When to Speak, When to Abstain: Contrastive Decoding with Abstention

Hyuhng Joon Kim[1], **Youna Kim**[1], **Sang-goo Lee**[1 2], **Taeuk Kim**[3 *]
[1]Seoul National University, [2]IntelliSys, Korea, [3]Hanyang University
{heyjoonkim, anna9812, sglee}@europa.snu.ac.kr
kimtaeuk@hanyang.ac.kr

Abstract

Large Language Models (LLMs) demonstrate exceptional performance across diverse tasks by leveraging pre-trained (*i.e., parametric*) and external (*i.e., contextual*) knowledge. While substantial efforts have been made to enhance the utilization of both forms of knowledge, situations in which models lack relevant information remain underexplored. To investigate this challenge, we first present a controlled testbed featuring four distinct knowledge access scenarios, including the aforementioned edge case, revealing that conventional LLM usage exhibits insufficient robustness in handling all instances. Addressing this limitation, we propose **Contrastive Decoding with Abstention (CDA)**, a novel training-free decoding method that allows LLMs to generate responses when relevant knowledge is available and to abstain otherwise. CDA estimates the relevance of both knowledge sources for a given input, adaptively deciding which type of information to prioritize and which to exclude. Through extensive experiments, we demonstrate that CDA can effectively perform accurate generation and abstention simultaneously, enhancing reliability and preserving user trust.

1 Introduction

Large Language Models (LLMs) (Team et al., 2023; Achiam et al., 2023; Dubey et al., 2024) acquire extensive *parametric knowledge* during pre-training, enabling them to attain remarkable performance across a wide range of tasks. Although parametric knowledge can be comprehensive and highly informative, it is inherently bounded by the scope of the pre-training corpus. Consequently, LLMs become less reliable when processing inputs from underrepresented domains, such as those including domain-specific (Kandpal et al., 2023; Raja et al., 2024; Feng et al., 2024a) or outdated data (Lazaridou et al., 2024; Kasai et al., 2023; Zhao et al., 2024a).

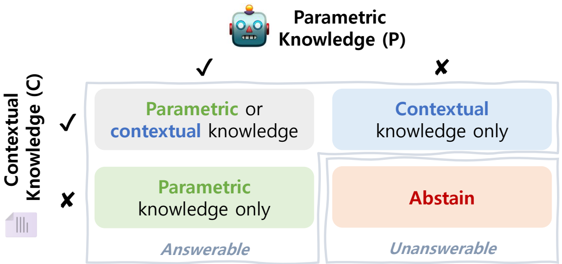

Parametric Knowledge (P)
✓
✗
Contextual Knowledge (C)
✓
✗
Parametric or contextual knowledge
Contextual knowledge only
Parametric knowledge only
Abstain
Answerable
Unanswerable

Figure 1: This study considers four possible scenarios based on the existence of the model's parametric and contextual knowledge. The model is expected to respond *reliably* by either (1) generating correct responses leveraging any form of relevant knowledge available or (2) abstaining from producing potentially inaccurate or misleading outputs when no relevant knowledge exists.

To overcome this challenge, approaches that integrate previously unseen information during inference have emerged (Buttcher et al., 2016; Yin et al., 2016; Karpukhin et al., 2020). They provide external information to LLMs as *contextual knowledge*, expanding the knowledge boundary beyond what is learned from training.

Since LLMs are generally exposed to two distinct sources of information—*parametric* and *contextual* knowledge—they are expected to adaptively leverage both to maximize performance. Despite efforts to enhance such desired behavior, scenarios where neither parametric nor contextual knowledge is available—often encountered in real-world settings—remain largely underexplored. Compelling models to respond imprudently in such cases heightens the risk of hallucination, diminishes reliability, and introduces potential dangers in high-stakes applications.

Therefore, it is crucial for LLMs to abstain from responding when necessary information is inaccessible (Varshney et al., 2024; Zhang et al., 2024a; Wen et al., 2024) while preserving performance

[*] Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9710–9730
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

when relevant knowledge is available. However, such behavior requires a precise assessment of the knowledge and the integration of this assessment into the generation process, both of which are inherently challenging.

In this work, we first present a controlled testbed, where the accessibility of both types of knowledge for a query is explicitly determined. In contrast to typical scenarios where definitively confirming the availability of parametric or contextual knowledge is imprecise, our experimental setup facilitates controlled investigations, covering all scenarios depicted in Figure 1. Experimental results on this testbed indicate that existing methods for LLMs lack sufficient robustness in effectively handling all the considered scenarios.

To this end, we propose **Contrastive Decoding with Abstention (CDA)**, a novel, training-free decoding method that enables LLMs to not only leverage relevant parametric or contextual knowledge during generation but also abstain when no appropriate knowledge is available. During the decoding process, CDA assesses the relevance of both forms of knowledge, adaptively determining the knowledge to attend to during generation. Moreover, CDA steers the models towards abstention if no relevant knowledge is available. The relevancy is estimated as the uncertainty associated with the knowledge in response to a specific query.

Extensive experiments with four LLMs on three question-answering (QA) datasets (Zhang et al., 2023; Etezadi and Shamsfard, 2023) demonstrate that CDA effectively enables LLMs to abstain in the absence of relevant knowledge while maintaining existing capabilities without additional training. Further validation against training-based methods demonstrated CDA's robust generalization capabilities, while evaluations in retrieval-augmented generation (RAG) setting highlight its effectiveness across practical scenarios.

2 Related Work

2.1 Contrastive Decoding

Contrastive decoding (CD) controls text generation by contrasting different output distributions and steers the model in the desired direction. DExperts (Liu et al., 2021) employs an ensemble of an "expert" and an "anti-expert" for tasks such as detoxification. Li et al. (2023) contrasts the output distributions of a large LM and a small LM for open-ended text generation. CD is also proven effective in domains such as reasoning (O'Brien and Lewis, 2023) and machine translation (Waldendorf et al., 2024). Recently, there has been growing interest in context-aware contrastive decoding (CCD) (Zhao et al., 2024b; Kim et al., 2024b; Qiu et al., 2024; Shi et al., 2024b), which enables the model to leverage both parametric and contextual knowledge during decoding, tackling tasks such as knowledge conflicts (Longpre et al., 2021; Chen et al., 2022; Zhou et al., 2023). Despite the promising results, existing approaches assume that at least one knowledge source is always available. In practice, LLMs frequently encounter situations with no relevant knowledge, a gap that these methods fail to bridge. To address this limitation, we expand the scope to include such edge cases and propose a novel approach of integrating abstention to CCD.

2.2 Abstention in LLMs

LLMs often generate unintended or undesirable responses, such as hallucinations (Maynez et al., 2020; Ji et al., 2023; Jiang et al., 2024), biases (Sap et al., 2020; Feng et al., 2023), and harmful or unsafe outputs (Anwar et al., 2024; Ye et al., 2024; Zhang et al., 2024b). In such instances, it is appropriate for the model to abstain (Kamath et al., 2020; Feng et al., 2024b; Srinivasan et al., 2024) from generating unintended content. Abstention can be employed for unanswerable (Sulem et al., 2022; Amayuelas et al., 2024) or ambiguous (Min et al., 2020; Kim et al., 2024a) queries. Furthermore, models may abstain when relevant parametric knowledge is absent (Ahdritz et al., 2024; Kim and Thorne, 2024). Abstention can be facilitated by utilizing confidence scores of generations (Sun et al., 2022; Kuhn et al., 2023; Duan et al., 2024) or training the model for abstention capabilities (Zhang et al., 2024a; Sun et al., 2024; Cohen et al., 2024). Unlike previous approaches, this work proposes a training-free decoding method, enabling off-the-shelf models to abstain when necessary.

3 Testbed Design for Controlled Analysis

The primary objective of this work is to enable the model to dynamically adjust its behavior based on the presence and absence of its knowledge. Specifically, the model must effectively address all four scenarios depicted in Figure 1. However, as we lack prior information regarding the model's possessed knowledge, it is challenging to determine and evaluate whether the model should provide an answer

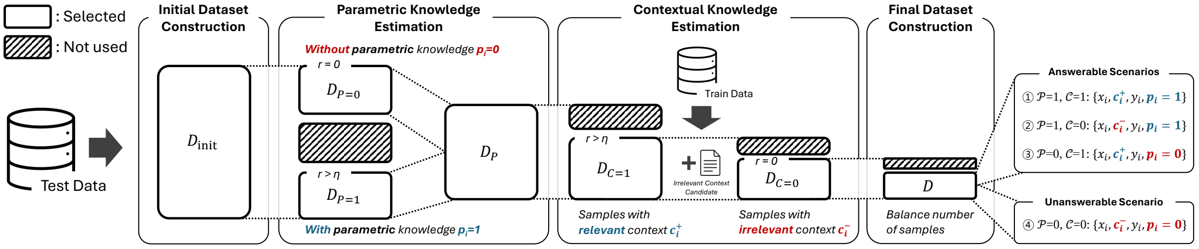

: Selected
: Not used
Test Data
Initial Dataset Construction
D_{init}
Parametric Knowledge Estimation
Without parametric knowledge p_i=0
r = 0
$D_{P=0}$
r > η
$D_{P=1}$
With parametric knowledge p_i=1
D_P
Contextual Knowledge Estimation
Train Data
r > η
$D_{C=1}$
Irrelevant Context Candidate
r = 0
$D_{C=0}$
Samples with relevant context c_i^+
Samples with irrelevant context c_i^-
Final Dataset Construction
D
Balance number of samples
Answerable Scenarios
① $\mathcal{P}$=1, $\mathcal{C}$=1: $\{x_i, c_i^+, y_i, p_i = 1\}$
② $\mathcal{P}$=1, $\mathcal{C}$=0: $\{x_i, c_i^-, y_i, p_i = 1\}$
③ $\mathcal{P}$=0, $\mathcal{C}$=1: $\{x_i, c_i^+, y_i, p_i = 0\}$
Unanswerable Scenario
④ $\mathcal{P}$=0, $\mathcal{C}$=0: $\{x_i, c_i^-, y_i, p_i = 0\}$

Figure 2: The overall process of dataset construction for the testbed.

or abstain from a given query. Thus, we construct a testbed by explicitly controlling the accessibility of the knowledge to simulate all the scenarios. This section first formulates the problem and describes the setup process as illustrated in Figure 2. Further details are in Appendix A.

3.1 Problem Formulation

This paper focuses on QA tasks, which facilitate a clear assessment of the knowledge usage of the model. *Parametric* knowledge ($\mathcal{P}$) is defined as the knowledge the model acquires during pre-training, and *contextual* knowledge ($\mathcal{C}$) refers to the external knowledge provided within the input at inference time. The knowledge is deemed *relevant* if it contains information capable of generating an accurate response to the query. For a given query x and a context c, the objective is to produce the ground-truth answer y when relevant knowledge is available or to abstain otherwise.

Figure 1 illustrates the scenarios addressed in this work. Inputs are defined *answerable* if one or more relevant knowledge are present ($\mathcal{P}$=1 or $\mathcal{C}$=1). With relevant parametric knowledge ($\mathcal{P}$=1), the model is expected to generate the correct answer regardless of c. On the other hand, the model should generate grounded on c given relevant contextual knowledge ($\mathcal{C}$=1). When no relevant knowledge is available ($\mathcal{P}$=0 and $\mathcal{C}$=0), the query is considered *unanswerable*, and the model should refuse to generate incorrect responses. Thus, a *reliable* model should properly generate an accurate answer or abstain grounded on the possessed knowledge.

3.2 Initial Dataset Construction

The testbed utilizes three extractive QA datasets from the MRQA benchmark (Fisch et al., 2019): Natural Questions (NQ) (Kwiatkowski et al., 2019), HotpotQA (Yang et al., 2018), and TriviaQA (Joshi et al., 2017). Each dataset consists of a query x_i,

Answer the following question.
<few-shot demonstrations>
Question: <question>
Answer:

(a) Parametric template $\mathcal{T}_p(\cdot)$.

Answer the following question based on the given context.
<few-shot demonstrations>
Context: <context>
Question: <question>
Answer:

(b) Contextual template $\mathcal{T}_c(\cdot)$.

Answer the following question based on the given context. If you are unsure or don't know the correct answer, simply respond with "Unknown".
<few-shot demonstrations>
Context: <context>
Question: <question>
Answer:

(c) Explicit abstention template $\mathcal{T}_a(\cdot)$.

Figure 3: List of inference templates.

an answer y_i, and a pre-defined context c_i[2] containing one or more answer spans. We split c_i into 100-word spans containing y_i to avoid excessively long contexts. Through preprocessing, we construct $\mathcal{D}_{\text{init}} = \{(x_i, c_i, y_i)\}_{i=1}^{\mathcal{N}_{\text{init}}}$.

3.3 Parametric Knowledge Estimation

To estimate the model's parametric knowledge, we assess the generation consistency (Wang et al., 2023)[3] for a query $x_i \in \mathcal{D}_{\text{init}}$. We prompt the model with the parametric template $\mathcal{T}_p(x_i)$ from Figure 3a, which relies solely on the model's parametric knowledge for the prediction. By sampling n responses, we compute the consistency rate $r = \frac{m}{n}$, where m is the number of correct responses. If $r = 0$, we assume the model lacks relevant parametric knowledge for x_i. These samples are collected as $\mathcal{D}_{\mathcal{P}=0} = \{(x_i, c_i, y_i, p_i = 0)\}_{i=1}^{\mathcal{N}_{\mathcal{P}=0}}$. On the other hand, the model is con-

[2] We assume the context is always factual and only focus on the relevance to the query. While incorporating the context's factuality is practical, we consider it orthogonal to this study.

[3] Estimating parametric knowledge is challenging (Shi et al., 2024a) due to various influencing factors. Since considering all potential factors is infeasible, we employ a fixed inference setting for all the experiments, which we contend best approximates the knowledge within a controlled setting.

sidered to pose relevant parametric knowledge if $r > \eta$ for a pre-defined threshold η, given its consistent accuracy. These samples are grouped into $\mathcal{D}_{\mathcal{P}=1} = \{(x_i, c_i, y_i, p_i = 1)\}_{i=1}^{\mathcal{N}_{\mathcal{P}=1}}$. The resulting dataset is defined as $\mathcal{D}_{\mathcal{P}} = \mathcal{D}_{\mathcal{P}=0} + \mathcal{D}_{\mathcal{P}=1}$.

3.4 Contextual Knowledge Estimation

In this stage, we select relevant and irrelevant contexts for a given query.

Relevant Context Selection We provide the model with a contextual template $\mathcal{T}_c(c_i, x_i)$ from Figure 3b, where $(x_i, c_i) \in \mathcal{D}_{\mathcal{P}}$. The model can leverage contextual knowledge by providing c_i as the input. We further verify the relevance of c_i by computing the consistency rate r. Samples with $r > \eta$ are defined as the relevant context c_i^+ and are grouped into $\mathcal{D}_{\mathcal{C}=1} = \{(x_i, c_i^+, y_i, p_i)\}_{i=1}^{\mathcal{N}_{\mathcal{C}=1}}$.

Irrelevant Context Selection For each sample $(x_i, c_i^+) \in \mathcal{D}_{\mathcal{C}=1}$, we select an irrelevant context candidate c_j^{train} from the training set. The candidate is selected with the highest SBERT (Reimers and Gurevych, 2019) embedding similarity to c_i^+ to avoid overly unrelated contexts. We then prompt the model with $\mathcal{T}_c(c_j^{\text{train}}, x_i)$ and measure the consistency rate r. Only candidates with $r = 0$ are considered irrelevant context c_i^-, ensuring that c_i^- does not provide any unintended relevant information. The resulting dataset is defined as $\mathcal{D}_{\mathcal{C}=0} = \{(x_i, c_i^+, c_i^-, y_i, p_i)\}_{i=1}^{\mathcal{N}_{\mathcal{C}=0}}$.

3.5 Final Dataset Construction

Finally, we randomly select an equal number of samples with $p_i = 0$ and $p_i = 1$, constructing $\mathcal{D} = \{(x_i, c_i^+, c_i^-, y_i, p_i)\}_{i=1}^{\mathcal{N}}$. Note that the number of selected data varies across models due to their distinct knowledge boundaries.

4 Contrastive Decoding with Abstention

Contrastive Decoding with Abstention (CDA) is a novel decoding method integrating abstention within the CCD process. This section provides a detailed description of the overall process.

4.1 Preliminary

Given a model θ at decoding step t, the parametric knowledge distribution d_t^p and the contextual knowledge distribution d_t^c is defined as follows:

$$\begin{aligned} d_t^p &= \text{logit}_\theta(y_t \mid \mathcal{T}_p(x, y_{<t})), \\ d_t^c &= \text{logit}_\theta(y_t \mid \mathcal{T}_c(c, x, y_{<t})) \end{aligned} \tag{1}$$

where $y_{<t}$ are previously generated tokens. CCD measures the final output distribution d_t^o as an ensemble of d_t^p and d_t^c as:

$$d_t^o = d_t^p + w_t^c \, (d_t^c - d_t^p) \tag{2}$$

The weight w_t^c should precisely quantify the relevance of c, ensuring a higher weight when c is deemed relevant.

4.2 Incorporating Abstention

To enable CDA to properly abstain, we incorporate the abstention distribution d_t^a computed from an explicit abstention instruction $\mathcal{T}_a(\cdot)$ in Figure 3c.

$$d_t^a = \text{logit}_\theta(y_t \mid \mathcal{T}_a(c, x, y_{<t})) \tag{3}$$

CDA expands Eq. 2 by applying d_t^a for the final output distribution, where the weight for d_t^a is defined as $w_t^a = 1 - w_t^p - w_t^c$.

$$\begin{aligned} d_t^o &= d_t^p + w_t^c \, (d_t^c - d_t^p) + w_t^a \, (d_t^a - d_t^p) \\ &= (1 - w_t^c - w_t^a) \, d_t^p + w_t^c \, d_t^c + w_t^a \, d_t^a \\ &= w_t^p \, d_t^p + w_t^c \, d_t^c + (1 - w_t^p - w_t^c) \, d_t^a \end{aligned} \tag{4}$$

Intuitively, w_t^a decreases when the model is confident of a possessed knowledge, whereas it increases when both knowledge are uncertain.

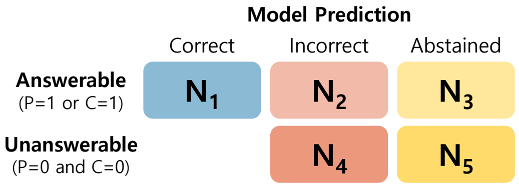
4.3 Knowledge Relevance Assessment via Uncertainty Calibration

A key requirement for CDA is that the weights w_t^p and w_t^c should effectively quantify the relevance of the corresponding knowledge. We assess the relevance as the uncertainty of the corresponding knowledge regarding x. Specifically, we utilize the entropy (Malinin and Gales, 2021; Abdar et al., 2021), a widely used measure to assess the uncertainty of the knowledge for a query (Kuhn et al., 2023; Duan et al., 2024), particularly prevalent in CCD (Kim et al., 2024b; Qiu et al., 2024). For an output distribution d, the entropy $\mathcal{H}$ is defined as:

$$\mathcal{H} = -\sum_{i=1}^{|\mathcal{V}|} d_i \log d_i, \tag{5}$$

where d_i is the i^{th} token of the vocabulary $\mathcal{V}$. The parametric uncertainty $\mathcal{H}_t^p$ and contextual uncertainty $\mathcal{H}_t^c$ are derived from their respective distributions d_t^p and d_t^c.

Nonetheless, directly comparing $\mathcal{H}_t^p$ and $\mathcal{H}_t^c$ are imprecise since they are conditioned on distinct inputs and possibly miscalibrated. To address this,



	Correct	Incorrect	Abstained
Answerable (P=1 or C=1)	N_1	N_2	N_3
Unanswerable (P=0 and C=0)		N_4	N_5

Figure 4: Illustration of all possible results. The model should generate correct answers for answerable queries (N_1) and abstain for unanswerable queries (N_5). Any other responses (N_2, N_3, N_4) are classified as incorrect.

we "calibrate" (Zhao et al., 2021; Holtzman et al., 2021; He et al., 2024) the uncertainty measures by accounting for the model's intrinsic bias. Specifically, we estimate the bias with a "content-free" `null` prompt, replacing specific inputs x_i and c_i with placeholder tokens $\bar{x}$ and $\bar{c}$ to remove any specific semantic information. Applying the templates $\mathcal{T}_p(\bar{x})$ and $\mathcal{T}_c(\bar{c}, \bar{x})$ yields the parametric `null` distribution $\bar{d}_t^p$ and the contextual `null` distribution $\bar{d}_t^c$, along with their corresponding entropy values $\bar{\mathcal{H}}_t^p$ and $\bar{\mathcal{H}}_t^c$. The confidence for the knowledge is quantified as the additional information provided by the input relative to the `null` prompt.

$$r_t^p = \frac{\max(\mathcal{H}_t^p - \bar{\mathcal{H}}_t^p,\, 0)}{\bar{\mathcal{H}}_t^p},\ r_t^c = \frac{\max(\mathcal{H}_t^c - \bar{\mathcal{H}}_t^c,\, 0)}{\bar{\mathcal{H}}_t^c}. \quad (6)$$

We obtain the final weights w_t^p and w_t^c from Eq. 4 by normalizing r_t^p and r_t^c, respectively.

$$w_t^p = \frac{r_t^p}{r_t^p + r_t^c}\, r_t^p,\ w_t^c = \frac{r_t^c}{r_t^p + r_t^c}\, r_t^c \quad (7)$$

When a particular knowledge source provides substantial additional information relative to the `null` prompt, it is assigned a higher weight, whereas the absence of knowledge results in a reduced weight.

4.4 CDA with Momentum (CDA-M)

At each decoding step, previous content may unintentionally steer the model toward irrelevant knowledge. To mitigate this, we apply momentum, updating the current weight w_t as a convex combination with the previous weight w_{t-1}.

$$w_t \leftarrow \alpha\, w_{t-1} + (1 - \alpha)\, w_t \quad (8)$$

Here, the hyperparameter α controls the influence of the previous step on the current step. Applying momentum helps stabilize the decoding process by smoothing abrupt weight changes.

5 Experiments

5.1 Experimental Setting

The experiments utilize the testbed from Section 3 and four instruction-tuned models including LLAMA3 8B INSTRUCT (Dubey et al., 2024), LLAMA2 7B & 13B CHAT (Touvron et al., 2023), and MISTRAL 7B INSTRUCT (Jiang et al., 2023). Results are averaged over three different random seeds. Further details are stated in Appendix B.

5.2 Evaluation Metric

To measure the overall performance across two distinct tasks, we adopt three metrics, each reflecting unique aspects of performance, based on the five possible results in Figure 4.

Answerable Prediction F1 ($\mathbf{F1_{ans}}$) For answerable queries, we compute $\mathrm{F1_{ans}}$ (Kim et al., 2024a) as the harmonic mean of precision ($\frac{N_1}{N_1+N_2+N_4}$) and recall ($\frac{N_1}{N_1+N_2+N_3}$). The prediction is considered correct if it contains the ground-truth answer (Mallen et al., 2023; Schick et al., 2023).

Abstention F1 ($\mathbf{F1_{abs}}$) The model should abstain from incorrect responses for unanswerable queries while minimizing over-abstention. $\mathrm{F1_{abs}}$ (Kim et al., 2024a) measures such behaviors by incorporating both precision ($\frac{N_5}{N_3+N_5}$) and recall ($\frac{N_5}{N_4+N_5}$). A prediction is deemed as an abstention if it contains any pre-defined abstention phrases (Amayuelas et al., 2024; Kim et al., 2024a).

Reliability Score (RS) RS (Xu et al., 2024) is the weighted sum of accuracy (Acc., $\frac{N_1}{N}$) and coverage (Cov., $\frac{N_1+N_3+N_5}{N}$), where N is the total number of samples and α is set as the answer rate ($1 - \frac{N_3+N_5}{N}$).

$$\mathrm{RS}(\alpha) = \alpha \times \mathrm{Cov.} + (1 - \alpha) \times \mathrm{Acc.} \quad (9)$$

RS prioritizes accuracy at a lower answer rate while avoiding errors with a higher coverage at a high answer rate. We also report the accuracy and coverage for a more thorough analysis.

5.3 Baselines

To evaluate the effectiveness of our approach, we compare different inference methods as baselines.

Direct Prompting includes **contextual prompting (CONTEXT)** employing $\mathcal{T}_c(\cdot)$ and **abstention prompting (ABSTAIN)** utilizing $\mathcal{T}_a(\cdot)$, with an explicit instruction for abstention.

Backbone	LLAMA3 8B INSTRUCT					LLAMA2 7B CHAT					LLAMA2 13B CHAT					MISTRAL 7B INSTRUCT				
Method	$F1_{ans}$	$F1_{abs}$	RS	Acc.	Cov.	$F1_{ans}$	$F1_{abs}$	RS	Acc.	Cov.	$F1_{ans}$	$F1_{abs}$	RS	Acc.	Cov.	$F1_{ans}$	$F1_{abs}$	RS	Acc.	Cov.
										NQ										
CONTEXT	57.26	2.18	50.22	49.95	50.23	57.13	0.09	49.98	49.97	49.98	57.14	0.37	50.00	49.95	50.00	57.04	0.25	49.88	49.85	49.88
CAD	55.34	2.14	48.56	48.29	48.56	54.78	0.17	47.94	47.92	47.94	45.57	0.48	39.90	39.84	39.90	55.03	1.15	48.21	48.07	48.21
ACD	71.36	0.92	62.50	**62.39**	62.50	64.55	0.15	56.48	**56.46**	56.48	66.66	0.18	58.33	**58.31**	58.33	68.48	0.00	59.93	**59.93**	59.93
ABSTAIN	60.22	52.05	48.17	38.36	57.27	38.73	45.16	26.73	20.54	42.11	50.49	46.62	37.83	29.70	48.93	61.86	53.29	52.85	42.83	59.70
SELF-ASK	57.23	48.11	43.91	35.06	53.70	56.81	10.95	50.17	48.66	50.23	59.37	20.80	52.98	49.89	53.21	62.52	49.03	57.49	48.76	59.54
ENTROPY	64.06	53.34	55.53	45.15	60.90	58.00	41.09	51.98	44.47	54.08	59.23	42.61	52.75	44.85	55.19	61.91	56.29	55.48	44.63	60.33
ACD-A	63.56	52.46	53.99	43.82	60.11	48.82	39.52	37.55	30.41	45.66	57.96	46.34	49.28	40.27	54.43	61.46	54.02	55.80	45.58	59.56
FSB	69.27	54.94	59.64	49.02	65.09	55.04	47.26	43.26	34.47	52.20	62.44	47.60	53.66	44.46	58.19	66.71	55.51	58.95	48.32	63.65
CDA	72.06	**55.49**	62.95	52.28	67.51	66.86	47.52	59.86	51.22	62.38	68.62	47.14	61.63	53.16	63.76	69.68	56.47	61.45	50.61	66.09
CDA-M	**73.15**	55.47	**63.72**	53.16	**68.30**	**69.99**	**47.60**	**62.28**	53.62	**64.81**	**70.66**	**48.12**	**63.18**	54.46	**65.48**	**71.00**	**56.46**	**62.30**	51.45	**67.03**
										HotpotQA										
CONTEXT	57.15	1.19	50.08	49.93	50.08	57.14	0.00	49.98	49.98	49.98	57.16	0.10	50.00	49.98	50.00	56.99	0.09	49.88	49.87	49.88
CAD	55.78	1.88	48.89	51.99	48.89	54.68	0.04	47.84	47.84	47.84	54.60	0.07	47.76	47.75	47.76	53.74	0.20	47.04	47.01	47.04
ACD	74.36	0.57	65.09	**65.02**	65.09	69.27	0.00	60.60	60.60	60.60	69.80	0.07	61.15	**61.14**	61.15	72.72	0.09	63.64	**63.62**	63.64
ABSTAIN	66.88	56.58	56.23	45.21	63.55	47.29	48.65	33.49	26.18	48.18	57.17	51.95	43.02	33.89	55.10	61.12	54.66	53.06	42.55	59.23
SELF-ASK	50.64	48.63	33.86	26.80	49.70	58.58	17.35	52.19	49.70	52.33	58.22	14.36	51.87	49.89	51.95	61.33	43.36	56.36	48.95	57.70
ENTROPY	67.08	56.44	57.92	46.94	63.88	59.00	45.11	52.64	44.16	55.51	59.24	45.47	53.01	44.49	55.81	63.31	**60.07**	57.19	45.65	62.40
ACD-A	65.88	54.67	57.41	46.80	62.63	57.79	51.11	47.90	38.19	55.61	62.00	52.13	54.03	43.95	59.16	61.02	50.93	57.01	48.10	58.95
FSB	74.89	58.51	66.21	55.05	70.55	65.63	**53.82**	55.73	45.37	62.04	68.68	54.16	60.41	50.05	64.79	74.32	52.18	68.20	59.27	69.94
CDA	78.71	62.50	70.20	58.36	74.52	73.39	42.41	66.96	60.15	67.82	**73.69**	56.81	68.98	59.53	**70.44**	76.28	55.84	69.43	59.50	71.83
CDA-M	**79.32**	**62.59**	**70.64**	58.78	**74.99**	**74.09**	42.31	**67.50**	**60.70**	**68.37**	73.66	**56.89**	**68.92**	59.42	70.42	**76.94**	56.67	**69.98**	59.85	**72.49**
										TriviaQA										
CONTEXT	57.29	2.24	50.23	49.95	50.23	57.17	0.26	50.02	49.99	50.02	57.15	0.47	50.03	49.97	50.03	57.13	0.35	50.00	49.96	50.00
CAD	55.83	0.65	48.87	48.79	48.87	55.11	0.25	48.22	48.19	48.22	55.81	0.32	48.84	48.80	48.84	54.71	0.57	47.89	47.82	47.89
ACD	76.79	3.09	67.38	**66.99**	67.39	72.86	0.18	63.74	**63.72**	63.74	72.49	0.42	63.43	**63.38**	63.43	75.01	0.05	65.62	**65.62**	65.62
ABSTAIN	67.46	57.31	56.21	45.07	64.10	59.19	50.45	47.04	37.87	56.27	48.53	46.74	35.11	27.52	47.88	60.73	53.16	51.25	41.22	58.42
SELF-ASK	52.40	48.38	36.73	29.01	50.61	57.93	8.97	51.05	49.84	51.09	58.09	12.07	51.53	49.85	51.59	62.18	48.60	57.19	48.55	59.20
ENTROPY	66.17	57.26	56.50	45.33	63.35	60.21	47.84	53.99	45.00	57.10	60.21	48.31	54.31	45.32	57.27	62.21	56.59	56.23	45.48	60.71
ACD-A	66.67	56.73	58.81	57.88	63.87	61.21	50.52	53.09	43.46	58.18	58.42	49.72	49.22	39.54	55.61	61.62	51.62	56.81	47.48	59.48
FSB	77.02	59.84	68.55	57.24	72.62	69.50	51.88	60.41	50.59	64.78	66.21	52.08	56.12	45.98	61.91	**77.67**	47.53	**70.69**	62.72	72.07
CDA	80.39	**65.67**	72.35	60.01	76.67	**73.70**	51.29	**67.29**	58.43	**69.06**	71.08	51.44	64.08	54.78	66.55	75.76	56.35	67.57	57.21	71.21
CDA-M	**80.93**	65.66	**72.74**	60.40	**77.07**	73.47	**52.10**	67.11	58.04	69.00	**73.12**	**53.46**	**65.82**	56.09	**68.52**	76.95	**57.06**	68.38	57.84	**72.21**

Table 1: Experimental results on three different datasets. For each dataset, the **best method** is highlighted in bold, and the second-best method is underlined. CDA(-M) outperforms all the baselines across different metrics.

SELF-ASK prompts the model with $\mathcal{T}_c(\cdot)$ and further verifies the generation (Kadavath et al., 2022). Predictions verified as "unknown" are abstained.

Context-aware Decoding (CAD) amplifies contextual influence by gauging $d_t^o = d_t^c + w_t^c\,(d_t^c - d_t^p)$ with a fixed w_t^c during decoding (Shi et al., 2024b).

ENTROPY measures the entropy (Eq. 5) of the generated tokens when prompted with $\mathcal{T}_c(\cdot)$ (Huang et al., 2025). We measure four different variants — first-token, average, maximum, and minimum entropy — and report the first-token entropy with the best performance. If the measure exceeds a pre-defined threshold, the prediction is deemed uncertain and thus abstained.

Adaptive Contrastive Decoding (ACD) follows Eq. 2 where $w_t^c = 1 - \frac{\mathcal{H}_t^c}{\mathcal{H}_t^p + \mathcal{H}_t^c}$ (Kim et al., 2024b).

ACD with Abstention (ACD-A) expands ACD to perform abstention where $w_t^c = 1 - \frac{\mathcal{H}_t^c}{\mathcal{H}_t^p + \mathcal{H}_t^c + \mathcal{H}_t^a}$ and $w_t^a = 1 - \frac{\mathcal{H}_t^a}{\mathcal{H}_t^p + \mathcal{H}_t^c + \mathcal{H}_t^a}$ following Eq. 4.

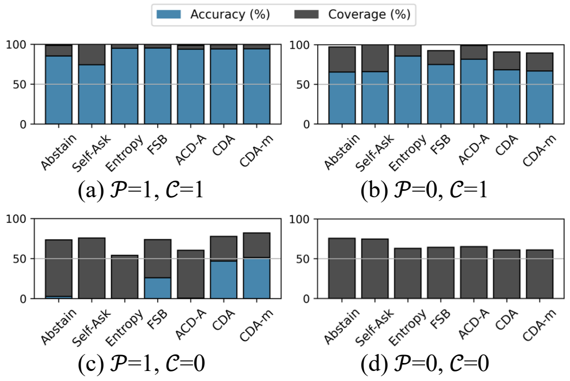
First Step Branching (FSB) compares the first-token entropy $\mathcal{H}_1^p$, $\mathcal{H}_1^c$, and $\mathcal{H}_1^a$ when prompted with $\mathcal{T}_p(\cdot)$, $\mathcal{T}_c(\cdot)$, and $\mathcal{T}_a(\cdot)$, respectively. We select the most certain method and continue the generation with the selected method.

5.4 Main Results

The main results are presented in Table 1.

Methods not accounting for abstention fail to handle unanswerable queries. Methods such as CONTEXT, CAD, and ACD exhibit near-zero $F1_{abs}$, indicating their inability to handle unanswerable queries. Moreover, the negligible gap between their accuracy and coverage further confirms their incapability to abstain.

Incorporating abstention enhances the handling of unanswerable queries. ABSTAIN and SELF-ASK exhibit biased abstentions, resulting in low accuracy and high coverage. ACD-A and ENTROPY perform abstention to some extent, but they struggle to balance between accurate generation and abstention. FSB emerges as the strongest among the baseline, effectively addressing (un)answerable queries. Overall, incorporating abstention does provide the model with the ability to abstain, but the baselines fail to effectively balance the trade-off



(a) $\mathcal{P}$=1, $\mathcal{C}$=1

(b) $\mathcal{P}$=0, $\mathcal{C}$=1

(c) $\mathcal{P}$=1, $\mathcal{C}$=0

(d) $\mathcal{P}$=0, $\mathcal{C}$=0

Figure 5: The accuracy and coverage for all the scenarios regarding both knowledge. CDA(-M) effectively balances between correct predictions and abstentions, especially in the presence of irrelevant contexts ($\mathcal{C}$=0).

between accurate generation and appropriate abstention.

CDA(-M) exhibits superior performance across all datasets. CDA(-M) outperforms all the baselines on F1$_{\text{abs}}$ and RS, properly abstaining unanswerable queries. Moreover, its effective handling of answerable queries exhibits the highest F1$_{\text{ans}}$ and the second-best accuracy following ACD.

6 Ablation Study

This section presents ablation studies of CDA(-M). Unless otherwise specified, all experiments are conducted on LLAMA3 8B INSTRUCT with the testbed from Section 3. Methods capable of abstention, including ABSTAIN, SELF-ASK, ENTROPY, FSB, and ACD-A, are utilized for comparison. Further details and results are in Appendix C.

6.1 Analysis of Different Scenarios

Figure 5 depicts the accuracy (in blue) and coverage (in gray) for each scenario in the NQ dataset. Note that the main objective is to balance between accurate generation when relevant knowledge is present ($\mathcal{P}$=1 or $\mathcal{C}$=1) and abstention when such knowledge is absent ($\mathcal{P}$=0 and $\mathcal{C}$=0). Most baselines exhibit over-abstention, particularly with irrelevant context. These methods fail to utilize relevant parametric knowledge, resulting in either biased abstention or incorrect generation. In contrast, CDA(-M) robustly leverages proper knowledge while maintaining balanced abstention capability.

6.2 Ablation on Momentum Weight

In this section, we investigate the impact of momentum weight α on CDA-M. Figure 6 displays F1$_{\text{ans}}$ and F1$_{\text{abs}}$ of α from 0.0 to 1.0, along with the best-performing baselines in the NQ dataset. We can observe that F1$_{\text{abs}}$ remains stable while F1$_{\text{ans}}$ improves when applying momentum. As intended, momentum reduces hallucinations while preserving abstention capabilities. Overall, CDA-M outperforms the strongest baselines regardless of α.

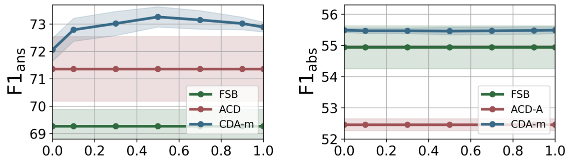


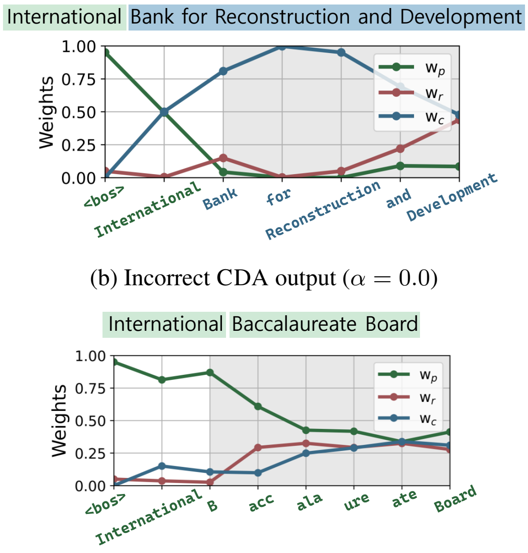
Figure 6: F1$_{\text{ans}}$ and F1$_{\text{abs}}$ according to different α values. Applying momentum significantly improves F1$_{\text{ans}}$.

<Query>
What is the full form of IB board

<Irrelevant Context>
The **International Bank for Reconstruction and Development** (IBRD) is an international financial institution that offers loans to ...

<Ground-truth Answer>
International Baccalaureate

(a) Input query, irrelevant context, and ground-truth answer.



(b) Incorrect CDA output ($\alpha = 0.0$)

(c) Correct CDA-M output ($\alpha = 0.7$)

Figure 7: Example generation of CDA(-M) with an irrelevant context. Irrelevant context contains phrases that are similar to the ground-truth answer, making it easier for the model to hallucinate. The weights of CDA shift to the irrelevant context, resulting in incorrect generation. On the other hand, the momentum applied to CDA-M mitigates abrupt shifts in attention, resulting in correct generation.

We further conduct case studies of how this find-

Dataset	Method	F1$_{\text{ans}}$	F1$_{\text{abs}}$	RS	Acc.	Cov.
NQ	CDA	**72.06**	**55.49**	**62.95**	**52.28**	**67.51**
	w/o calibration	59.35	52.06	47.89	38.04	56.79
HotpotQA	CDA	**78.71**	**62.50**	**70.20**	**58.36**	**74.52**
	w/o calibration	66.94	56.39	56.42	45.43	63.55
TriviaQA	CDA	**80.39**	**65.66**	**72.35**	**60.01**	**76.67**
	w/o calibration	67.56	57.24	56.46	45.32	64.17

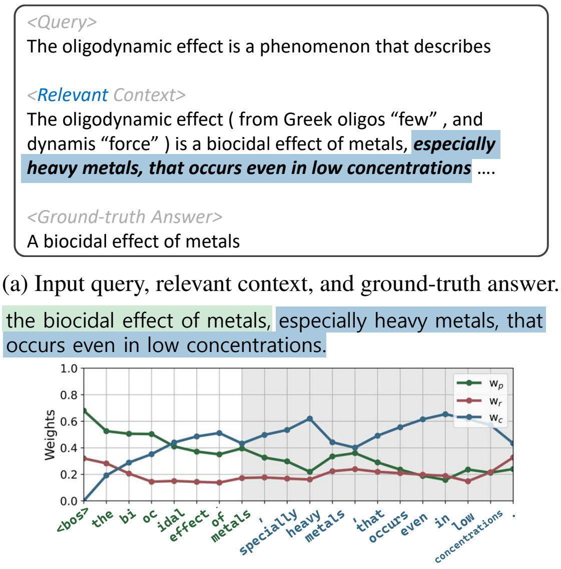
Table 2: The effects of applying calibration. We can observe significant degradation without calibration.

ing relates to the momentum weights. Figure 7a demonstrates an example of a question and an irrelevant context. Figure 7b displays the generation result of CDA and the weights measured for the knowledge at every decoding step. CDA initially assigns more weight to relevant parametric knowledge, generating the correct span up to "International Bank". However, the model's attention shifts toward contextual knowledge, incorporating incorrect information from the context span "International Bank for Reconstruction and Development." We presume that the phrase "International Bank" from the irrelevant context causes this shift in weight. In contrast, Figure 7c demonstrates that CDA-M, leveraging momentum, mitigates abrupt shifts in attention toward irrelevant knowledge, resulting in accurate generation.

Notably, CDA-M does not persistently focus on a single knowledge source. A case where CDA-M fully utilizes both knowledge appropriately is shown in Figure 8. Figure 8a displays a case where a relevant context is provided to the model. We can observe from Figure 8b that CDA-M initially focuses on the relevant parametric knowledge, and over time, it transitions to incorporate contextual knowledge, producing richer and more nuanced answers.

6.3 Effect of Calibration

CDA(-M) leverages calibrated uncertainty measures to quantify the relevance of different knowledge. To evaluate the effect of calibration, we modify Eq. 4 to employ non-calibrated measure by directly setting $r_t^p = \mathcal{H}_t^p$ and $r_t^c = \mathcal{H}_t^c$. As shown in Table 2, this change significantly degrades the performance across all metrics by up to 14 points. Directly utilizing uncalibrated measures yields suboptimal results, thus highlighting the importance of the additional calibration step in CDA(-M).



<Query>
The oligodynamic effect is a phenomenon that describes

<Relevant Context>
The oligodynamic effect (from Greek oligos "few" , and dynamis "force") is a biocidal effect of metals, ***especially heavy metals, that occurs even in low concentrations*** ....

<Ground-truth Answer>
A biocidal effect of metals

(a) Input query, relevant context, and ground-truth answer.

the biocidal effect of metals, especially heavy metals, that occurs even in low concentrations.

(b) Correct CDA-M output ($\alpha = 0.7$)

Figure 8: Example generation of CDA-M for a relevant context. CDA-M initially focuses on the parametric knowledge, and the attention shifts to incorporate contextual knowledge, generating richer output.

6.4 Comparison with Training-based Methods

Training models to abstain has demonstrated notable performance. Hence, we compare CDA-M with instruction-tuning (Ouyang et al., 2022; Yang et al., 2024), which explicitly trains the model to abstain when necessary.

Experimental Setting The training data are labeled according to parametric and contextual knowledge. Following the procedure described in Section 3, we verify whether the model possesses relevant knowledge for each training sample. Samples with relevant knowledge are labeled with the ground-truth answer y, while samples without any relevant knowledge are labeled with a pre-defined abstention response y_{abs} (e.g., "unknown"). The model is then trained to generate the label given $\mathcal{T}_c(c, x)$. For evaluation, we utilize the instruction-tuned model to generate an output given $\mathcal{T}_c(c, x)$.

Experimental Results Table 3 presents the in-domain (IND) and out-of-domain (OOD) results of instruction-tuning. CDA-M consistently outperforms instruction-tuning, even in IND scenarios. Furthermore, while training often tailors the model to specific domains, resulting in significant performance drops in OOD settings, CDA-M demonstrates superior generalization capabilities. This robustness makes CDA(-M) a more applicable solution for practical scenarios.

Target (→)	NQ			HotpotQA			TriviaQA		
Source (↓)	F1$_{\text{ans}}$	F1$_{\text{abs}}$	RS	F1$_{\text{ans}}$	F1$_{\text{abs}}$	RS	F1$_{\text{ans}}$	F1$_{\text{abs}}$	RS
NQ	66.37	43.87	61.12	64.43	48.79	59.65	67.31	49.00	62.14
HotpotQA	64.86	42.73	59.80	74.06	57.57	68.76	65.26	50.69	61.17
TriviaQA	65.15	46.37	57.70	66.68	49.82	60.58	67.60	53.68	61.84
CDA-M	**73.15**	**55.47**	**63.72**	**79.32**	**62.59**	**70.64**	**80.93**	**65.66**	**72.74**

Table 3: The results of instruction-tuning. In-domain results, the **best results**, and the second-based results are highlighted. CDA-M displays superior performance across all the scenarios.

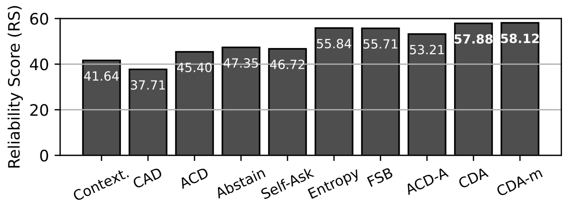


Figure 9: Average Reliability Score (RS) in RAG settings. CDA(-M) outperforms all the baselines.

6.5 Evaluation on RAG Setting

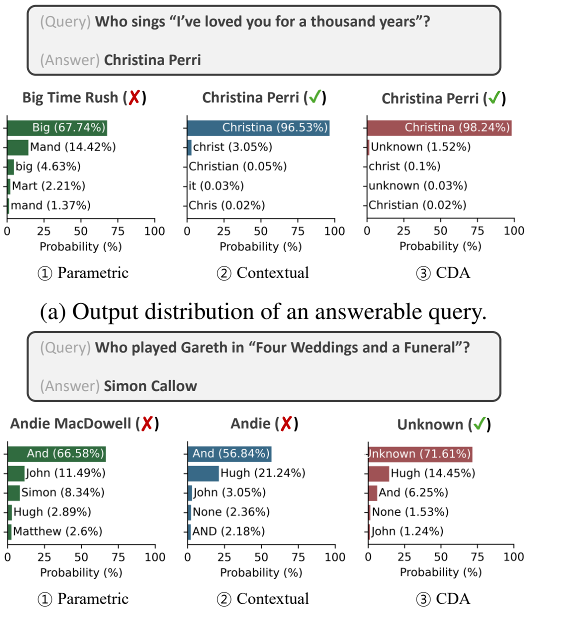
To evaluate CDA in practical, real-world scenarios, we conduct experiments within the RAG setting.

Experimental Setting We utilize CONTRIEVER-MSMARCO (Izacard et al., 2022) as a retriever, and the top-1 context is retrieved from the Wikipedia contexts.[4] Unlike the controlled setting, where the presence of both knowledge is precisely estimated, the prior knowledge of the given query is unknown in the RAG setting. Since it is difficult to determine the answerablility of the given query, we evaluate solely based on the Reliability Score (RS).

Experimental Results Figure 9 displays the average results in the RAG setting. Similar to the main experiments, methods without abstention capabilities demonstrate poor performance, while FSB and ENTROPY emerge as strong baselines. Overall, CDA(-M) outperform all the baselines, highlighting the effectiveness in the practical RAG setting.

6.6 Output Distribution Analysis

This section analyzes how the output distribution shifts from parametric and contextual distributions to the final distribution of CDA. Figure 10 displays the top-5 softmax probabilities along with their corresponding tokens for d_1^p, d_1^c, and d_1^o from Eq. 4. Figure 10a illustrates an output distribution for an answerable query. CDA successfully attends to the relevant contextual knowledge "Christina" and amplifies the probability from 96.53% to 98.24%, effectively mitigating the influence of the parametric knowledge. Furthermore, Figure 10b illustrates how CDA handles unanswerable queries. While both knowledge erroneously generates the token "Andie", CDA successfully shifts the distribution towards abstention, preventing hallucinations.



(a) Output distribution of an answerable query.

(b) Output distribution of an unanswerable query.

Figure 10: Top-5 probabilities and their corresponding tokens for parametric, contextual, and CDA distribution. CDA (a) amplifies the relevant knowledge for answerable queries while (b) shifting the distribution to abstain from unanswerable queries.

7 Conclusion

This work addresses the challenge of generating reliable responses leveraging parametric and contextual knowledge when available, while abstaining when both are absent. To evaluate these scenarios, we construct a testbed based on the model's approximated knowledge. Furthermore, we present **Contrastive Decoding with Abstention (CDA)**, a novel, training-free decoding method that incorporates abstention in the generation process. CDA quantifies the relevance of both knowledge and dynamically attends to the relevant one. When no relevant knowledge is available, CDA guides the model to abstain. Through extensive experiments, CDA exhibits accurate generation when relevant knowledge is available and abstention otherwise, reducing the risks of hallucination.

[4] Wikipedia dump from Dec. 2018.

Limitations

Our study acknowledges a few limitations that present opportunities for future research.

Computation Cost Contrastive decoding inherently involves comparing multiple outputs, inevitably increasing the overall cost. CDA also requires additional computations, costing roughly double that of greedy decoding. However, CDA enables reliable generation through abstention, a capability that is enhanced through this additional computation. We believe that the increased cost is a reasonable trade-off for achieving a more dependable and safe model. Nonetheless, reducing the cost is an essential factor, which will be addressed as a primary objective in future work. A detailed analysis of the inference cost is provided in Appendix D.

Limitations in Task Scope Our study primarily focuses on single-context scenarios, providing a relatively clear distinction between the presence and absence of knowledge, facilitating precise analysis. However, extending the scope to multi-context scenarios would be an important direction for future work. Additionally, this work focuses on short-form QA tasks, which are relatively easier to assess the knowledge usage. However, expanding the task to reasoning-intensive, long-form generation tasks would be a meaningful advancement.

Advanced Abstention The current work primarily focuses on the model's ability to simply express abstention, which is often lacking in user-friendliness. Future research could explore incorporating reasoning capabilities to explain the rationale behind abstention decisions.

Acknowledgement

This work was partly supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University), No.RS-2020-II201373, Artificial Intelligence Graduate School Program (Hanyang University), NO.RS-2021-II212068, Artificial Intelligence Innovation Hub (Artificial Intelligence Institute, Seoul National University)] and the Mid-Career Bridging Program through Seoul National University.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Gustaf Ahdritz, Tian Qin, Nikhil Vyas, Boaz Barak, and Benjamin L Edelman. 2024. Distinguishing the knowable from the unknowable with language models. *arXiv preprint arXiv:2402.03563*.

Alfonso Amayuelas, Kyle Wong, Liangming Pan, Wenhu Chen, and William Yang Wang. 2024. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6416–6432, Bangkok, Thailand. Association for Computational Linguistics.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric J Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Chenyu Zhang, Ruiqi Zhong, Sean O hEigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schroeder de Witt, Sumeet Ramesh Motwani, Yoshua Bengio, Danqi Chen, Philip Torr, Samuel Albanie, Tegan Maharaj, Jakob Nicolaus Foerster, Florian Tramèr, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. 2024. Foundational challenges in assuring alignment and safety of large language models. *Transactions on Machine Learning Research*. Survey Certification, Expert Certification.

Stefan Buttcher, Charles LA Clarke, and Gordon V Cormack. 2016. *Information retrieval: Implementing and evaluating search engines*. Mit Press.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Roi Cohen, Konstantin Dobler, Eden Biran, and Gerard de Melo. 2024. I don't know: Explicit modeling of uncertainty with an [IDK] token. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Roi Cohen, Mor Geva, Jonathan Berant, and Amir Globerson. 2023. Crawling the internal knowledge-base of language models. *arXiv preprint arXiv:2301.12810*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand. Association for Computational Linguistics.

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*.

Romina Etezadi and Mehrnoush Shamsfard. 2023. The state of the art in open domain complex question answering: a survey. *Applied Intelligence*, 53(4):4124–4144.

Ruitao Feng, Xudong Hong, Mayank Jobanputra, Mattes Warning, and Vera Demberg. 2024a. Retrieval-augmented modular prompt tuning for low-resource data-to-text generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14053–14062, Torino, Italia. ELRA and ICCL.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024b. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China. Association for Computational Linguistics.

Kang He, Yinghan Long, and Kaushik Roy. 2024. Prompt-based bias calibration for better zero/few-shot learning of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12673–12691, Miami, Florida, USA. Association for Computational Linguistics.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2025. Look before you leap: An exploratory study of uncertainty analysis for large language models. *IEEE Transactions on Software Engineering*, pages 1–18.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Preprint*, arXiv:2112.09118.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

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*.

Che Jiang, Biqing Qi, Xiangyu Hong, Dayuan Fu, Yang Cheng, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. 2024. On large language models' hallucination with regard to known facts. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1041–1053, Mexico City, Mexico. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. 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, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick 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)*, pages 6769–6781, Online. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, yoichi takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime QA: What's the answer right now? In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Hyuhng Joon Kim, Youna Kim, Cheonbok Park, Junyeob Kim, Choonghyun Park, Kang Min Yoo, Sang-goo Lee, and Taeuk Kim. 2024a. Aligning language models to explicitly handle ambiguity. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1989–2007, Miami, Florida, USA. Association for Computational Linguistics.

Minsu Kim and James Thorne. 2024. Epistemology of language models: Do language models have holistic knowledge? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12644–12669, Bangkok, Thailand. Association for Computational Linguistics.

Youna Kim, Hyuhng Joon Kim, Cheonbok Park, Choonghyun Park, Hyunsoo Cho, Junyeob Kim, Kang Min Yoo, Sang-goo Lee, and Taeuk Kim. 2024b. Adaptive contrastive decoding in retrieval-augmented generation for handling noisy contexts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2421–2431, Miami, Florida, USA. Association for Computational Linguistics.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

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*, 7:452–466.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liška, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2024. Mind the gap: assessing temporal generalization in neural language models. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. 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*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

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)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

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

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Zexuan Qiu, Zijing Ou, Bin Wu, Jingjing Li, Aiwei Liu, and Irwin King. 2024. Entropy-based decoding for retrieval-augmented large language models. *arXiv preprint arXiv:2406.17519*.

Mahimai Raja, E Yuvaraajan, et al. 2024. A rag-based medical assistant especially for infectious diseases. In *2024 International Conference on Inventive Computation Technologies (ICICT)*, pages 1128–1133. IEEE.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pages 68539–68551. Curran Associates, Inc.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024a. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*.

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

Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Chandu. 2024. Selective "selective prediction": Reducing unnecessary abstention in vision-language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12935–12948, Bangkok, Thailand. Association for Computational Linguistics.

Elior Sulem, Jamaal Hay, and Dan Roth. 2022. Yes, no or IDK: The challenge of unanswerable yes/no questions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1075–1085, Seattle, United States. Association for Computational Linguistics.

Meiqi Sun, Wilson Yan, Pieter Abbeel, and Igor Mordatch. 2022. Quantifying uncertainty in foundation models via ensembles. In *NeurIPS 2022 Workshop on Robustness in Sequence Modeling*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110, Bangkok, Thailand. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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*.

Neeraj Varshney, Pavel Dolin, Agastya Seth, and Chitta Baral. 2024. The art of defending: A systematic evaluation and analysis of LLM defense strategies on safety and over-defensiveness. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13111–13128, Bangkok, Thailand. Association for Computational Linguistics.

Jonas Waldendorf, Barry Haddow, and Alexandra Birch. 2024. Contrastive decoding reduces hallucinations in large multilingual machine translation models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2539, St. Julian's, Malta. Association for Computational Linguistics.

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

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2024. Know your limits: A survey of abstention in large language models. *Preprint*, arXiv:2407.18418.

Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. 2024. Rejection improves reliability: Training LLMs to refuse unknown questions using RL from knowledge feedback. In *First Conference on Language Modeling*.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024. Alignment for honesty. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Junjie Ye, Sixian Li, Guanyu Li, Caishuang Huang, Songyang Gao, Yilong Wu, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. ToolSword: Unveiling safety issues of large language models in tool learning across three stages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2181–2211, Bangkok, Thailand. Association for Computational Linguistics.

Dawei Yin, Yuening Hu, Jiliang Tang, Tim Daly, Mianwei Zhou, Hua Ouyang, Jianhui Chen, Changsung Kang, Hongbo Deng, Chikashi Nobata, Jean-Marc Langlois, and Yi Chang. 2016. Ranking relevance in yahoo search. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 323–332, New York, NY, USA. Association for Computing Machinery.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024a. R-tuning: Instructing large language models to say 'I don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.

Lingxi Zhang, Jing Zhang, Xirui Ke, Haoyang Li, Xinmei Huang, Zhonghui Shao, Shulin Cao, and Xin Lv. 2023. A survey on complex factual question answering. *AI Open*, 4:1–12.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2024b. SafetyBench: Evaluating the safety of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15537–15553, Bangkok, Thailand. Association for Computational Linguistics.

Bowen Zhao, Zander Brumbaugh, Yizhong Wang, Hannaneh Hajishirzi, and Noah Smith. 2024a. Set the clock: Temporal alignment of pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15015–15040, Bangkok, Thailand. Association for Computational Linguistics.

Tony Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*.

Zheng Zhao, Emilio Monti, Jens Lehmann, and Haytham Assem. 2024b. Enhancing contextual understanding in large language models through contrastive decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4225–4237, Mexico City, Mexico. Association for Computational Linguistics.

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*, pages 14544–14556, Singapore. Association for Computational Linguistics.

A Details of Testbed Setting

In this section, we provide the details of the testbed setup process.

A.1 Dataset Details

Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), HotpotQA (Yang et al., 2018) are open-domain question answering datasets, structured to include a question, a short form answer, and a pre-defined context. The answer span, including the ground-truth answer, can be found within the context. Specifically, NQ is composed of information-seeking queries from the Google search engine, and the contexts are Wikipedia pages retrieved by Crowdworkers. HotpotQA is a multi-hop reasoning dataset comprising two entity-linked paragraphs from Wikipedia and questions collected from crowdworkers. Unlike its original setting, which includes distractor paragraphs, we use the processed version from MRQA (Fisch et al., 2019), where the distractors have been removed. TriviaQA utilizes question-and-answer pairs collected from trivia and quiz-league websites. We use the web version of TriviaQA from MRQA.

A.2 Data Preprocessing

The dataset consists of a query x_i, a ground-truth answer y_i, and a pre-defined context c_i. To keep both inputs and outputs concise, we only utilize samples with lengths of the x_i and y_i limited to 50 and 10 words, respectively. In cases where x_i appears multiple times within the context c_i, we extract multiple corresponding spans $(c_i^1, ..., c_i^k)$ for the same query x_i. Each resulting triplets $\{(x_i, c_i^1, y_i), ..., (x_i, c_i^k, y_i)\}$ are included in the dataset $\mathcal{D}_{\text{init}}$.

Relevant Knowledge Estimation Although the pre-defined context always contains the answer span, further validation is necessary since it is split into 100-word spans, which may not contain sufficient information to answer the query. To address this, we measure the sampling consistency by applying a temperature of 1.0 and generating $n = 10$ samples for each query. Each generated sample is compared with the ground-truth answer y to determine its correctness. The consistency rate is compared with a pre-defined threshold η set to 0.7. The knowledge is considered relevant if the consistency rate exceeds the threshold value. In other words, the model is deemed to have relevant knowledge for the input x if at least eight samples are correct out of ten generations.

Backbone	Seed	NQ	HotpotQA	TriviaQA
LLAMA3 8B INSTRUCT	1	2,338	3,614	7,524
	2	1,912	3,400	9,924
	3	2,432	3,462	8,122
LLAMA2 7B CHAT	1	1,462	2,748	13,702
	2	1,758	3,538	14,034
	3	2,088	3,560	14,648
LLAMA2 13B CHAT	1	2,150	3,508	9,806
	2	2,108	3,978	10,274
	3	2,460	3,912	11,284
MISTRAL 7B INSTRUCT	1	764	1,110	9,364
	2	520	1,624	11,578
	3	894	1,438	11,252

Table 4: Number of samples for each dataset constructed for the testbed.

Irrelevant Knowledge Estimation For irrelevant context selection, we utilize SBERT (Reimers and Gurevych, 2019) embedding to measure the cosine similarity between the training set contexts and the relevant context. The context with the highest cosine similarity is selected as the irrelevant context candidate. This process is necessary to avoid selecting contexts that are overly unrelated to the query. Finally, we select contexts with a consistency rate of $r = 0$ to ensure they do not provide any unintended information or hints.

Final Dataset Construction To ensure a balanced dataset, we randomly sample equal number of sample with relevant (i.e., $p_i = 1$) and irrelevant (i.e., $p_i = 0$) parametric knowledge, matching the size of the smaller set. The final number of selected samples for each model is presented in Table 4. Note that the number of selected data points varies across models, reflecting the differences in the possessed knowledge.

A.3 Evaluation Details

For evaluation, we utilize the final dataset D $= \{(x_i, c_i^+, c_i^-, y_i, p_i)\}_{i=1}^{\text{N}}$. Specifically, for $x_i \in D$, we evaluate on both input pairs with relevant (x_i, c_i^+) and irrelevant (x_i, c_i^-) contexts. Only the input (x_i, c_i^-) where $p_i = 0$ is deemed unanswerable. All other inputs are considered answerable by leveraging either parametric or contextual knowledge. For example, the input (x_i, c_i^+) where $p_i = 0$ is answerable since the model can utilize the relevant context c_i^+, even though the model does not pose relevant parametric knowledge (i.e., $p_i = 0$). (x_i, c_i^-) where $p_i = 1$ is also answerable since x_i can be answered by utilizing the model's parametric knowledge (i.e., $p_i = 1$). All the experiments are averaged over three different random seeds. The full results of LLAMA3 8B INSTRUCT, LLAMA2 7B CHAT, LLAMA2 13B CHAT, and MISTRAL 7B INSTRUCT are reported in Table 10, Table 11, Table 12, and Table 13, respectively.

B Experiential Setting Details

In this section, we describe implementation details for the experiment settings.

B.1 Implementation Details of CDA(-M)

CDA(-M) utilize the templates $\mathcal{T}_p(\cdot)$, $\mathcal{T}_c(\cdot)$, and $\mathcal{T}_a(\cdot)$ from Table 3. For the calibration, we set the query placeholder token $\bar{x}$ to "`[QUESTION]`" and the context placeholder token $\bar{c}$ to "`[CONTEXT]`". The output distribution $\bar{d}_t^p$ and $\bar{d}_t^c$ for `null` prompts are computed as follows.

$$\begin{aligned} \bar{d}_t^p &= \mathrm{logit}_\theta(y_t \mid \mathcal{T}_p(\bar{x}, y_{<t})), \\ \bar{d}_t^c &= \mathrm{logit}_\theta(y_t \mid \mathcal{T}_c(\bar{c}, \bar{x}, y_{<t})), \end{aligned} \tag{10}$$

For CDA-M, momentum is applied to each weight as follows:

$$\begin{aligned} w_t^c &\leftarrow \alpha\, w_{t-1}^c + (1-\alpha)\, w_t^c, \\ w_t^p &\leftarrow \alpha\, w_{t-1}^p + (1-\alpha)\, w_t^p, \\ w_t^a &\leftarrow \alpha\, w_{t-1}^a + (1-\alpha)\, w_t^a, \end{aligned} \tag{11}$$

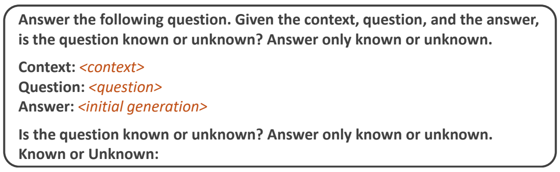
where the momentum weight α is set to 0.7.

B.2 Evaluation Details

The model is provided with a 2-shot demonstration from the training set and evaluated with greedy generation. For answerable queries, the prediction is considered correct if it contains the ground-truth answer (Mallen et al., 2023; Schick et al., 2023). Furthermore, the model is expected to appropriately abstain from generating hallucinations for unanswerable queries. Since there are various ways to abstain, we determine proper abstention by detecting the presence of any pre-defined abstention phrases in the model's output (Amayuelas et al., 2024; Kim et al., 2024a). The pre-defined phrases are the following: `[unknown answer, answer is unknown, unable to answer, no answer, cannot answer, don't know, do not know]`

B.3 Baselines

This section provides details of some baselines.



Answer the following question. Given the context, question, and the answer, is the question known or unknown? Answer only known or unknown.

Context: *<context>*
Question: *<question>*
Answer: *<initial generation>*

Is the question known or unknown? Answer only known or unknown.
Known or Unknown:

Table 5: Verification template $\mathcal{T}_v(\cdot)$ for SELF-ASK. With the generated answer, original question, and context, the model is prompted to verify whether the question is (un)known.

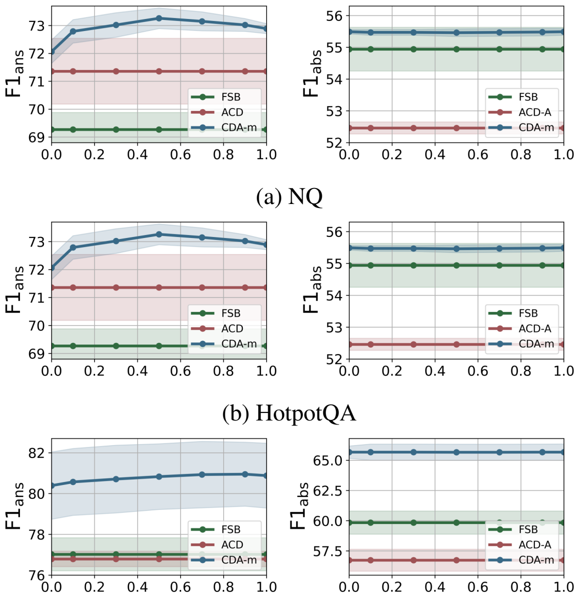
SELF-ASK uses $\mathcal{T}_c(\cdot)$ from Table 3b for the initial generation $\hat{y}$. The initial generation is "self-asked" to the identical model and is verified using the template $\mathcal{T}_v(c, x, \hat{y})$ from Table 5. The prediction is abstained if the model generates "unknown" as the verification result.

CAD computes the output distribution by amplifying the influence of the contextual knowledge as $d_t^o = d_t^c + w_t^c\,(d_t^c - d_t^p)$. A fixed weight w_t^c controls the amount of contextual knowledge applied to the final output distribution. Following the original work (Shi et al., 2024b), we evaluate the performance with w_t^c set to 0.5 and 1.0 and report the best result.

ENTROPY measures the entropy of the generated tokens when prompted with $\mathcal{T}_c(\cdot)$. Specifically, for a prediction $\hat{y} = \{\hat{y}^1, ..., \hat{y}^L\}$ with L-tokens, we leverage the output distribution of each token (i.e., $d^1, ..., d^L$) to measure the token entropy (i.e., $\mathcal{H}^1, ..., \mathcal{H}^L$) following Eq. 5. We measure four different variants: first-token ($\mathcal{H}^1$), average ($\frac{1}{L}\sum_{i=1}^{L}\mathcal{H}^i$), maximum ($\max(\mathcal{H}^1, ..., \mathcal{H}^L)$), and minimum ($\min(\mathcal{H}^1, ..., \mathcal{H}^L)$) entropy.

We compare the entropy measure with a threshold value to perform abstention. Specifically, if the entropy measure exceeds the threshold value, the prediction is considered uncertain and is abstained. We utilize the threshold value, which demonstrates the best Reliability Score (RS) from the training set. We report the first-token entropy, which yields the best performance as the main result, and the results of other variants are reported only in the Appendix.

FSB utilizes $\mathcal{T}_p(\cdot)$, $\mathcal{T}_c(\cdot)$, and $\mathcal{T}_a(\cdot)$ to measure $\mathcal{H}_1^p$, $\mathcal{H}_1^c$, and $\mathcal{H}_1^a$, respectively. We compare the entropy values at the first decoding step and select the prompting method with the lowest entropy (highest confidence). For example, if $\mathcal{H}_1^a$ displays the highest confidence, the model continues to generate the same prediction as ABSTAIN.



(a) NQ

(b) HotpotQA

(c) TriviaQA

Figure 11: $\mathrm{F1}_{\text{ans}}$ and $\mathrm{F1}_{\text{abs}}$ according to different α values. Applying momentum significantly improves $\mathrm{F1}_{\text{ans}}$.

Dataset	$\mathrm{F1}_{\text{abs}}$	$\mathrm{F1}_{\text{ans}}$	RS	Acc.	Cov.
NQ	59.35 (3.43)	52.06 (1.18)	47.89 (4.43)	38.04 (3.87)	56.79 (2.78)
HotpotQA	66.94 (0.84)	56.39 (0.21)	56.42 (1.33)	45.43 (1.29)	63.55 (0.69)
TriviaQA	67.56 (1.27)	57.24 (0.63)	56.46 (1.19)	45.32 (1.14)	64.17 (0.99)

Table 6: Average and standard deviation (in parentheses) of CDA without calibration.

C Details on Ablation Experiments

C.1 Ablation on Momentum Weight

Figure 11 demonstrates the results of $\mathrm{F1}_{\text{ans}}$ and $\mathrm{F1}_{\text{abs}}$ with α values from 0.0 to 1.0 in NQ, HotpotQA, and TriviaQA. We can observe a consistent pattern across the datasets. The results indicate that incorporating momentum enhances the performance of $\mathrm{F1}_{\text{ans}}$ while $\mathrm{F1}_{\text{abs}}$ remains consistent.

C.2 Results of Calibration

Full ablation results regarding the effect of calibration are reported in Table 6. All the experiments are averaged over three different random seeds.

C.3 Implementation Details of Training-based Methods

In this section, we describe the implementation details of training-based methods. Besides **instruction-tuning**, we also utilize **external verifier** (Cobbe et al., 2021; Cohen et al., 2023) for comparison. We report the average performance over three different random seeds. The full results are displayed in Table 7.

Source	Target	$\mathrm{F1}_{\text{abs}}$	$\mathrm{F1}_{\text{ans}}$	RS
NQ	NQ	64.86 (1.49)	46.56 (3.17)	58.03 (1.06)
	HotpotQA	66.62 (1.90)	52.36 (2.61)	58.50 (1.16)
	TriviaQA	65.07 (2.09)	53.52 (3.07)	57.39 (2.17)
HotpotQA	NQ	61.46 (3.13)	38.42 (8.06)	54.60 (2.15)
	HotpotQA	68.41 (0.22)	56.92 (1.32)	60.80 (1.12)
	TriviaQA	60.09 (2.37)	38.53 (11.96)	53.28 (1.04)
TriviaQA	NQ	63.30 (0.74)	45.10 (5.07)	55.55 (0.27)
	HotpotQA	66.97 (0.19)	54.88 (0.98)	59.24 (0.26)
	TriviaQA	65.31 (1.14)	52.63 (2.56)	58.13 (0.43)

(a) Results of external verifier.

Source	Target	$\mathrm{F1}_{\text{ans}}$	$\mathrm{F1}_{\text{abs}}$	RS
NQ	NQ	66.37 (0.55)	43.87 (2.02)	61.12 (0.46)
	HotpotQA	64.43 (0.42)	48.79 (3.88)	59.65 (0.42)
	TriviaQA	67.31 (1.82)	49.00 (4.74)	62.14 (1.43)
HotpotQA	NQ	64.86 (2.38)	42.73 (1.49)	59.80 (1.92)
	HotpotQA	74.06 (0.76)	57.57 (0.73)	68.76 (0.81)
	TriviaQA	65.26 (4.92)	50.69 (2.06)	61.17 (4.10)
TriviaQA	NQ	65.15 (1.66)	46.37 (4.00)	59.70 (1.10)
	HotpotQA	66.68 (0.13)	49.82 (3.52)	60.58 (0.25)
	TriviaQA	67.60 (1.73)	53.68 (2.92)	61.84 (0.70)

(b) Results of instruction-tuning.

Table 7: Average and standard deviation (in parentheses) of training-based methods over three different random seeds. In-domain results are highlighted in blue.

External Verifier We utilize RoBERTa-base as an external verifier. The overall training process is as follows. First, we prompt the inference model (e.g., LLAMA3 8B INSTRUCT) with the prompt $\mathcal{T}_c(c, x)$ and generate a prediction $\hat{y}$. Then, we measure the correctness of $\hat{y}$ by comparing it with the ground-truth answer y. The label for the verifier $\bar{y}$ is assigned based on the correctness of $\hat{y}$. Specifically, we assign $\bar{y} = 1$ when $\hat{y}$ is correct and $\bar{y} = 0$ when $\hat{y}$ is incorrect. We feed $\mathcal{T}_c(c, x, \hat{y})$ into the verifier and pass the [CLS] token through a single MLP layer for the final prediction. The verifier is trained to predict $\bar{y}$ given $\mathcal{T}_c(c, x, \hat{y})$. During inference, we first generate $\hat{y}$ from the inference model. Then, we pass $\mathcal{T}_c(c, x, \hat{y})$ to the verifier to predict the correctness of the prediction $\hat{y}$. Samples classified as incorrect by the verifier are abstained.

Instruction-tuning For instruction-tuning, we first re-label the training data based on the model's knowledge. Specifically, we follow the same knowledge estimation process from Section 3 and estimate the parametric and contextual knowledge of the given training sample. If the model possesses at least one relevant knowledge for the given x, it is labeled with the ground-truth answer y. However, if the model does not have any relevant knowledge, x is labeled with a pre-defined abstention response y_{abs} (e.g., "unknown"). The model is then trained to generate the label given $\mathcal{T}_c(c, x)$. For evaluation, we utilize the instruction-tuned model for the prediction with $\mathcal{T}_c(c, x)$ as the input.

Training Details For instruction-tuning, we employ QLoRA (Dettmers et al., 2023) from Huggingface PEFT library (Mangrulkar et al., 2022) with $r = 4$ and $alpha = 16$ for efficient training. We select the model with the best Reliability Score (RS) performance in the validation set from the learning rates `[1e-3, 5e-4, 1e-4, 5e-5, 1e-5, 5e-6, 1e-6]` and training epochs `[2, 3, 4, 5]`. All results are averaged over three different random seeds. Figure 7 displays the full results of the external verifier and instruction-tuning in both in-domain (IND) and out-of-domain (OOD) scenarios. Overall, instruction-tuning displays better performance than the external verifier in IND scenarios. However, both methods exhibit a **notable degradation in OOD inputs**, limiting their generalizability.

C.4 Evaluation on RAG Setting

Table 8 presents the Reliability Score (RS) results across all the datasets and backbones. Results of Entropy variants (average, maximum, minimum entropy) are also reported. CDA(-M) consistently outperform the baselines, underscoring their effectiveness in practical scenarios.

Method	NQ	HotpotQA	TriviaQA
Llama3 8B Instruct			
Context	33.35	31.00	65.30
CAD	28.86	28.24	59.32
ACD	38.40	34.05	73.91
Abstain	50.74	50.71	76.66
Self-Ask	45.67	39.61	64.91
Entropy (*first-token*)	50.84	50.24	76.20
Entropy (*average*)	48.46	47.67	72.12
Entropy (*max*)	49.40	48.34	74.24
Entropy (*min*)	44.76	43.58	68.61
FSB	53.66	51.07	79.14
ACD-A	51.80	50.35	75.58
CDA	**54.33**	51.77	80.62
CDA-M	54.32	**51.80**	**80.67**
Llama2 7B Chat			
Context	30.11	28.28	61.01
CAD	27.20	25.86	54.44
ACD	33.02	30.04	67.35
Abstain	22.06	20.02	70.70
Self-Ask	31.72	33.99	63.12
Entropy (*first-token*)	41.94	44.71	71.87
Entropy (*average*)	42.88	45.01	70.12
Entropy (*max*)	41.36	**45.69**	71.88
Entropy (*min*)	41.30	41.65	65.67
FSB	37.36	39.58	73.13
ACD-A	33.56	35.05	70.55
CDA	**47.09**	45.07	73.18
CDA-M	46.40	45.02	**73.90**
Llama2 13B Chat			
Context	31.15	30.57	66.81
CAD	28.40	27.59	61.80
ACD	35.65	32.73	72.09
Abstain	41.54	24.54	47.11
Self-Ask	34.15	33.68	67.73
Entropy (*first-token*)	48.28	44.31	74.56
Entropy (*average*)	47.63	43.31	73.88
Entropy (*max*)	48.61	43.93	75.11
Entropy (*min*)	43.73	39.88	70.18
FSB	49.88	**45.28**	72.79
ACD-A	48.30	41.62	65.78
CDA	51.42	44.06	77.89
CDA-M	**51.95**	44.73	**78.74**
Mistral 7B Instruct			
Context	30.68	29.00	62.40
CAD	28.60	26.08	56.09
ACD	32.42	28.91	66.24
Abstain	49.11	41.99	73.00
Self-Ask	39.34	38.78	67.94
Entropy (*first-token*)	48.65	46.82	71.69
Entropy (*average*)	46.84	46.69	71.54
Entropy (*max*)	48.38	47.06	71.57
Entropy (*min*)	42.58	42.76	66.19
FSB	49.05	**47.32**	70.26
ACD-A	48.42	47.14	70.39
CDA	49.36	46.94	72.86
CDA-M	**49.51**	47.13	**73.22**

Table 8: Results of Reliability Score (RS) across all the scenarios in the RAG setting. CDA(-M) consistently outperform all the baselines.

D Computation Cost Analysis

Contrastive decoding enables the model to utilize various knowledge and abilities by leveraging multiple output distributions. However, the process incurs additional costs, which is also an inherent limitation of CDA. In this section, we analyze the computation cost of CDA in detail.

Let the lengths of the template, context, and query be L_t, L_c, and L_q, respectively. The computational cost for a single inference (e.g., Context) is proportional to the square of the total input length $O((L_t + L_c + L_q)^2)$. CDA performs five inferences of contextual prompting $O((L_t + L_c + L_q)^2)$, parametric prompting $O((L_t + L_q)^2)$, abstention prompting $O((L_t + L_c + L_q)^2)$, contextual `null` prompting $O((L_t + 2)^2)$, and parametric `null` prompting $O((L_t + 1)^2)$. While this may appear computationally heavy, in practice, the template and the query are relatively short, while the context is typically long, making the overall computation

Dataset	CONTEXT	CDA	CDA-M
NQ	258.69	622.81 (× 2.41)	645.78 (× 2.50)
HotpotQA	194.87	486.82 (× 2.50)	535.67 (× 2.75)
TriviaQA	188.92	482.65 (× 2.55)	461.87 (× 2.44)

Table 9: Total computation time (in seconds) for LLAMA3 8B INSTRUCT generating 100 samples. We can observe that the total computation time of CDA(-M) is roughly two times that of CONTEXT.

roughly $2 * O(L_c^2)$. Consequently, the additional cost required in CDA is **approximately twice that of a single inference**.

To validate these estimates, we conduct an experiment on LLAMA3 8B INSTRUCT using 100 randomly selected samples from all the datasets. We compare the total generation time of CONTEXT with CDA in seconds on a single RTX 3090 GPU. Table 9 demonstrates the overall inference time. Empirical results indicate that the inference time of CDA does not exceed three times that of a single inference.

Dataset	NQ					HotpotQA					TriviaQA				
Method	$F1_{ans}$	$F1_{abs}$	RS	Acc.	Cov.	$F1_{ans}$	$F1_{abs}$	RS	Acc.	Cov.	$F1_{ans}$	$F1_{abs}$	RS	Acc.	Cov.
CONTEXT	57.26	2.18	50.22	49.95	50.23	57.15	1.19	50.08	49.93	50.08	57.29	2.24	50.23	49.95	50.23
	(0.04)	(0.57)	(0.08)	(0.01)	(0.08)	(0.04)	(0.70)	(0.04)	(0.05)	(0.04)	(0.12)	(1.23)	(0.17)	(0.01)	(0.17)
CAD	55.34	2.14	48.56	48.29	48.56	55.78	1.88	48.89	51.99	48.89	55.83	0.65	48.87	48.79	48.87
	(0.96)	(0.79)	(0.89)	(0.80)	(0.89)	(0.43)	(0.71)	(0.38)	(4.69)	(0.39)	(0.40)	(0.36)	(0.37)	(0.34)	(0.37)
ACD	71.36	0.92	62.50	**62.39**	62.50	74.36	0.57	65.09	**65.02**	65.09	76.79	3.09	67.38	**66.99**	67.39
	(1.18)	(0.35)	(1.06)	(1.02)	(1.06)	(0.57)	(0.17)	(0.50)	(0.50)	(0.50)	(0.38)	(1.56)	(0.44)	(0.26)	(0.44)
ABSTAIN	60.22	52.05	48.17	38.36	57.27	66.88	56.58	56.23	45.21	63.55	67.46	57.31	56.21	45.07	64.10
	(1.85)	(1.18)	(3.52)	(2.94)	(1.84)	(0.84)	(0.31)	(1.35)	(1.29)	(0.71)	(1.35)	(0.56)	(1.34)	(1.28)	(1.06)
SELF-ASK	57.23	48.11	43.91	35.06	53.70	50.64	48.63	33.86	26.80	49.70	52.40	48.38	36.73	29.01	50.61
	(0.87)	(0.35)	(0.48)	(0.46)	(0.64)	(0.78)	(0.30)	(0.57)	(0.49)	(0.51)	(0.28)	(0.39)	(0.39)	(0.34)	(0.06)
ENTROPY	64.06	53.34	55.53	45.15	60.90	67.08	56.44	57.92	46.94	63.88	66.17	57.26	56.50	45.33	63.35
(*first-token*)	(2.23)	(1.11)	(2.181)	(2.00)	(1.96)	(0.29)	(0.21)	(0.89)	(0.95)	(0.25)	(0.71)	(0.36)	(0.82)	(0.74)	(0.50)
ENTROPY	54.91	48.12	41.79	33.19	52.31	58.86	50.94	44.05	35.09	55.72	43.52	47.00	29.53	23.13	45.57
(*average*)	(4.80)	(1.44)	(5.43)	(4.65)	(3.70)	(3.14)	(0.86)	(4.69)	(3.95)	(2.54)	(10.45)	(2.98)	(9.55)	(7.60)	(6.67)
ENTROPY	55.80	49.76	43.06	34.38	53.67	61.44	52.93	46.88	37.29	58.12	33.49	44.95	20.42	15.88	39.65
(*max*)	(7.06)	(2.57)	(8.61)	(6.92)	(5.49)	(3.31)	(1.69)	(4.10)	(3.38)	(2.84)	(6.09)	(1.31)	(4.48)	(3.58)	(3.38)
ENTROPY	54.94	26.45	47.21	42.42	48.45	56.44	25.65	49.25	44.80	50.18	57.42	24.73	50.57	46.44	51.28
(*min*)	(0.38)	(1.03)	(0.58)	(0.44)	(0.58)	(0.47)	(3.97)	(0.54)	(1.41)	(0.25)	(0.30)	(5.82)	(0.44)	(1.29)	(0.65)
FSB	69.27	54.94	59.64	49.02	65.09	74.89	58.51	66.21	55.05	70.55	77.02	59.84	68.55	57.24	72.62
	(0.61)	(0.69)	(1.98)	(2.08)	(0.77)	(0.86)	(0.21)	(0.91)	(0.96)	(0.72)	(0.81)	(0.96)	(1.07)	(1.27)	(0.68)
ACD-A	63.56	52.46	53.99	43.82	60.11	65.88	54.67	57.41	46.80	62.63	66.67	56.73	58.81	57.88	63.87
	(0.14)	(0.19)	(1.72)	(1.67)	(0.36)	(0.71)	(0.17)	(0.92)	(0.95)	(0.60)	(0.65)	(0.91)	(0.65)	(0.79)	(0.39)
CDA	<u>72.06</u>	**55.49**	<u>62.95</u>	52.28	<u>67.51</u>	<u>78.71</u>	<u>62.50</u>	<u>70.20</u>	58.36	<u>74.52</u>	<u>80.39</u>	**65.67**	<u>72.35</u>	60.01	<u>76.67</u>
	(0.42)	(1.46)	(0.23)	(0.30)	(0.17)	(1.05)	(0.51)	(1.07)	(1.03)	(0.96)	(1.65)	(0.67)	(2.33)	(2.43)	(1.60)
CDA-M	**73.15**	<u>55.47</u>	**63.72**	<u>53.16</u>	**68.30**	**79.32**	**62.59**	**70.64**	<u>58.78</u>	**74.99**	**80.93**	<u>65.66</u>	**72.74**	<u>60.40</u>	**77.07**
	(0.34)	(0.12)	(0.33)	(0.55)	(0.12)	(0.91)	(0.52)	(0.99)	(0.97)	(0.85)	(1.63)	(0.67)	(2.30)	(2.39)	(1.59)

Table 10: Average and standard deviation (in parentheses) of the results over three different random seeds in LLAMA3 8B INSTRUCT. The **best result** is highlighted in bold, and the <u>second-best result</u> is underlined.

Dataset	NQ					HotpotQA					TriviaQA				
Method	$F1_{ans}$	$F1_{abs}$	RS	Acc.	Cov.	$F1_{ans}$	$F1_{abs}$	RS	Acc.	Cov.	$F1_{ans}$	$F1_{abs}$	RS	Acc.	Cov.
CONTEXT	57.13	0.09	49.98	49.97	49.98	57.14	0.00	49.98	49.98	49.98	57.17	0.26	50.02	49.99	50.02
	(0.05)	(0.13)	(0.04)	(0.03)	(0.04)	(0.02)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.15)	(0.01)	(0.01)	(0.01)
CAD	54.78	0.17	47.94	47.92	47.94	54.68	0.04	47.84	47.84	47.84	55.11	0.25	48.22	48.19	48.22
	(0.07)	(0.12)	(0.05)	(0.06)	(0.05)	(0.33)	(0.05)	(0.30)	(0.29)	(0.30)	(0.13)	(0.13)	(0.12)	(0.11)	(0.12)
ACD	64.55	0.15	56.48	**56.46**	56.48	69.27	0.00	60.60	<u>60.60</u>	60.60	72.86	0.18	63.74	**63.72**	63.74
	(1.38)	(0.11)	(1.20)	(1.21)	(1.20)	(0.78)	(0.00)	(0.68)	(0.68)	(0.68)	(0.50)	(0.05)	(0.43)	(0.44)	(0.43)
ABSTAIN	38.73	45.16	26.73	20.54	42.11	47.29	48.65	33.49	26.18	48.18	59.19	50.45	47.04	37.87	56.27
	(8.61)	(1.83)	(7.70)	(6.13)	(5.03)	(8.00)	(2.77)	(8.91)	(7.14)	(5.69)	(5.98)	(2.03)	(8.41)	(7.35)	(4.85)
SELF-ASK	56.81	10.95	50.17	48.66	50.23	58.58	17.35	52.19	49.70	52.33	57.93	8.97	51.05	49.84	51.09
	(1.47)	(1.13)	(1.34)	(1.22)	(1.35)	(0.14)	(2.29)	(0.38)	(0.04)	(0.41)	(0.06)	(1.06)	(0.15)	(0.02)	(0.16)
ENTROPY	58.00	41.09	51.98	44.47	54.08	59.00	45.11	52.64	44.16	55.51	60.21	47.84	53.99	45.00	57.10
(*first-token*)	(0.17)	(1.81)	(0.14)	(0.64)	(0.33)	(0.67)	(1.92)	(0.87)	(0.66)	(0.95)	(0.10)	(1.06)	(0.42)	(0.66)	(0.24)
ENTROPY	54.30	40.06	46.38	38.66	50.23	50.07	44.54	39.15	30.93	47.99	56.46	44.75	48.08	39.48	52.93
(*average*)	(2.15)	(0.65)	(2.96)	(3.03)	(2.03)	(2.37)	(0.85)	(2.57)	(2.32)	(1.73)	(0.72)	(0.66)	(1.35)	(1.43)	(0.74)
ENTROPY	54.96	40.96	47.82	43.34	51.21	38.94	42.71	29.66	23.15	41.41	56.94	46.74	49.30	40.49	54.13
(*max*)	(1.27)	(1.20)	(2.04)	(7.10)	(1.02)	(11.39)	(2.27)	(12.32)	(10.62)	(7.57)	(3.75)	(0.61)	(5.73)	(5.63)	(3.24)
ENTROPY	38.59	21.56	48.25	44.60	48.91	57.30	22.53	50.56	46.93	51.05	57.01	17.77	50.22	47.48	50.49
(*min*)	(23.18)	(5.37)	(0.54)	(1.65)	(0.13)	(0.31)	(3.98)	(0.57)	(0.79)	(0.65)	(0.04)	(3.78)	(0.06)	(0.70)	(0.17)
FSB	55.04	47.26	43.26	34.47	52.20	65.63	**53.82**	55.73	45.37	62.04	69.50	<u>51.88</u>	60.41	50.59	64.78
	(3.59)	(0.84)	(4.01)	(3.60)	(2.57)	(2.31)	(0.93)	(4.11)	(4.07)	(2.22)	(2.21)	(1.57)	(4.05)	(4.28)	(2.23)
ACD-A	48.82	39.52	37.55	30.41	45.66	57.79	<u>51.11</u>	47.90	38.19	55.61	61.21	50.52	53.09	43.46	58.18
	(4.58)	(7.69)	(4.90)	(4.68)	(3.53)	(3.06)	(1.22)	(4.90)	(4.56)	(2.64)	(1.78)	(1.43)	(3.82)	(3.89)	(1.73)
CDA	<u>66.86</u>	<u>47.52</u>	<u>59.86</u>	51.22	<u>62.38</u>	<u>73.39</u>	42.41	<u>66.96</u>	60.15	<u>67.82</u>	**73.70**	51.29	**67.29**	<u>58.43</u>	**69.06**
	(2.51)	(1.36)	(2.94)	(3.55)	(2.01)	(0.76)	(3.80)	(1.00)	(0.78)	(1.08)	(0.71)	(1.42)	(1.04)	(1.34)	(0.69)
CDA-M	**69.99**	**47.60**	**62.28**	<u>53.62</u>	**64.81**	**74.09**	42.31	**67.50**	**60.70**	**68.37**	<u>73.47</u>	**52.10**	<u>67.11</u>	58.04	<u>69.00</u>
	(1.40)	(1.32)	(2.12)	(2.71)	(1.23)	(0.94)	(3.54)	(1.13)	(0.87)	(1.21)	(0.43)	(1.75)	(1.01)	(1.24)	(0.78)

Table 11: Average and standard deviation (in parentheses) of the results over three different random seeds in LLAMA2 7B CHAT. The **best result** is highlighted in bold, and the <u>second-best result</u> is underlined.

Dataset	NQ					HotpotQA					TriviaQA				
Method	F1$_{\text{ans}}$	F1$_{\text{abs}}$	RS	Acc.	Cov.	F1$_{\text{ans}}$	F1$_{\text{abs}}$	RS	Acc.	Cov.	F1$_{\text{ans}}$	F1$_{\text{abs}}$	RS	Acc.	Cov.
Context	57.14	0.37	50.00	49.95	50.00	57.16	0.10	50.00	49.98	50.00	57.15	0.47	50.03	49.97	50.03
	(0.02)	(0.26)	(0.00)	(0.03)	(0.00)	(0.04)	(0.08)	(0.03)	(0.02)	(0.03)	(0.05)	(0.09)	(0.04)	(0.03)	(0.04)
CAD	45.57	0.48	39.90	39.84	39.90	54.60	0.07	47.76	47.75	47.76	55.81	0.32	48.84	48.80	48.84
	(12.58)	(0.08)	(11.01)	(11.02)	(11.01)	(0.51)	(0.05)	(0.45)	(0.46)	(0.45)	(0.14)	(0.11)	(0.11)	(0.12)	(0.11)
ACD	66.66	0.18	58.33	**58.31**	58.33	69.80	0.07	61.15	**61.14**	61.15	72.49	0.42	63.43	**63.38**	63.43
	(0.65)	(0.01)	(0.56)	(0.56)	(0.56)	(0.63)	(0.05)	(0.49)	(0.48)	(0.49)	(0.67)	(0.24)	(0.61)	(0.60)	(0.61)
Abstain	50.49	46.62	37.83	29.70	48.93	57.17	51.95	43.02	33.89	55.10	48.53	46.74	35.11	27.52	47.88
	(1.98)	(0.33)	(2.63)	(2.19)	(1.30)	(2.74)	(0.89)	(3.56)	(3.05)	(1.95)	(4.52)	(1.01)	(5.28)	(4.37)	(3.14)
Self-Ask	59.37	20.80	52.98	49.89	53.21	58.22	14.36	51.87	49.89	51.95	58.09	12.07	51.53	49.85	51.59
	(0.06)	(1.48)	(0.21)	(0.03)	(0.23)	(0.04)	(1.32)	(0.16)	(0.02)	(0.17)	(0.11)	(1.21)	(0.14)	(0.07)	(0.16)
Entropy	59.23	42.61	52.75	44.85	55.19	59.24	45.47	53.01	44.49	55.81	60.21	48.31	54.34	45.32	57.27
(*first-token*)	(0.45)	(1.11)	(0.58)	(0.48)	(0.56)	(0.38)	(1.84)	(0.77)	(0.53)	(0.76)	(0.21)	(0.07)	(0.15)	(0.17)	(0.14)
Entropy	56.18	42.85	47.54	39.22	52.18	56.62	46.18	48.00	39.01	53.37	57.58	46.10	49.57	40.65	54.25
(*average*)	(1.25)	(0.54)	(1.91)	(1.96)	(1.07)	(0.92)	(0.97)	(1.68)	(1.53)	(1.09)	(0.81)	(2.06)	(1.22)	(1.46)	(0.97)
Entropy	57.69	41.64	50.81	42.96	53.59	54.24	44.75	45.49	37.21	51.61	57.75	47.09	50.18	41.26	54.80
(*max*)	(0.62)	(0.46)	(1.10)	(1.12)	(0.75)	(3.25)	(2.55)	(5.71)	(6.05)	(2.23)	(2.01)	(0.37)	(3.85)	(4.29)	(1.95)
Entropy	55.30	24.55	48.25	44.04	49.05	57.17	15.07	50.60	48.39	50.76	56.74	14.04	49.57	47.42	49.78
(*min*)	(0.40)	(0.92)	(0.59)	(0.76)	(0.48)	(0.19)	(4.99)	(0.22)	(0.80)	(0.29)	(0.12)	(4.43)	(0.26)	(0.54)	(0.35)
FSB	62.44	47.60	53.66	44.46	58.19	68.68	54.16	60.41	50.05	64.79	66.21	52.08	56.12	45.98	61.91
	(1.68)	(0.54)	(2.21)	(2.37)	(1.31)	(1.01)	(2.21)	(1.77)	(2.24)	(0.83)	(1.07)	(0.99)	(1.83)	(1.84)	(1.04)
ACD-A	57.96	46.34	49.28	40.27	54.43	62.00	52.13	54.03	43.95	59.16	58.42	49.72	49.22	39.54	55.61
	(0.80)	(0.27)	(1.55)	(1.59)	(0.72)	(0.44)	(2.44)	(1.40)	(1.56)	(0.91)	(0.69)	(0.57)	(1.26)	(1.13)	(0.74)
CDA	68.62	47.14	61.63	53.16	63.76	**73.69**	56.81	**68.98**	59.53	**70.44**	71.08	51.44	64.08	54.78	66.55
	(0.65)	(1.79)	(1.14)	(0.94)	(1.02)	(0.76)	(2.25)	(0.19)	(0.53)	(0.15)	(1.13)	(1.66)	(0.85)	(0.69)	(0.85)
CDA-m	**70.66**	**48.12**	**63.18**	54.46	**65.48**	73.66	**56.89**	68.92	59.42	70.42	**73.12**	**53.46**	**65.82**	56.09	**68.52**
	(0.63)	(2.06)	(1.15)	(0.87)	(1.06)	(0.65)	(3.88)	(0.25)	(0.52)	(0.31)	(1.22)	(1.19)	(0.58)	(0.66)	(0.54)

Table 12: Average and standard deviation (in parentheses) of the results over three different random seeds in Llama2 13B Chat. The **best result** is highlighted in bold, and the second-best result is underlined.

Dataset	NQ					HotpotQA					TriviaQA				
Method	F1$_{\text{ans}}$	F1$_{\text{abs}}$	RS	Acc.	Cov.	F1$_{\text{ans}}$	F1$_{\text{abs}}$	RS	Acc.	Cov.	F1$_{\text{ans}}$	F1$_{\text{abs}}$	RS	Acc.	Cov.
Context	57.04	0.25	49.88	49.85	49.88	56.99	0.09	49.88	49.87	49.88	57.13	0.35	50.00	49.96	50.00
	(0.07)	(0.36)	(0.06)	(0.06)	(0.06)	(0.04)	(0.13)	(0.04)	(0.03)	(0.04)	(0.02)	(0.15)	(0.03)	(0.01)	(0.03)
CAD	55.03	1.15	48.21	48.07	48.21	53.74	0.20	47.04	47.01	47.04	54.71	0.57	47.89	47.82	47.89
	(0.25)	(0.26)	(0.21)	(0.24)	(0.21)	(0.46)	(0.15)	(0.40)	(0.41)	(0.40)	(0.10)	(0.09)	(0.09)	(0.10)	(0.09)
ACD	68.48	0.00	59.93	**59.93**	59.93	72.72	0.09	63.64	**63.62**	63.64	75.01	0.05	65.62	**65.62**	65.62
	(2.18)	(0.00)	(1.86)	(1.86)	(1.86)	(0.12)	(0.13)	(0.10)	(0.11)	(0.10)	(0.30)	(0.05)	(0.27)	(0.27)	(0.27)
Abstain	61.86	53.29	52.85	42.83	59.70	61.12	54.66	53.06	42.55	59.23	60.73	53.16	51.25	41.22	58.42
	(8.11)	(1.45)	(9.08)	(9.42)	(6.00)	(2.58)	(0.83)	(3.88)	(3.89)	(1.88)	(1.94)	(2.11)	(4.36)	(4.36)	(1.84)
Self-Ask	62.52	49.03	57.49	48.76	59.54	61.33	43.36	56.36	48.95	57.70	62.18	48.60	57.19	48.55	59.20
	(0.80)	(3.29)	(0.69)	(0.23)	(0.82)	(0.27)	(2.23)	(0.47)	(0.13)	(0.55)	(0.21)	(0.75)	(0.15)	(0.07)	(0.15)
Entropy	61.91	56.29	55.48	44.63	60.33	63.31	60.07	57.19	45.65	62.40	62.21	56.59	56.23	45.48	60.71
(*first-token*)	(0.37)	(1.15)	(0.66)	(0.57)	(0.11)	(0.26)	(1.91)	(1.51)	(1.50)	(0.71)	(0.42)	(0.95)	(1.22)	(1.59)	(0.43)
Entropy	60.03	50.49	52.81	43.04	57.34	60.98	55.02	54.19	43.52	59.28	59.87	52.83	53.82	43.80	58.04
(*average*)	(0.32)	(2.70)	(0.52)	(0.06)	(0.85)	(0.56)	(3.32)	(0.98)	(0.50)	(1.33)	(0.17)	(2.02)	(1.47)	(2.03)	(0.36)
Entropy	61.73	**56.68**	55.73	44.92	60.37	63.44	**61.65**	57.50	45.58	62.94	62.82	**61.16**	56.79	45.06	62.32
(*max*)	(0.66)	(2.22)	(1.54)	(1.86)	(0.29)	(0.27)	(2.06)	(0.44)	(0.09)	(0.72)	(0.35)	(1.26)	(2.06)	(1.96)	(0.62)
Entropy	56.45	23.04	49.40	45.56	50.01	55.86	26.93	49.26	44.67	50.13	57.01	17.86	50.36	47.62	50.63
(*min*)	(0.23)	(0.96)	(0.33)	(0.48)	(0.27)	(0.59)	(4.41)	(0.81)	(0.95)	(0.99)	(0.06)	(4.62)	(0.02)	(0.83)	(0.20)
FSB	66.71	55.51	58.95	48.32	63.65	74.32	52.18	68.20	59.27	69.94	**77.67**	47.53	**70.69**	62.72	72.07
	(3.35)	(1.18)	(3.89)	(3.83)	(2.90)	(1.15)	(4.47)	(1.17)	(2.38)	(0.49)	(0.63)	(9.54)	(0.81)	(1.25)	(1.57)
ACD-A	61.46	54.02	55.80	45.58	59.56	61.02	50.93	57.01	48.10	58.95	61.62	51.62	56.81	47.48	59.48
	(1.38)	(1.66)	(1.23)	(1.68)	(0.65)	(0.35)	(4.60)	(0.35)	(0.79)	(1.03)	(1.17)	(6.00)	(0.81)	(1.16)	(1.86)
CDA	69.68	56.47	61.45	50.61	66.09	76.28	55.84	69.43	59.50	71.83	75.76	56.35	67.57	57.21	71.21
	(2.26)	(1.31)	(2.31)	(2.36)	(1.93)	(1.37)	(1.97)	(1.68)	(2.30)	(1.11)	(1.62)	(5.43)	(2.79)	(3.74)	(1.73)
CDA-m	**71.00**	56.46	**62.30**	51.45	**67.03**	**76.94**	56.67	**69.98**	59.85	**72.49**	76.95	57.06	68.38	57.84	**72.21**
	(2.27)	(1.50)	(2.30)	(2.39)	(1.91)	(1.54)	(1.95)	(1.87)	(2.53)	(1.21)	(1.49)	(5.18)	(2.94)	(3.94)	(1.59)

Table 13: Average and standard deviation (in parentheses) of the results over three different random seeds in Mistral 7B Instruct. The **best result** is highlighted in bold, and the second-best result is underlined.