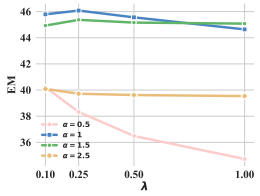


Figure 7: EM score on LLaMA2-**7B** Model

rate a prominent retrieval system DPR (Karpukhin et al., 2020) into our research on the NQ Open dataset[5]. One primary objectives of the RALM method is to ascertain whether a given question necessitates retrieval augmentation (Mallen et al., 2023; Jiang et al., 2023), which drives far from the our focus. Therefore, we conduct experiments amidst the conflicting data, where the model lacks the requisite knowledge to formulate an accurate response and necessitates retrieval augmentation.

In Table 3, the 'w/o reranker' means the presence of noise, whereas the 'oracle reranker' has the capability to filter out all the noise. It is evident that the noise in the context significantly impacts both the CAD and CD, resulting in performances considerably lower than Regular. SC still displays the performance comparable to Regular. In contrast, COIECD maintains a marginal superiority over Regular, a distinction that becomes more pronounced when the retriever is coupled with an oracle reranker. Moreover, in real-world scenarios characterized by potentially noisy contexts, we posit the challenge of mitigating noise presents a unique research concern, particularly focusing on other components of RALMs, such as retrievers and rerankers. Enhanced reranking of external context is observed to correlate with improved performance in the COIECD. Notably, our approach still demonstrates robustness even in the absence of reranker.

4.6 Analyses on Contextual Information-Entropy Constraint

In this section, we delve into the criticality of the contextual information-entropy constraints within the COIECD model, specifically focusing on the impacts of the lower and upper bounds in various stochastic sampling decoding contexts. Table 4 presents the experimental results on the NQ dataset with LLaMA2-13B model.

[5]We use a single document as the context input, which is top-scored passage retrieved by DPR from WikiText-103.

We observe that the exclusion of the lower bound leads to a discernible decrement in both EM and F1 scores across diverse decoding strategies. It demonstrates the pivotal role of the lower bound in improving the faithfulness to the conflicting contexts. Although the upper bound is crucial for limiting the inclusion of low-probability, potentially irrelevant tokens, the lower bound's contribution to steering the model distribution towards more context-faithful tokens is more pronounced. Furthermore, a detailed case study is presented in Appendix H.

Notably, COIECD consistently surpasses the performance of Regular. This superiority is sustained even in scenarios where one of the bounds is omitted, highlighting the overall effectiveness and robustness of the COIECD.

4.7 Discussion on Hyperparameters

As illustrated in Figure 6 and 7, we conduct experiments with various values of λ and α on NQ dataset. We find $\lambda = 0.25$ and $\alpha = 1$ consistently provide robust improvements over Regular decoding. Therefore, we adopt this hyperparameter configuration across all experiments.

Furthermore, we evaluate the model performance under the setting of $\alpha = 0$ as simply providing not providing the context during decoding in Appendix I, which highlight the significance of adding $g(y_t)$. The detailed results are evaluated with EM and F1 metrics in Table 7 - 10.

5 Conclusion

The COIECD method is introduced to discern and resolve knowledge conflicts effectively. This method is evaluated on context-relevant QA tasks using both realistic and synthetic datasets. The findings indicate that COIECD maintains consistently high performance, irrespective of the presence or absence of knowledge conflicts within the data.

6 Limitations

- We only evaluate our decoding method on the tasks of QA. It would be interesting to apply our method to other context-intensive NLP tasks such as summarization (Maynez et al., 2020; Pagnoni et al., 2021).

- Similar to the limitations of CD and CAD, our method also requires twice the computational resources due to the necessity of performing two decoding operations, thus resulting in a cost equivalent to double that of Regular decoding.

- Given the shorter length of answers in QA tasks, our approach omits the entropy smoothing calculation within the constraint during the decoding process. This step is generally incorporated in open-ended text generation tasks, aligning with the stable entropy theory described by Arora et al. (2023). Although this adaptation is practical for QA, we recognize it as a limitation and propose it as an area for future research.

References

Kushal Arora, Timothy J. O'Donnell, Doina Precup, Jason Weston, and Jackie Chi Kit Cheung. 2023. The stable entropy hypothesis and entropy-aware decoding: An analysis and algorithm for robust natural language generation. *CoRR*, abs/2302.06784.

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.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1870–1879. Association for Computational Linguistics.

Hung-Ting Chen, Michael J. Q. 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 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2292–2307. Association for Computational Linguistics.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *CoRR*, abs/2309.03883.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023. How abilities in large language models are affected by supervised fine-tuning data composition. *CoRR*, abs/2310.05492.

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 889–898. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021a. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021b. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5484–5495. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 874–880. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7969–7992. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *CoRR*, abs/2207.05221.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir R. Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2022. Realtime QA: what's the answer right now? *CoRR*, abs/2207.13332.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomás Kociský, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 29348–29363.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix X. Yu, and Sanjiv Kumar. 2023a. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1774–1793. Association for Computational Linguistics.

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 (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12286–12312. Association for Computational Linguistics.

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 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7052–7063. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9802–9822. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1906–1919. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023. Locally typical sampling. *Trans. Assoc. Comput. Linguistics*, 11:102–121.

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 (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10056–10070. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4812–4829. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits

of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5418–5426. Association for Computational Linguistics.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *CoRR*, abs/2302.04761.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023a. Trusting your evidence: Hallucinate less with context-aware decoding. *CoRR*, abs/2305.14739.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023b. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3784–3803. Association for Computational Linguistics.

Hao Sun, Xiao Liu, Yeyun Gong, Yan Zhang, and Nan Duan. 2023. Beamsearchqa: Large language models are strong zero-shot qa solver. *arXiv preprint arXiv:2305.14766*.

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*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023a. SCOTT: self-consistent chain-of-thought distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5546–5558. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023c. Resolving knowledge conflicts in large language models. *CoRR*, abs/2310.00935.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023d. Self-knowledge guided retrieval augmentation for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10303–10315. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. *CoRR*, abs/2303.11315.

A Analyses of the Performances of Existing Decoding Methods

We compare the performance of three baseline methods (introduced in Section 4.1) on LLaMA2-13B model: Regular takes the context and question as input with greedy decoding, and the other two methods are specialized in conflict-solving decoding strategies. We experiment on three subsets of NQ (Kwiatkowski et al., 2019): data without conflict(∼1K), data with all conflicts (∼3K), and random sampled data with half-ratio conflicts. The details of these datasets is introduced in Appendix. E and the detailed experimental results are illustrated in the NQ dataset of Table 1 and Figure 4. We use Exact Match (EM) as the major evaluation metric in the comparison.

In the table presented below, it is observed that when conflicts occur 100% of the time, both conflict-solving decoding methods address the issue more effectively than Regular. Notably, CAD exhibits a pronounced improvement, achieving a significant increase of up to 2.08 in the EM score. Nevertheless, when the ratio of conflict decreases, there is a discernible decrease in those methods' efficacy. Especially, the performance of CAD noticeably deteriorates, trailing behind the Regular by a margin of 5.55.

B Information-Theoretic Properties of Language Models

B.1 Locally Typical Set

Meister et al. (2023) posit that the language modeling can be conceptualized as a discrete stochastic process and build its notion on the concept of **typical set**. Informally, the typical set, derived from information theory, is the set of all samples that we would expect when sampling from the language model distribution. But it relies on a stationary and ergodic language process which contradicts with the non-ergodic language process. So they define a more restrictive notion of typical set - termed as **locally typical set** - for the language process, from which each token generates in a natural and error-minimizing manner.

Definition B.1 (Locally Typical Set)**.** *Let $\boldsymbol{Y} = \{Y_t\}_{t=1}^{\infty}$ be a discrete stochastic process under distribution p. The (T, ε)-**locally typical set** of $\boldsymbol{Y}$ is the set of all sequences of length exactly T such that*

$$\mathcal{L}_{\varepsilon}^{(T)} = \Big\{ \boldsymbol{y} = \boldsymbol{y}_0 \cdots \boldsymbol{y}_T \mid \forall 1 \le t \le T, \quad \Big| \log p(y_t \mid \boldsymbol{y}_{<t}) + \mathrm{H}(Y_t \mid \boldsymbol{Y}_{<t} = \boldsymbol{y}_{<t}) \Big| < \varepsilon \Big\} \tag{14}$$

The relationship can be formalized as the following hypothesis, which has been verified empirically using data from human language process.

Hypothesis B.2. *Samples $\boldsymbol{y} = \boldsymbol{y}_0 \cdots \boldsymbol{y}_T$ from a human language process with distribution p tend to belong to the process's locally typical set $\mathcal{L}_{\varepsilon}^{(T)}$ for large enough T and some $\varepsilon > 0$. In words, this means that we should expect every word in natural-sounding sentences to be close to the* expected *information content under p, i.e., the conditional entropy given prior context.*

The H represents the entropy rate of $\boldsymbol{Y}$, which is equivalent to the standard definition of (Shannon) entropy $\mathcal{H}$ for a random variable $\boldsymbol{Y}$. The locally typical set restricts the set of tokens to those for which each has an information context——measured by its negative log probability——close to the *expected* information content given prior context, i.e., the entropy of the distribution $p(\cdot \mid \boldsymbol{y}_{<t})$.

B.2 Stable Entropy Hypothesis

Arora et al. (2023) postulate that natural language generations usually lie in a narrow and nearly flat entropy band. In the empirical analyses, they observe that, the mean entropy of a language model remains stable over the length of the generation, which is defined as the **stable entropy baseline**[6] in Eq.15. Under the context distribution at time t, an input $\boldsymbol{x}$ and vocabulary $\mathcal{V}$, $y_t \in \mathcal{V}$:

$$\mu_{\mathcal{H}}(t; \mathcal{V}) = \mathbb{E}_{y_t \in \mathcal{V}}\big[\mathcal{H}(y_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t})\big]. \tag{15}$$

Then a **stable entropy zone** is defined as the zone around the stable entropy baseline that covers a major fraction of entropy of the model under the target distribution. They define it by standard deviation ($\sigma_{\mathcal{H}}(t; \mathcal{V})$) around the stable entropy baseline as the stable entropy zone and posit the following hypothesis:

Hypothesis B.3. *Decoding algorithms whose generation's smoothed entropy stays mostly enclosed within the stable entropy zone will produce higher quality, coherent, less repetitive, and more "human-like" text.*

[6] Here we drop the smoothing step for brevity.

C Detailed Proofs of Propositions

Assumption C.1. *If a task-specific context $\boldsymbol{c}$ is contained by parametric knowledge (denoted as $\mathcal{K}$) without triggering any conflicts in model p, then it also can be the natural generation of model.*

$$\textit{if } \boldsymbol{c} \in \mathcal{K}, \textit{ then } \boldsymbol{c} \in \bigcup \boldsymbol{y} \sim p(\cdot \mid \boldsymbol{x})$$

Here, $\bigcup \boldsymbol{y}$ indicates the sampling set of all natural generations by the model conditioning by the question $\boldsymbol{x}$. Then we define the entropy of the model following Arora et al. (2023) as

$$\mathcal{H}(\boldsymbol{y}_t|\boldsymbol{x}, \boldsymbol{y}_{<t}) = \mathop{\mathbb{E}}_{y_t \sim p(\cdot|\boldsymbol{y}_{<t})} - \log p(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}) \tag{16}$$

For brevity, we use $\mathcal{H}_1(\boldsymbol{y_t})$ to represent the entropy of conditional distribution over question $\boldsymbol{x}$ and generation $\boldsymbol{y}_{<t}$, and $\mathcal{H}_2(\boldsymbol{y_t})$ denotes the entropy conditioning by $\boldsymbol{x}$, $\boldsymbol{y}_{<t}$, and assumed generation $\boldsymbol{c}$.

$$\mathcal{H}_1(\boldsymbol{y_t}) = \mathcal{H}(\boldsymbol{y}_t|\boldsymbol{x}, \boldsymbol{y}_{<t}) \tag{17}$$

$$\mathcal{H}_2(\boldsymbol{y_t}) = \mathcal{H}(\boldsymbol{y}_t|\boldsymbol{x}, \boldsymbol{c}, \boldsymbol{y}_{<t}) \tag{18}$$

where $\mathcal{H}_2(\boldsymbol{y_t})$ denotes the entropy conditioning by previously generated tokens $\boldsymbol{c}$ and $\boldsymbol{y}_{<t}$, and $\mathcal{H}_1(\boldsymbol{y_t})$ represents the entropy of conditional distribution over generation $\boldsymbol{y}_{<t}$.

Proposition C.2 (Bound on Entropy Shift)**.** *The entropy shift denoted as $\mathcal{H}_2(\boldsymbol{y_t}) - \mathcal{H}_1(\boldsymbol{y_t})$ is bounded within the width of the **stable entropy zone**.*

Proof. Note that context c is the natural generation of a language model in the setting, both the entropy $\mathcal{H}_1(\boldsymbol{y_t})$ and $\mathcal{H}_2(\boldsymbol{y_t})$ should fall into a stable entropy zone around the mean entropy $\mu_{\mathcal{H}}$. Let β be the threshold of a certain standard deviation around the mean entropy. According to Eq. 15, it can be deduced that

$$\left|\mathcal{H}_1(\boldsymbol{y_t}) - \mu_{\mathcal{H}_1}\right| < \frac{\beta}{2}, \left|\mathcal{H}_2(\boldsymbol{y_t}) - \mu_{\mathcal{H}_2}\right| < \frac{\beta}{2} \tag{19}$$

Stable entropy baseline demonstrates that mean entropy of a model under the target context distribution remains stable. Since the length of context is limited, the mean entropy $\mu_{\mathcal{H}_1}$ and $\mu_{\mathcal{H}_2}$ can be equated if smoothed, denoted as $\mu_{\mathcal{H}}$. Considering inequalities in Eq.(19) jointly, we can obtain the bound on the entropy shift using the triangle inequality:

$$\begin{aligned} &\left|\mathcal{H}_2(\boldsymbol{y_t}) - \mathcal{H}_1(\boldsymbol{y_t})\right| \\ &= \left|\left(\mathcal{H}_2(\boldsymbol{y_t}) - \mu_{\mathcal{H}}\right) - \left(\mathcal{H}_1(\boldsymbol{y_t}) - \mu_{\mathcal{H}}\right)\right| \\ &< \left|\mathcal{H}_2(\boldsymbol{y_t}) - \mu_{\mathcal{H}}\right| + \left|\mathcal{H}_1(\boldsymbol{y_t}) - \mu_{\mathcal{H}}\right| < \beta \end{aligned} \tag{20}$$

■

Proposition C.3 (Bound on information-entropy shift)**.** *As the information content of a random variable is quantified as its negative log-probability. Let the information content $I(y_t) = -\log p(y_t \mid \boldsymbol{x}, \boldsymbol{c}, \boldsymbol{y}_{<t})$, we denote the **information-entropy shift** as: $I(y_t) - \mathcal{H}_1(\boldsymbol{y_t})$. The following bound holds for a constant:*

$$\left|I(y_t) - \mathcal{H}_1(\boldsymbol{y_t})\right| < \gamma \tag{21}$$

where $\gamma > 0$.

Proof. **Locally typicality** demonstrates that the information content of y should is quite close to a specific value of the entropy under model distribution p. It means that there exists a sufficiently small constant $\epsilon > 0$:

$$\left|I(y_t) - \mathcal{H}_2(\boldsymbol{y_t})\right| < \epsilon \tag{22}$$

which bounds the information of y into a coherent and contextual generation. Applying triangle inequality on Eq.(20) and Eq.(22), the following inequality holds for a constant:

$$\begin{aligned} &\left|I(y_t) - \mathcal{H}_1(\boldsymbol{y_t})\right| \\ &= \left|\left(I(y_t) - \mathcal{H}_2(\boldsymbol{y_t})\right) + \left(\mathcal{H}_2(\boldsymbol{y_t}) - \mathcal{H}_1(\boldsymbol{y_t})\right)\right| \\ &< \left|I(y_t) - \mathcal{H}_2(\boldsymbol{y_t})\right| + \left|\mathcal{H}_2(\boldsymbol{y_t}) - \mathcal{H}_1(\boldsymbol{y_t})\right| \\ &< \beta + \epsilon = \gamma \end{aligned} \tag{23}$$

■

D Empirical Study of Assumption C.1

In this section, we show that the distribution entropy of non-conflicting context remains more stable than the non-conflicting one. Then the assumption C.1 can be proved with the **stable entropy hypothesis** B.3.

To demonstrate our assumption, we follow a similar setup as Arora et al. (2023) in a text completion setup. We use the LLaMA2-13B model and NQ data. We sample 500 pieces of data from Conf. and Non-Conf. sub-datasets respectively, then compute the mean smoothed entropy at each step and calculate the standard deviation (std) for each generation. Figure 8 visualizes the std of smoothed entropy for conflicting and non-conflicting generation. The vertical axis, labeled 'Std of Smoothed Entropy', represents the std of each step's entropy in the generation. The horizontal axis represents the NQ samples from Conf. and Non-Conf. data. From the violin plot, it can be observed that the

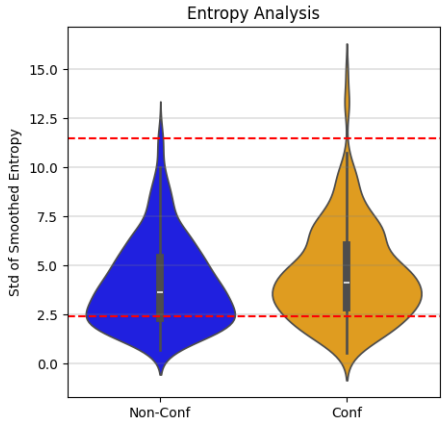


Figure 8: Conf. or Non-Conf. distributions of the 'Std of Smoothed Entropy' for NQ dataset.

entropy distribution for Conf. data exhibit a bimodal nature, suggesting that quite a few samples are characterized by large variances. Furthermore, the box line of Conf. is higher than the one of Non-Conf., which demonstrates that the Non-Conf. is more likely to be a natural generation of the model due to its more stable entropy levels. This is deduced by the **stable entropy hypothesis**, which posits that "*generation's smoothed entropy stays mostly enclosed within the stable entropy zone will produce higher quality, coherent, less repetitive, and more 'human-like' text.*".

E Dataset Details

We use three realistic QA datasets (NaturalQuestions (Kwiatkowski et al., 2019), SQuAD 1.1 (Rajpurkar et al., 2016)), and StrategyQA (Geva et al., 2021a)) and one conflicting QA dataset (Counterfacts dataset (Longpre et al., 2021)) for evaluating our method.

NaturalQuestions consists of real-world information-seeking queries issued to the Google search engine and their corresponding long answers (gold evidence passage) and short answers (one or more entities). In our study, we employ the long answers as the input context and short answers as the ground truth, and conduct evaluations on the dev set.

The SQuAD 1.1 is a common QA benchmark. It includes questions posed by human annotators on a given Wikipedia paragraph, where the answer to each question is a segment of text (or span) from the paragraph. In our experiments, we conduct experiments on the dev for evaluation.

StrategyQA is a fact reasoning benchmark that necessitates the implicit question decomposition into reasoning steps. Built around Wikipedia terms, these questions are accompanied by multiple evidence paragraphs. The model is expected to provide a True or False answer. We concatenate question-relevant evidences to form the input context. We adopt the training set for evaluation, considering the volume of data.

Counterfacts is based on the NaturalQuestions (Kwiatkowski et al., 2019) dataset. To generate conflicting contextual knowledge, Longpre et al. (2021) first identify questions with named entity answers, find the supportive document for each question and then replace the gold answer entity in the document with a random entity.

E.1 Posteriori judgement

We delineates the process of identifying instances of knowledge conflicts. The evaluation of these conflicts is based on the accuracy[7] of the model's responses when context is not provided. The scenarios are divided into two categories:

- **Non-Conflicting (Non-Conf.)**: This category pertains to situations where the model is capable of accurately responding to a question without the need for its corresponding context. Such instances suggest that the model has internalized the context, thereby indicating a consistency between its parametric knowledge and the external contextual knowledge.

- **Conflicting (Conf.)**: When the model fails to provide the true answer without the aid of context, indicating a conflict between its inherent parametric knowledge and the external contextual knowledge. Following Wang et al. (2023d), incorrect responses reflects the model does not possess the knowledge equipped by the external context, which has a discrepancy with the model's parametric knowledge.

In this setting, the NQ, SQuAD and StrategyQA datasets can serve as suitable approximations of realistic scenarios where conflicts may not necessarily occur. Additionally, the synthetic dataset named Counterfacts, which is composed exclusively of conflicting data (Conf. data), serves as a unique

[7]Given the excessively rigid nature of EM for evaluation, an F1 score of 0.5 has been employed as a proxy for preliminary categorization.

case. This is because it contains randomly replaced answers that are not inherently known to the model, distinguishing it from the aforementioned datasets.

E.2 Data Statistic

Datasets		LLaMA2 7B	LLaMA2 13B	OPT 6.7B	OPT 13B	FLAN-T5 3B	FLAN-T5 11B
NQ (∼4K)	Total(%)	100	100	100	100	100	100
	Conf.(%)	81.91	76.79	99.34	97.21	88.07	85.80
	Non-Conf.(%)	18.09	23.21	0.64	2.79	11.93	14.20
SQuAD (∼6K)	Total(%)	100	100	100	100	100	100
	Conf.(%)	84.18	82.06	97.48	95.41	92.30	90.55
	Non-Conf.(%)	15.82	17.94	2.56	4.59	7.70	9.45
StrategyQA (∼2K)	Total(%)	100	100	100	100	100	100
	Conf.(%)	40.31	39.43	94.98	88.91	36.11	33.23
	Non-Conf.(%)	59.69	60.57	5.02	11.09	63.89	66.77
Counterfacts (∼6K)	Conf.(%)	100	100	100	100	100	100

Table 5: The data distributions of the datasets

As illustrated in Table 5, a discernible trend emerges wherein an escalation in the model's parameters is accompanied by a corresponding increase in the percentage of non-conflicting data, signifying a greater degree of internalized knowledge within larger models. Notably, among this cohort of models, the OPT series models exhibit the lowest parametric knowledge, yet they demonstrate substantial enhancements across most datasets when the COIECD method is applied. It is also noteworthy to observe that even in the case of the popular LLaMA2 models, the proportion of non-conflicting data does not surpass 25% in the NQ and SQuAD datasets. This observation necessitate the further research for the inherent parametric knowledge enhancement of the model.

F Baseline Methods

Contrastive Decoding (CD) In our experiments, we employ the distribution $g(y_t)$ with a certain threshold as a baseline decoding method, referred to as the CD (Li et al., 2023b) method. We modify the original object of CD (computes the distribution discrepancy between an small amateur model and an expert larger model) to simulate the form of $g(y_t)$.

$$\begin{aligned} CD_{\text{original}} &= \log p_{\text{EXP}}(y_t|x, y<t) - \\ &\quad \log p_{\text{AMA}}(y_t|x, y<t) \\ CD_{\text{modify}} &= \log p(y_t|x, y<t) - p(y_t|y<t) \\ &= \log g(y_t) \end{aligned}$$

The threshold is same as in the original CD method:

$$\mathcal{V}_{\text{head}}(y_{<t}) = \left\{ y_t \in \mathcal{V} : p(y_t|y_{<t}) \geq 0.1 \cdot \max_y p(y|y_{<t}) \right\}$$

Here, we represent the input context as x. CD adopts the object of difference between the output likelihood when inputs are presented with and without input context. It enhances the influence of the context for high-probability words within a crude threshold. Therefore, it cannot obtain consistent improvement in performance, particularly with non-conflicting data.

And the Section 3.1 aims to explore a delicate constraint for output distribution to find out whether the context is in conflict. Then we propose a contextual information-entropy constraint on fine-grained token level based on the perspective of information theory.

$$C(y_{<t}) = \{ y \in \mathcal{V} : l_{p_s} \leq p_s(y_t) \leq u_{p_s} \} \quad (7)$$

Context-Aware Decoding (CAD) In CAD (Shi et al., 2023a) method, the output probability is a product-of-experts of the original output probability and PMI weighted by α as follow:

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

Since they set α = 1 for all models evaluated on the knowledge conflict datasets, this method can be regarded as an unconstrained ($\lambda = 1\ in\ C(y_{<t})$) decoding method when α is set to 1. If so, CAD can be considered as a specific case of our approach.

Furthermore, CAD, as evidenced in their experimental evaluation, necessitates the different hyperparameter values (the adjustment level of CAD is 0.5 and 1) for realistic datasets and counterfacts. The absence of such specific adjustments results in a substantial decline in performance. This aspect of our findings underscores the superiority and robustness of our method.

G Maximization v.s. Sampling Strategies

Recall that prior experiments are conducted based on greedy strategy that maximizes the distribution probability, except for SC with a fixed sampling strategy. We explore other strategies like sampling alternatives based on the same baselines. Table 6 represents the results on maximization-based strategies: greedy decoding, and stochastic sampling: nucleus (Holtzman et al., 2019), top-k (Fan et al., 2018), typical (Meister et al., 2023) on the NQ dataset of LLaMA2-13B.

We observe that COIECD consistently produces the higher EM and F1 score the than Regular irrespective of the choice of decoding strategy. In

contrast, both the CD and CAD exhibit a lack of stability in performance among diverse decoding strategies. Additionally, the result points to the significant value of beam search, particularly in relation to CD, in boosting performance. It can be attributed to the increasing search width, a feature of beam search which effectively eliminates disturbing tokens brought by contrastive object.

Decoding Methods	Regular		CD		CAD		COIECD	
	EM	F1	EM	F1	EM	F1	EM	F1
Greedy	46.48	61.51	46.19	61.97	46.79	62.29	**47.42**	**62.89**
Nucleus								
($p = 0.9$)	43.80	59.75	46.14	61.73	44.37	60.50	**46.19**	**62.13**
($p = 0.95$)	43.82	60.05	45.77	62.03	43.17	59.45	**46.53**	**62.80**
Top-k								
($k = 30$)	42.46	58.64	46.03	61.86	41.88	58.37	**46.98**	**62.14**
($k = 50$)	41.64	58.27	45.37	61.42	41.82	58.54	**46.16**	**61.87**
Typical								
($\tau = 0.2$)	45.08	61.03	46.06	61.93	45.14	60.70	**47.08**	**62.75**
($\tau = 0.9$)	44.32	60.24	46.37	61.97	43.77	60.01	**46.74**	**62.03**

Table 6: **Decoding on maximization-based and stochastic sampling strategies.** The red cell indicates superior performance than the Regular decoding, and green denotes degeneration.

H Case Study

As illustrated in Figure 9, we look closer into two cases of conflicting and non-conflicting one.

Lower-bound & Upper-bound violation. The conflicting case mainly shows the function of lower bound. For token y_t, if $p_\delta(y_t) \leq l_{p_\delta}$, it represents a sufficiently low information content $I(y_t)$ compared to the entropy $\mathcal{H}_1(\boldsymbol{y_t})$. This indicates that the generation (like *Russia*) may be overconfident and other informative gains (like *Qatar*) may be ignored, then the conflict occurs. The upper bound serves to filter some disturbing low-probability distribution, which plays a role in stochastic sampling decoding. In the non-conflicting case, if $p_\delta(y_t) \geq u_{p_\delta}$, the high information context represents a lower probability, indicating that the model is less certain about current token (like *Germany*). The decreased confidence might also be attributed to a potential conflict within the context.

No violation. In the non-conflicting case, it is observed that no tokens fall into the lower-violation zone. This can be attributed to the model's pronounced confidence in a solitary high-probability token, identified as *Russia*. Such a high degree of confidence leads to the assignment of a zero value to l_{p_δ}. The rationale behind this assignment stems from the understanding that a heightened level of confidence effectively indicates the non-existence of any conflict.

I Detailed Results on Hyperparameter Analysis

Here we display the detailed results about hyperparameter analysis on different sizes of LLaMA2 model with EM and F1 metrics.

EM Score	λ=0.1	λ=0.25	λ=0.5	λ=1
α=0	19.30	16.82	15.40	14.25
α=0.5	38.88	36.49	34.12	31.70
α=1.0	47.08	**47.42**	47.21	46.48
α=1.5	46.19	46.82	46.85	46.79
α=2.0	36.51	36.38	36.07	35.75

Table 7: Exact Match score on LLaMA2-13B Model.

F1 Score	λ=0.1	λ=0.25	λ=0.5	λ=1
α=0	31.67	27.76	25.40	23.09
α=0.5	56.49	53.97	51.04	47.30
α=1.0	62.51	**62.89**	62.43	61.51
α=1.5	61.72	62.29	62.28	62.28
α=2.0	54.52	54.34	53.90	53.56

Table 8: F1 score on LLaMA2-13B Model.

EM Score	λ=0.1	λ=0.25	λ=0.5	λ=1
α=0	15.19	13.46	12.28	11.51
α=0.5	40.22	38.33	36.49	34.73
α=1.0	45.79	**46.08**	45.56	44.64
α=1.5	44.93	45.37	45.16	45.08
α=2.0	40.09	39.72	39.62	39.54

Table 9: Exact Match score on LLaMA2-7B Model.

F1 Score	λ=0.1	λ=0.25	λ=0.5	λ=1
α=0	25.36	22.51	20.13	18.66
α=0.5	55.00	53.40	51.50	49.50
α=1.0	59.44	**59.67**	59.12	58.57
α=1.5	58.71	59.06	58.86	58.89
α=2.0	55.37	55.12	54.97	54.97

Table 10: F1 score on LLaMA2-7B Model.

The different values of alpha can measure the importance of adding $g(y_t)$ in Eq. 12 (§ 3.2). The results highlight the significance of adding $g(y_t)$. The performance declines dramatically when α equals 0. It's under a decoding strategy where simply providing or not providing the context during decoding.

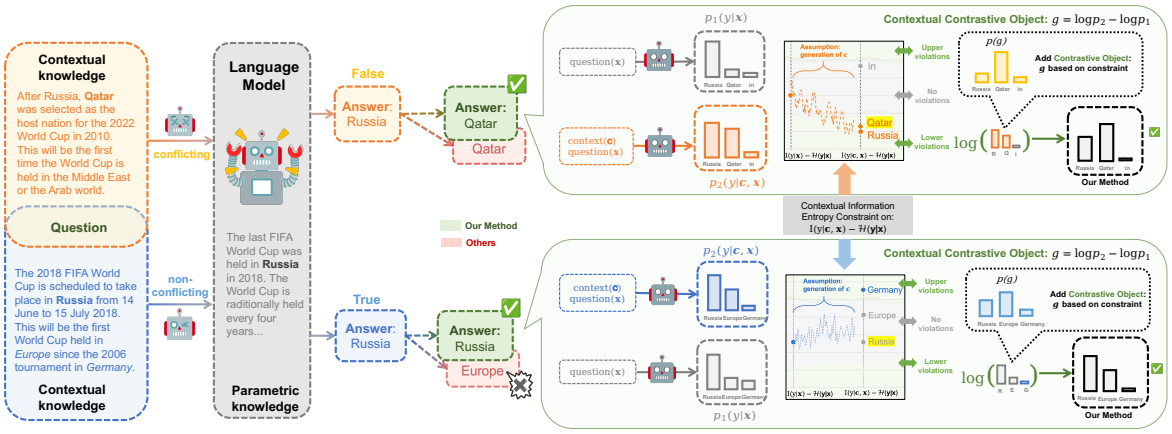


Figure 9: Left: The illustration of conflicting and non-conflicting scenarios. Existing methods adeptly handle conflicts but struggle to address non-conflicting context. In contrast, COIECD exhibits the capability to effectively handle both scenarios. Right: The detailed process of COIECD method. Utilizing a contextual information-entropy constraint, we discern the tokens that violate this constraint, which are typically triggered by conflicting contexts. For these tokens, situated in different zones, we employ distinct decoding strategies.

J More Results

We present the experimental results of GPT-3.5 and GPT-4 in Table 11, as well as the results on the other size of LLaMA2, OPT and FLAN-T5 models in Table 12-14.

J.1 The performances of GPT-3.5 and GPT-4

In general, the GPT-4 model displays a modestly superior performance in comparison to the models utilized in our experiments, whereas GPT-3.5 attains a level of performance that aligns with our best results achieved by the LLaMA2-13B model.

Datasets		GPT-3.5 EM	GPT-3.5 F1	GPT-4 EM	GPT-4 F1
NQ	Total	44.45	61.63	47.36	65.28
	Conf.	31.46	50.07	35.04	54.68
	Non-Conf.	70.16	84.51	78.44	89.66
SQuAD	Total	58.16	75.74	63.02	78.42
	Conf.	52.39	71.40	57.58	75.92
	Non-Conf.	78.07	90.71	82.63	93.36
StrategyQA	Total	82.75	82.75	91.22	91.22
	Conf.	68.29	68.29	78.83	78.83
	Non-Conf.	91.46	91.46	96.67	96.67
Counterfacts	Total (Conf.)	61.69	66.40	64.66	71.11

Table 11: The Performances of GPT-3.5 and GPT-4

J.2 The Performances of Models in Different Size

Owing to the constraints of experimental resources, we confined our model within the scope of a maximum parameter capacity of 13B for the experiments. In addition to the main results in experiment section, additional outcomes are illustrated in the Table 12-14.

Datasets		Decoding	LLaMA2-7B		LLaMA2-13B	
			EM	F1	EM	F1
NQ	Total	Regular	44.64	58.60	46.48	61.51
		SC	44.72 (+0.08)	58.47 (−0.13)	46.66 (+0.18)	61.76 (+0.25)
		CD	45.35 (+0.71)	59.21 (+0.61)	46.19 (−0.29)	61.97 (+0.46)
		CAD	45.08 (+0.44)	58.89 (+0.29)	46.79 (+0.31)	62.29 (+0.78)
		COIECD	**46.08** (+1.44)	**59.67** (+1.07)	**47.42** (+0.94)	**62.89** (+1.38)
	Conf.	Regular	39.06	53.66	38.45	54.37
		SC	38.99 (−0.07)	53.43 (−0.23)	38.65 (+0.20)	54.64 (+0.27)
		CD	40.69 (+1.63)	55.30 (+1.64)	39.64 (+1.19)	56.50 (+2.13)
		CAD	40.82 (+1.76)	**55.45** (+1.79)	**40.53** (+2.08)	**57.15** (+2.78)
		COIECD	**40.95** (+1.89)	55.24 (+1.58)	39.88 (+1.43)	56.59 (+2.22)
	Non-Conf.	Regular	69.91	80.93	73.05	85.15
		SC	**70.64** (+0.73)	**81.26** (+0.33)	**73.16** (+0.11)	**85.30** (+0.15)
		CD	66.42 (−3.49)	76.93 (−4.00)	67.84 (−5.21)	80.06 (−5.09)
		CAD	64.39 (−5.52)	74.46 (−6.47)	67.50 (−5.55)	79.19 (−5.96)
		COIECD	69.33 (−0.58)	79.71 (−1.22)	72.37 (−0.68)	83.75 (−1.40)
SQuAD	Total	Regular	54.75	68.92	54.46	68.92
		SC	55.02 (+0.27)	69.04 (+0.12)	54.55 (+0.09)	68.85 (−0.07)
		CD	**57.56** (+2.81)	**70.94** (+2.02)	53.89 (−0.57)	68.04 (−0.88)
		CAD	56.98 (+2.23)	70.12 (+1.20)	56.46 (+2.00)	70.52 (+1.60)
		COIECD	57.32 (+2.57)	70.39 (+1.47)	**57.10** (+2.64)	**70.86** (+1.94)
	Conf.	Regular	50.11	65.17	48.78	64.34
		SC	50.32 (+0.21)	65.25 (+0.08)	48.87 (+0.09)	64.24 (−0.10)
		CD	**54.36** (+4.25)	**68.51** (+3.34)	50.68 (+1.90)	66.01 (+1.67)
		CAD	53.33 (+3.22)	67.43 (+2.26)	51.64 (+2.86)	**67.09** (+2.75)
		COIECD	53.41 (+3.30)	67.42 (+2.25)	**51.95** (+3.17)	66.91 (+2.57)
	Non-Conf.	Regular	79.44	88.84	80.50	89.88
		SC	**80.09** (+0.65)	**89.20** (+0.36)	80.57 (+0.07)	**89.96** (+0.08)
		CD	74.57 (−4.87)	83.85 (−4.99)	68.64 (−11.86)	77.35 (−12.53)
		CAD	76.41 (−3.03)	84.44 (−4.40)	78.53 (−1.97)	86.19 (−3.69)
		COIECD	78.14 (−1.30)	86.21 (−2.63)	**80.69** (+0.19)	88.93 (−0.95)
StrategyQA	Total	Regular	79.69	79.69	81.09	81.09
		SC	**79.34** (−0.35)	**79.34** (−0.35)	81.05 (−0.04)	81.05 (−0.04)
		CD	69.96 (−9.73)	69.96 (−9.73)	83.58 (+2.49)	83.58 (+2.49)
		CAD	74.93 (−4.76)	74.93 (−4.76)	85.50 (+4.41)	85.50 (+4.41)
		COIECD	78.91 (−0.78)	78.91 (−0.78)	**85.76** (+4.67)	**85.76** (+4.67)
	Conf.	Regular	61.11	61.11	57.36	57.36
		SC	61.11 (+0.00)	61.11 (+0.00)	57.59 (+0.23)	57.59 (+0.23)
		CD	59.15 (−1.96)	59.15 (−1.96)	**81.15** (+23.79)	**81.15** (+23.79)
		CAD	**64.57** (+3.46)	**64.57** (+3.46)	77.31 (+19.95)	77.31 (+19.95)
		COIECD	63.71 (+2.60)	63.71 (+2.60)	80.29 (+22.93)	80.29 (+22.93)
	Non-Conf.	Regular	92.25	92.25	96.54	96.54
		SC	**91.66** (−0.59)	**91.66** (−0.59)	**96.47** (−0.07)	**96.47** (−0.07)
		CD	77.25 (−15.00)	77.25 (−15.00)	85.16 (−11.38)	85.16 (−11.38)
		CAD	81.93 (−10.32)	81.93 (−10.32)	89.33 (−7.21)	89.33 (−7.21)
		COIECD	89.17 (−3.08)	89.17 (−3.08)	90.80 (−5.74)	90.80 (−5.74)
Counterfacts	Total (Conf.)	Regular	67.86	68.77	61.67	62.63
		SC	68.30 (+0.44)	69.23 (+0.46)	61.76 (+0.09)	62.76 (+0.13)
		CD	72.94 (+5.08)	74.29 (+5.52)	67.96 (+6.29)	69.16 (+6.53)
		CAD	**73.11** (+5.25)	**75.99** (+7.22)	**68.76** (+7.09)	**71.20** (+8.57)
		COIECD	71.57 (+3.71)	68.86 (+0.09)	68.30 (+6.63)	69.33 (+6.70)

Table 12: The results of LLaMA2-7B and LLaMA2-13B.

Datasets		Decoding	OPT-6.7B		OPT-13B	
			EM	F1	EM	F1
NQ	Total	Regular	19.74	26.25	21.11	30.14
		SC	24.24 (+4.50)	29.78 (+3.53)	24.40 (+3.29)	33.31 (+3.17)
		CD	22.90 (+3.16)	34.48 (+8.23)	17.30 (−3.81)	27.63 (−2.51)
		CAD	29.15 (+9.41)	40.16 (+13.91)	24.76 (+3.65)	36.37 (+6.23)
		COIECD	**30.07** (+10.33)	**40.77** (+14.52)	**27.08** (+5.97)	**38.87** (+8.73)
	Conf.	Regular	19.79	26.24	20.72	29.33
		SC	24.26 (+4.47)	29.75 (+3.51)	23.82 (+3.10)	32.54 (+3.21)
		CD	22.96 (+3.17)	34.54 (+8.30)	17.25 (−3.47)	27.58 (−1.75)
		CAD	29.21 (+9.42)	40.19 (+13.95)	24.55 (+3.83)	36.19 (+6.86)
		COIECD	**30.13** (+10.34)	**40.78** (+14.54)	**26.80** (+6.08)	**38.63** (+9.30)
	Non-Conf.	Regular	12.40	27.03	40.57	56.81
		SC	**21.79** (+9.39)	34.07 (+7.04)	**44.34** (+3.77)	**60.36** (+3.55)
		CD	12.51 (+0.11)	25.23 (−1.80)	18.87 (−21.70)	29.18 (−27.63)
		CAD	20.83 (+8.43)	**36.05** (+9.02)	32.08 (−8.49)	42.77 (−14.04)
		COIECD	19.01 (+6.61)	35.82 (+8.79)	36.79 (−3.78)	47.30 (−9.51)
SQuAD	Total	Regular	21.49	28.50	27.91	37.37
		SC	23.64 (+2.15)	30.97 (+2.47)	30.13 (+2.22)	40.08 (+2.71)
		CD	26.35 (+4.86)	37.90 (+9.40)	28.03 (+0.12)	37.51 (+0.14)
		CAD	29.46 (+7.97)	40.31 (+11.81)	35.01 (+7.10)	47.34 (+9.97)
		COIECD	**29.93** (+8.44)	**40.47** (+11.97)	**35.13** (+7.22)	**47.48** (+10.11)
	Conf.	Regular	21.49	28.50	27.31	36.27
		SC	23.14 (+1.65)	30.18 (+1.68)	29.57 (+2.26)	39.07 (+2.80)
		CD	26.33 (+4.84)	37.61 (+9.11)	27.42 (+0.11)	36.38 (+0.11)
		CAD	29.32 (+7.83)	39.97 (+11.47)	34.79 (+7.48)	**46.93** (+10.66)
		COIECD	**29.78** (+8.29)	**40.13** (+11.63)	**34.95** (+7.64)	46.84 (+10.57)
	Non-Conf.	Regular	35.62	57.10	40.30	60.28
		SC	**36.59** (+0.97)	**56.04** (−1.06)	**41.79** (+1.49)	**61.17** (+0.89)
		CD	26.90 (−8.72)	49.05 (−8.05)	40.67 (+0.37)	60.98 (+0.70)
		CAD	34.93 (−0.69)	53.44 (−3.66)	40.41 (+0.11)	58.93 (−1.35)
		COIECD	35.69 (+0.07)	53.71 (−3.39)	38.81 (−1.49)	60.88 (+0.60)
StrategyQA	Total	Regular	47.51	47.51	61.79	61.79
		SC	46.64 (−0.87)	46.64 (−0.87)	60.57 (−1.22)	60.57 (−1.22)
		CD	46.99 (−0.52)	46.99 (−0.52)	61.18 (−0.61)	61.18 (−0.61)
		CAD	53.10 (+5.59)	53.10 (+5.59)	62.31 (+0.52)	62.31 (+0.52)
		COIECD	**53.84** (+6.33)	**53.84** (+6.33)	**64.33** (+2.54)	**64.33** (+2.54)
	Conf.	Regular	47.86	47.86	61.48	61.48
		SC	47.03 (−0.83)	47.03 (−0.83)	60.28 (−1.20)	60.28 (−1.20)
		CD	47.26 (−0.60)	47.26 (−0.60)	60.86 (−0.62)	60.86 (−0.62)
		CAD	54.21 (+6.35)	54.21 (+6.35)	61.97 (+0.49)	61.97 (+0.49)
		COIECD	**54.90** (+7.04)	**54.90** (+7.04)	**62.06** (+0.58)	**62.06** (+0.58)
	Non-Conf.	Regular	40.87	40.87	82.35	82.35
		SC	39.13 (−1.74)	39.13 (−1.74)	79.41 (−2.94)	79.41 (−2.94)
		CD	**41.71** (+0.84)	**41.71** (+0.84)	82.35 (+0.00)	82.35 (+0.00)
		CAD	32.17 (−8.70)	32.17 (−8.70)	**85.29** (+2.94)	**85.29** (+2.94)
		COIECD	33.91 (−6.96)	33.91 (−6.96)	82.65 (+0.30)	82.65 (+0.30)
Counterfacts	Total (Conf.)	Regular	18.15	19.38	19.55	20.75
		SC	21.40 (+3.25)	22.62 (+3.24)	21.75 (+2.20)	22.90 (+2.15)
		CD	38.16 (+20.01)	42.78 (+23.40)	39.26 (+19.71)	42.89 (+22.14)
		CAD	**40.10** (+21.95)	**45.29** (+25.91)	**40.44** (+20.89)	**47.46** (+26.71)
		COIECD	37.35 (+19.20)	43.45 (+24.07)	38.68 (+19.13)	46.98 (+26.23)

Table 13: The results of OPT-6.7B and OPT-13B.

Datasets		Decoding	FLAN-T5-3B EM	FLAN-T5-3B F1	FLAN-T5-11B EM	FLAN-T5-11B F1
NQ	Total	Regular	46.00	62.78	44.98	65.02
		SC	46.14 (+0.14)	62.51 (−0.27)	44.28 (−0.70)	65.71 (+0.69)
		CD	37.62 (−8.38)	55.47 (−7.31)	39.06 (−5.92)	60.94 (−4.08)
		CAD	38.91 (−7.09)	57.77 (−5.01)	42.48 (−2.50)	64.20 (−0.82)
		COIECD	**48.84** (+2.84)	**64.45** (+1.67)	**45.14** (+0.16)	**65.98** (+0.96)
	Conf.	Regular	45.16	61.44	42.35	62.69
		SC	45.22 (+0.06)	61.02 (−0.42)	39.77 (−2.58)	62.25 (−0.44)
		CD	38.29 (−6.87)	55.97 (−5.47)	38.68 (−3.67)	60.59 (−2.10)
		CAD	39.34 (−5.82)	58.25 (−3.19)	41.71 (−0.64)	63.61 (+0.92)
		COIECD	**48.36** (+3.20)	**63.98** (+2.54)	**43.86** (+1.51)	**64.73** (+2.04)
	Non-Conf.	Regular	52.20	72.65	60.81	79.09
		SC	52.26 (+0.06)	**73.49** (+0.84)	51.39 (−9.42)	71.55 (−7.54)
		CD	33.40 (−18.80)	52.33 (−20.32)	41.40 (−19.41)	63.03 (−16.06)
		CAD	35.68 (−16.52)	54.22 (−18.43)	47.13 (−13.68)	67.75 (−11.34)
		COIECD	**52.42** (+0.22)	67.87 (−4.78)	**52.87** (−7.94)	**73.52** (−5.57)
SQuAD	Total	Regular	71.20	83.53	66.63	80.88
		SC	70.90 (−0.30)	83.28 (−0.25)	67.96 (+1.33)	81.51 (+0.63)
		CD	71.25 (+0.05)	83.10 (−0.43)	65.04 (−1.59)	79.12 (−1.76)
		CAD	68.62 (−2.58)	81.88 (−1.65)	68.88 (+2.25)	81.91 (+1.03)
		COIECD	**73.84** (+2.64)	**84.99** (+1.46)	**69.89** (+3.26)	**82.59** (+1.71)
	Conf.	Regular	70.51	83.09	65.34	80.01
		SC	70.25 (−0.26)	82.84 (−0.25)	62.07 (−3.27)	77.92 (−2.09)
		CD	71.31 (+0.80)	83.17 (+0.08)	64.57 (−0.77)	78.93 (−1.08)
		CAD	68.64 (−1.87)	81.92 (−1.17)	68.43 (+3.09)	81.73 (+1.72)
		COIECD	**73.51** (+3.00)	**84.76** (+1.67)	**69.20** (+3.86)	**82.22** (+2.21)
	Non-Conf.	Regular	79.56	88.81	78.99	89.15
		SC	**78.67** (−0.89)	**88.61** (−0.20)	**79.58** (+0.59)	**89.18** (+0.03)
		CD	70.60 (−8.96)	82.26 (−6.55)	69.57 (−9.42)	80.89 (−8.26)
		CAD	68.37 (−11.19)	81.42 (−7.39)	73.19 (−5.80)	83.62 (−5.53)
		COIECD	77.78 (−1.78)	87.76 (−1.05)	76.45 (−2.54)	86.14 (−3.01)
StrategyQA	Total	Regular	87.07	87.07	92.84	92.84
		SC	86.81 (−0.26)	86.81 (−0.26)	92.58 (−0.26)	92.58 (−0.26)
		CD	**89.34** (+2.27)	**89.34** (+2.27)	91.79 (−1.05)	91.79 (−1.05)
		CAD	88.69 (+1.62)	88.69 (+1.62)	92.45 (−0.39)	92.45 (−0.39)
		COIECD	88.78 (+1.71)	88.78 (+1.71)	**92.89** (+0.05)	**92.89** (+0.05)
	Conf.	Regular	69.41	69.41	83.44	83.44
		SC	68.80 (−0.61)	68.80 (−0.61)	83.18 (−0.26)	83.18 (−0.26)
		CD	**85.96** (+16.55)	**85.96** (+16.55)	**91.33** (+7.89)	**91.33** (+7.89)
		CAD	77.39 (+7.98)	77.39 (+7.98)	87.06 (+3.62)	87.06 (+3.62)
		COIECD	77.36 (+7.95)	77.36 (+7.95)	87.39 (+3.95)	87.39 (+3.95)
	Non-Conf.	Regular	97.06	97.06	97.51	97.51
		SC	**96.69** (−0.07)	**96.69** (−0.07)	**97.25** (−0.26)	**97.25** (−0.26)
		CD	91.26 (−5.80)	91.26 (−5.80)	92.02 (−5.49)	92.02 (−5.49)
		CAD	95.08 (−1.98)	95.08 (−1.98)	95.39 (−2.12)	95.39 (−2.12)
		COIECD	95.22 (−1.84)	95.22 (−1.84)	95.55 (−1.96)	95.55 (−1.96)
Counterfacts	Total (Conf.)	Regular	74.56	75.73	71.79	74.82
		SC	74.58 (+0.02)	75.64 (−0.09)	72.60 (+0.81)	74.30 (−0.52)
		CD	74.76 (+0.20)	77.30 (+1.57)	70.31 (−1.48)	75.67 (+0.85)
		CAD	68.23 (−6.33)	74.17 (−1.56)	67.39 (−4.40)	74.42 (−0.40)
		COIECD	**77.60** (+3.04)	**78.97** (+3.24)	**75.29** (+3.50)	**78.37** (+3.55)

Table 14: The results of FLAN-T5-3B and FLAN-T5-11B.