Adaptive Contrastive Decoding in Retrieval-Augmented Generation for Handling Noisy Contexts

Youna Kim[1], **Hyuhng Joon Kim**[1], **Cheonbok Park**[2 3], **Choonghyun Park**[1], **Hyunsoo Cho**[4], **Junyeob Kim**[1], **Kang Min Yoo**[1 2 5], **Sang-goo Lee**[1 6], **Taeuk Kim**[7 *]
[1]Seoul National University, [2]NAVER Cloud, [3]KAIST AI, [4]Ewha Womans University,
[5]NAVER AI LAB, [6]IntelliSys, Korea, [7]Hanyang University
{anna9812, heyjoonkim, pch330, juny116, sglee}@europa.snu.ac.kr
{cbok.park, kangmin.yoo}@navercorp.com, chohyunsoo@ewha.ac.kr
kimtaeuk@hanyang.ac.kr

Abstract

When using large language models (LLMs) in knowledge-intensive tasks, such as open-domain question answering, external context can bridge the gap between external knowledge and the LLMs' parametric knowledge. Recent research has been developed to amplify contextual knowledge over the parametric knowledge of LLMs with contrastive decoding approaches. While these approaches could yield truthful responses when relevant context is provided, they are prone to vulnerabilities when faced with noisy contexts. We extend the scope of previous studies to encompass noisy contexts and propose adaptive contrastive decoding (ACD) to leverage contextual influence effectively. ACD demonstrates improvements in open-domain question answering tasks compared to baselines, especially in robustness by remaining undistracted by noisy contexts in retrieval-augmented generation.

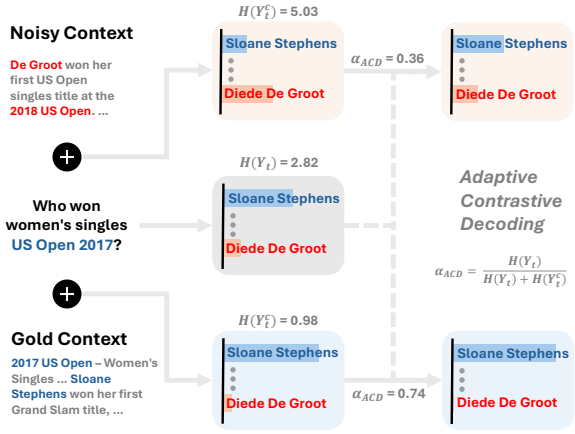

Figure 1: An illustration of adaptive contrastive decoding (ACD). Entropy (H) changes depending on context relevance, affecting the adaptive weight (α_{ACD}). Noisy context leads the model to incorrectly answer "Diede De Groot" when employing regular greedy decoding. ACD applies context-based adjustments, enabling the correct answer, "Sloane Stephens," despite the noise.

1 Introduction

While large language models (LLMs) (Touvron et al., 2023; Achiam et al., 2023) achieve remarkable performance levels across diverse benchmarks, they sometimes struggle to generalize to knowledge-intensive tasks, such as open-domain question-answering (QA; Chen et al., 2017), and may also fail to capture long-tail knowledge, leading to unfaithful output generation (Mallen et al., 2023; Kandpal et al., 2023). One common approach to address these limitations is fine-tuning the model, but this results in a quadratic rise in computational demands as the size of the LLMs increases exponentially (Longpre et al., 2023). To overcome this, researchers have been investigating strategies to combine non-parametric knowledge with LLMs during response generation without explicit re-training (Asai et al., 2023a). This approach leverages external information from knowledge bases and enhances the capability of the LLMs dynamically, ensuring that the information is both current and accurate.

Early studies in this field attempt to append query-relevant context to generate more accurate responses. Especially, contrastive decoding (Li et al., 2023; Malkin et al., 2022; Liu et al., 2021) yields significant enhancement in various tasks by amplifying the influence of the given context at decoding step (Shi et al., 2023; Zhao et al., 2024). While such methods work well when context information is correct and faithful, in real-world scenarios, context information is not always correct and may contain some noisy and unfaithful information. For instance, if the retrieval system pulls in irrelevant or contradictory information, it could lead to incorrect responses (Wang et al., 2024; Wu et al., 2024; Yu et al., 2024). This highlights the necessity for a generation model that can gauge the appropriateness of the context by itself, being robust to noise

*Corresponding author.

Findings of the Association for Computational Linguistics: EACL 2024, pages 2421–2431

and unfaithful data to ensure the output remains reliable (Yoran et al., 2024).

To assess whether the existing contrastive decoding approaches can be utilized in practice, we extend the setting to situations where the gold-standard context is not guaranteed, specifically in the retrieval-augmented generation (RAG) framework (Yao et al., 2022; Shi et al., 2024; Izacard et al., 2023). In this paper, we demonstrate that existing context-aware contrastive decoding approaches experience performance drops in open-domain question answering, especially when the retrieved context is noisy. To address this issue, we propose **adaptive contrastive decoding (ACD)**, adaptively weighting the contrastive contextual influence on the parametric knowledge, making it suitable for noisy context settings (Figure 1).

Incorporating the distinction between contextual and parametric knowledge, our approach aims to mitigate the dominance of potentially noisy contextual information in model output. We control contrastive contextual influence based on context's contribution to the LLM's uncertainty reduction, thereby minimizing its disruptive effect during decoding. Through in-depth experiments with three open-domain QA datasets, we demonstrate the potential of the proposed approach with increased overall performance. Moreover, ACD enhances the performance significantly on the noisy context scenario while minimizing performance degradation on the gold context scenario compared to the baselines.

2 Related Works

Context-Augmented Generation Approaches for context-augmented generation have been developed to enhance the model's limited parametric knowledge by providing external knowledge, enabling more factual and contextually accurate responses during inference (Zhou et al., 2023; He et al., 2024). To sufficiently incorporate the information from the context in model generation, contrastive decoding approaches are applied to overwrite the model's parametric knowledge with external knowledge (Shi et al., 2023; Zhao et al., 2024). These context-aware contrastive decoding methods to generate responses faithful to the given context show effective performance in summarization (See et al., 2017; Narayan et al., 2018), knowledge conflict (Longpre et al., 2022), and question answering with gold-standard contexts.

Robustness in RAG Frameworks While retrieval-augmented generation enables LLMs to become factual and reliable with the retrieved external knowledge, there are still concerns about incorrectly retrieved irrelevant contexts (Yoran et al., 2024). To address hallucination errors posed by irrelevant contexts, some researchers take an approach to train LLMs that can adaptively retrieve relevant context (Asai et al., 2023b; Wang et al., 2024). Another approach aims to selectively use retrieved contexts after assessing their truthfulness or relevance through context verification with prompting strategies or training untruthful context detectors (Yu et al., 2024; Zhang et al., 2024). These approaches highlight the ongoing efforts to advance the robustness and accuracy of LLMs in multiple directions to manage potentially misleading information.

3 Methodology

3.1 Problem Formulation

At decoding time step t, given the input x and preceding sequences $y_{<t}$, a pretrained auto-regressive LLM θ computes the logit $\mathbf{z}_t \in \mathbb{R}^{|V|}$, where V is the vocabulary, for the t-th token. In the open-domain QA task, a question q serves as the input x, and $\mathbf{z}_t$ relies solely on the LLM's parametric knowledge. When both q and the retrieved context c are provided as x, the logit is denoted as $\mathbf{z}_t^c \in \mathbb{R}^{|V|}$.

3.2 Contrastive Decoding

In cases where context cannot be blindly trusted, directly following the context-augmented distribution can increase the risk of being misled. Thus, we adopt the approach of adding the contextual influence, which contrasts with the LLM's parametric knowledge, to the parametric distribution $\mathbf{z}_t$. With the contrastive decoding objective, $\mathbf{z}_t^c$ and $\mathbf{z}_t$ are ensembled to reflect the influence of external context on the LLM's parametric knowledge at each decoding step t. The probability distribution $P_\theta(Y_t|x, y_{<t})$ is modified by weighted adjustment based on the difference between $\mathbf{z}_t^c$ and $\mathbf{z}_t$, as represented in the following equation.

$$P_\theta(Y_t \mid x, y_{<t}) = \text{softmax}(\mathbf{z}_t + \alpha\,(\mathbf{z}_t^c - \mathbf{z}_t)) \quad (1)$$

The contrastive adjustment enables the LLM to integrate external context c into its prediction, leveraging the weight α to control the impact of c on the final probability distribution.

3.3 Adaptive Weight on Contextual Influence

The degree to which contextual influence is incorporated into $\mathbf{z}_t$ needs to be controlled based on the provided context's informativeness. In practice, however, it is often unknown whether the context is gold or noisy. To address this, we investigate whether the model could adjust accordingly with a simple entropy-based approach.

The LLM's uncertainty is expressed with the entropy $H(Y_t)$ of its probability distribution $P_\theta(Y_t|x, y_{<t})$ (Huang et al., 2023; Kuhn et al., 2023). While $H(Y_t)$ reflects how much uncertainty the model has based on its parametric knowledge under the given question, $H(Y_t^c)$ is influenced by the external knowledge within the retrieved context c. Generally, when the context is added, the entropy decreases (Kendall and Gal, 2017). However, if the context is noisy, irrelevant, or provides no information to answer the given question, it may contribute to increased uncertainty instead.

Intuitively, if the retrieved context provides informative cues for answering the question, then $H(Y_t^c)$ is expected to be lowered compared to $H(Y_t)$. Conversely, if the context is non-helpful or even confusing the model prediction, $H(Y_t^c)$ in predicting the next token is likely to be higher. This scenario would be particularly evident when the model knows the answer with low $H(Y_t)$.

Considering the above scenarios, the motivation behind the adaptive weight α_{ACD} is to assign a relatively smaller weight in cases where the context increases uncertainty by being uninformative or confusing for the model in answering the given question. Thus, the value of α_{ACD} is set as the proportion of uncertainty contributed by $H(Y_t)$ relative to the total uncertainty when considering both $H(Y_t)$ and $H(Y_t^c)$:

$$\alpha_{ACD} = \frac{H(Y_t)}{H(Y_t) + H(Y_t^c)} \tag{2}$$

Under the condition where $H(Y_t) > H(Y_t^c)$, α_{ACD} value approaches to 1, indicating that when the context c is provided, the uncertainty associated with predicting the next token decreases. Conversely, when $H(Y_t) < H(Y_t^c)$, α_{ACD} value approaches to 0, reflecting minimal influence from c. Note that when $H(Y_t) = H(Y_t^c)$, α_{ACD} becomes 0.5, resulting in an ensemble of two distributions, $\mathbf{z}_t$ and $\mathbf{z}_t^c$, with equal weighting.

With α_{ACD}, the vocab v with maximum probability is selected as the next token under the following distribution:

$$\hat{P}_\theta(Y_t \mid x, y_{<t}) = \text{softmax}(\mathbf{z}_t + \alpha_{ACD}\,(\mathbf{z}_t^c - \mathbf{z}_t)) \tag{3}$$

Informed by α_{ACD} and contextual contrast, the adjustment process determines the degree to which the model's parametric knowledge is superseded, thus optimizing the assimilation of contextual information throughout decoding.

4 Experimental Results

4.1 Experimental Settings

Datasets and Models We conduct experiments on open-domain QA datasets, TriviaQA (Joshi et al., 2017), Natural Questions (NQ; Kwiatkowski et al., 2019), and PopQA (Mallen et al., 2022) with Wikipedia contexts.[2]

We use auto-regressive language models, LLAMA2 (7B & 13B, Touvron et al., 2023), LLAMA3 8B,[3] and MISTRAL 7B (Jiang et al., 2023). Utilizing CONTRIEVER-MSMARCO (Izacard et al., 2022) as a retriever, the top-1 retrieved context is appended to each question.

Evaluation Metric Following Zhao et al. (2024), we use few-shot prompts with 5 examples. We report Exact Match (EM) as an evaluation metric, which verifies whether the generated sequences precisely match one of the candidate answers.

Baselines As fundamental baselines, regular greedy decoding has been employed in open-book (Reg_{Opn}) and closed-book (Reg_{Cls}) settings. We compare our method against existing context-aware contrastive decoding methods, including Context-Aware Decoding (CAD; Shi et al., 2023) and Multi-Input Contrastive Decoding (MICD; Zhao et al., 2024). MICD uses inputs with and without context, along with an additional input with adversarial context, to generate the output distribution. MICD presents two methods, referred to as MICD_F and MICD_D, which offer fixed and dynamic α, respectively. Similar to our approach, to leverage the burden of hyperparameter search and dependency on fixed α, MICD_D also determines α dynamically. In MICD_D, α is assigned as the maximum token probability with context ($\max P_{wc}$) if $\max P_{wc}$ exceeds the maximum token probability without context ($\max P_{woc}$); otherwise, it is calculated as $1 - \max P_{woc}$.

[2] Wikipedia dump from Dec. 2018.

[3] https://github.com/meta-llama/llama3

	Dataset (→)	TriviaQA			NQ			PopQA		
Model	Method (↓)	All	$\text{Subset}_{\text{Gold}}$	$\text{Subset}_{\text{Noisy}}$	All	$\text{Subset}_{\text{Gold}}$	$\text{Subset}_{\text{Noisy}}$	All	$\text{Subset}_{\text{Gold}}$	$\text{Subset}_{\text{Noisy}}$
LLAMA2 7B	Reg_{Cls}	59.00	-	-	25.48	-	-	28.36	-	-
	Reg_{Opn}	60.23	87.40	33.50	31.39	**61.31**	12.40	38.49	81.21	7.77
	CAD	49.02	73.69	24.75	25.57	51.61	9.05	33.70	72.18	6.03
	MICD_F	60.36	85.72	35.39	29.45	56.10	12.54	35.73	74.25	8.03
	MICD_D	63.23	86.03	40.79	30.36	52.18	16.52	39.01	77.39	11.42
	ACD	**64.85**	**88.01**	**42.06**	**32.91**	56.60	**17.88**	**41.29**	**82.77**	**11.46**
LLAMA2 13B	Reg_{Cls}	63.77	-	-	30.80	-	-	32.70	-	-
	Reg_{Opn}	62.81	88.52	37.51	33.35	**62.96**	14.58	40.03	83.20	8.98
	CAD	52.62	76.78	28.85	27.87	55.96	10.05	35.86	76.38	6.71
	MICD_F	63.53	87.40	40.04	32.63	59.67	15.48	38.16	77.04	10.21
	MICD_D	66.52	87.68	45.69	34.38	57.32	19.83	41.65	79.27	**14.60**
	ACD	**67.37**	**89.36**	**45.74**	**36.12**	61.17	**20.24**	**43.35**	**83.98**	14.14
LLAMA3 8B	Reg_{Cls}	61.67	-	-	28.34	-	-	32.65	-	-
	Reg_{Opn}	61.27	86.94	36.02	33.30	**63.10**	14.40	39.73	82.95	8.64
	CAD	49.70	72.45	27.31	29.17	58.39	10.64	35.86	76.82	6.40
	MICD_F	61.01	85.40	37.00	27.62	51.89	12.22	37.99	77.12	9.85
	MICD_D	64.01	86.08	42.28	30.72	53.96	15.98	41.35	79.32	**14.04**
	ACD	**66.32**	**89.20**	**43.81**	**35.48**	62.03	**18.65**	**43.25**	**84.48**	13.60
MISTRAL 8B	Reg_{Cls}	63.72	-	-	29.64	-	-	29.04	-	-
	Reg_{Opn}	60.45	86.85	34.48	32.55	**64.67**	12.18	38.28	81.26	7.36
	CAD	44.69	66.89	22.85	24.10	52.25	6.25	33.93	73.95	5.15
	MICD_F	63.33	88.43	38.62	31.80	61.10	13.22	36.58	76.00	8.23
	MICD_D	66.97	89.24	45.05	33.24	57.89	17.61	39.87	78.46	**12.11**
	ACD	**67.82**	**90.16**	**45.83**	**35.37**	62.17	**18.38**	**41.47**	**82.90**	11.68

Table 1: EM accuracy of full data (All) and subsets with gold ($\text{Subset}_{\text{Gold}}$) and noisy contexts ($\text{Subset}_{\text{Noisy}}$). The highest score is in **bold**, and the second-best is underlined.

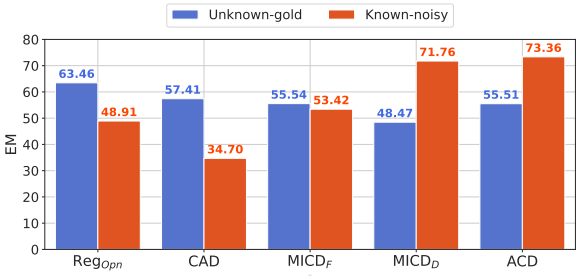

Figure 2: EM accuracy of each method in LLAMA2-7B. EM of three datasets used are averaged for each subset, *Unknown-gold* and *Known-noisy*.

4.2 Main Results

Performance on RAG As shown in Table 1, ACD outperforms the baselines across all datasets and models within the RAG framework, particularly when considering the full test data (All). When analyzing the performance by dividing the data into two subsets based on whether the retrieved context is gold ($\text{Subset}_{\text{Gold}}$) or not ($\text{Subset}_{\text{Noisy}}$), ACD achieves either the best or second-best performance. MICD_D demonstrates performance comparable to ACD on $\text{Subset}_{\text{Noisy}}$. However, it shows a significant drop on $\text{Subset}_{\text{Gold}}$, indicating a tendency to ignore gold context while handling noisy context. It is notable that both CAD and MICD_F exhibit a significant drop in their performance under noisy conditions.

Performance under Parametric Knowledge We aim to analyze the model's performance across various aspects, focusing specifically on its parametric knowledge. We estimate whether the model possesses relevant parametric knowledge for a given question based on its accuracy in a closed-book setting (Reg_{Cls}). We consider two subsets under the following conditions: (1) *Known-noisy*: the model has parametric knowledge of the given question and noisy context is retrieved. (2) *Unknown-gold*: the model does not have parametric knowledge of the given question and gold context is retrieved.

From Figure 2, we observe that ACD outperforms the baselines in *Known-noisy*. Notably, two approaches with adaptively adjusted weight, ACD and MICD_D, perform well in *Known-noisy*, while other baselines show a relative strength in *Unknown-gold*. However, these baselines also experience significant performance drops in *Known-noisy*, indicating distraction by noisy context despite correctly answering when only the question is provided. In both cases, ACD demonstrates better performance compared to MICD_D, overall showing a tendency towards reliability.

4.3 Analysis

Correlation between Adaptive Weight and Context Noisiness While other baselines rely on the fixed hyperparameter of weight α, ACD and

	α	NQ	TriviaQA	PopQA
Max	MICD_D	51.53	59.76	65.49
	ACD	**65.78**	**73.37**	**74.84**
Avg.	MICD_D	54.18	63.78	72.64
	ACD	**68.80**	**72.32**	**78.90**
First	MICD_D	53.92	62.95	68.81
	ACD	**73.27**	**80.45**	**80.08**

Table 2: AUROC between α used in each method and the noisiness of the retrieved context.

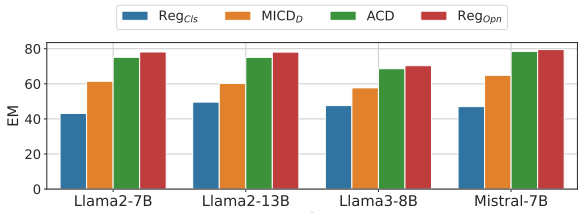

Figure 3: EM accuracy on NQ-swap with contexts replacing the gold answer with a random entity span.

MICD_D adjust α during the decoding step. It depends not only on the noisiness of the retrieved context but also on whether the model's parametric knowledge contains an answer to the given question. To exclude cases that are not directly related to the analysis of how weight is adjusted based on context quality and the model's parametric knowledge, we use the same subsets, *Known-noisy* and *Unknown-gold*.

Adaptive weights α_{ACD} and α_{MICD} are extracted at each decoding step and analyzed across three metrics: maximum, average, and the first within the generated sequence. As an evaluation metric, the area under the receiver operator characteristic curve (AUROC) between α and the noisiness of the retrieved context is measured. AUROC of each α for LLAMA 2-7B is reported in Table 2. Under every metric and dataset, ACD demonstrates a higher AUROC compared to MICD_D. Aligned with our motivation, when the model is knowledgeable and presented with noisy context, α_{ACD} tends to be lower, emphasizing greater reliance on parametric knowledge. Conversely, when the model lacks knowledge and is provided with gold context, α_{ACD} is adjusted to prioritize reliance on the provided context.

Handling Knowledge Conflict With a knowledge conflict QA dataset, NQ-swap (Longpre et al., 2022), we verify whether the two decoding methods with dynamic weight, ACD and MICD_D, can generate context-based responses without considering a conflicting context as a noisy context. The conflicting context in the NQ-swap dataset is constructed by replacing the answer entity span in the original gold context with a random entity of the same type. Figure 3 illustrates that ACD consistently exceeds the performance of MICD_D across all models and achieves results comparable to open-book regular decoding. The results indicate that the ACD's approach remains effective even in settings where the context is relevant to the question but contradicts the model's parametric knowledge.

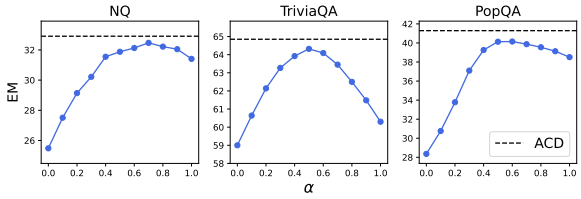

Figure 4: EM across alpha values ranges from 0.0 to 1.0. The dashed line indicates EM score with α_{ACD}.

Ablation on α_{ACD} To assess the impact of α_{ACD} on performance, we fix the value of α within a range $[0, 1]$ and examine whether employing ACD is more effective than optimizing a fixed weight. In Figure 4, it can be observed that using a fixed α results in degraded performance compared to ACD. Increasing the alpha value, which enhances the contextual influence on the output distribution, initially leads to a rise in the EM score. However, beyond a certain point, further increasing α results in a decline in the EM score. In scenarios with potential noisy context, a fixed α value may not ensure optimal performance. Therefore, employing an adaptive weight, α_{ACD}, to adjust the impact of contextual knowledge based on entropy is crucial for improving overall performance.

5 Conclusion

In this work, we mainly tackle handling noisy contexts in open-domain QA on the RAG framework. Our proposed method, ACD, dynamically adjusts contextual influence during decoding by quantifying the model's uncertainty that is either reduced or increased by the retrieved context. Our results show that ACD improves performances across various dimensions by considering the LLM's parametric knowledge and context noisiness. These findings highlight ACD's potential to enhance the reliability of retrieval-augmented generation.

Limitations

Similar to other contrastive decoding approaches, the inference cost of our approach is higher than the conventional greedy decoding. Specifically, while CAD incurs twice the inference cost and MICD incurs three times the cost, ACD also incurs twice the inference cost of conventional greedy decoding.

Our research is limited the base models and does not encompass chat or instruction-following models trained with reinforcement learning from human feedback (RLHF) or instruction fine-tuning (Ouyang et al., 2022; Chung et al., 2022). These aligned models often generate token distributions that vary significantly based on the presence or absence of contextual instruction or templates. For instance, an instruction-following model might start its generation with *"According to the given context ..."* when context is provided, while directly generating the answer in absence of context. This alignment with the provided instructions poses another challenge to be tackled when the contrastive decoding approach is utilized.

Our current focus is primarily on short-form QA tasks. Expanding to QA tasks with long-form generation will enable a wider range of applications. Under long-form QA tasks, our approach can be further developed to investigate scenarios where the context is only partially relevant to the question.

Acknowledgement

This work was partly supported by SNU-NAVER Hyperscale AI Center and 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)]

References

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*.

Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023a. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 41–46, Toronto, Canada. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023b. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *Preprint*, arXiv:2310.11511.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Zhenyu He, Zexuan Zhong, Tianle Cai, Jason D. Lee, and Di He. 2024. Rest: Retrieval-based speculative decoding. *Preprint*, arXiv:2311.08252.

Giwon Hong, Aryo Pradipta Gema, Rohit Saxena, Xiaotang Du, Ping Nie, Yu Zhao, Laura Perez-Beltrachini, Max Ryabinin, Xuanli He, Clémentine Fourrier, and Pasquale Minervini. 2024. The hallucinations leaderboard – an open effort to measure hallucinations in large language models. *Preprint*, arXiv:2404.05904.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023. Look before you leap: An exploratory study of uncertainty measurement for large language models. *Preprint*, arXiv:2307.10236.

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.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud,

Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *Preprint*, arXiv:1705.03551.

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*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *Preprint*, arXiv:2302.09664.

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

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. 2022. Entity-based knowledge conflicts in question answering. *Preprint*, arXiv:2109.05052.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2023. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *Preprint*, arXiv:2305.13169.

Nikolay Malkin, Zhen Wang, and Nebojsa Jojic. 2022. Coherence boosting: When your pretrained language model is not paying enough attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8214–8236, Dublin, Ireland. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*.

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.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. 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 Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *Preprint*, arXiv:1704.04368.

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

Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Gergely Szilvasy, Rich James, Xi Victoria Lin, Noah A. Smith, Luke Zettlemoyer, Scott Yih, and Mike Lewis. 2024. In-context pretraining: Language modeling beyond document boundaries. *Preprint*, arXiv:2310.10638.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller,

Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Yuhao Wang, Ruiyang Ren, Junyi Li, Wayne Xin Zhao, Jing Liu, and Ji-Rong Wen. 2024. Rear: A relevance-aware retrieval-augmented framework for open-domain question answering. *Preprint*, arXiv:2402.17497.

Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024. How easily do irrelevant inputs skew the responses of large language models? *Preprint*, arXiv:2404.03302.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. *Preprint*, arXiv:2310.01558.

Tian Yu, Shaolei Zhang, and Yang Feng. 2024. Truth-aware context selection: Mitigating hallucinations of large language models being misled by untruthful contexts. *Preprint*, arXiv:2403.07556.

Zihan Zhang, Meng Fang, and Ling Chen. 2024. Retrievalqa: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. *Preprint*, arXiv:2402.16457.

Zheng Zhao, Emilio Monti, Jens Lehmann, and Haytham Assem. 2024. Enhancing contextual understanding in large language models through contrastive decoding. *Preprint*, arXiv:2405.02750.

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.

```
Answer the following questions:

<few-shots>

Question: <question>
Answer:
```

Table 3: Template used in closed-book generation.

```
Answer the following questions:

<few-shots>

Context: <context>
Question: <question>
Answer:
```

Table 4: Template used in open-book generation.

Appendix

A Implementation Details

A.1 Instructions

The templates we use throughout the experiment are in Table 3 and Table 4. The template used in open-book generation (Table 4) is applied to get context-augmented distribution $\mathbf{z}_t^c$. Also, to obtain $\mathbf{z}_t$, the template in Table 3 is used.

A.2 Datasets

For NQ and TriviaQA, general world knowledge is required to answer the given question. In PopQA, tackling long-tailed information, less popular factual knowledge is asked. For NQ and TriviaQA, few-shot examples are adopted from train data. For PopQA, we randomly sample 5 examples with different relationship types for sample diversity. The number of test data in used is 3,610 for NQ, 11,313 for TriviaQA, and 14,262 for PopQA.

A.3 Baselines

Baselines using regular greedy decoding are evaluated under two different settings. In the closed-book setting, only the question is provided. In the open-book setting, the retrieved context is employed. The same top-1 retrieved context is utilized for every baseline and ACD.

CAD introduces a context-aware contrastive decoding approach that employs a contrastive output distribution to accentuate discrepancies in model predictions with and without context. This

	R@1	R@5	R@10	R@20	R@100
NQ	38.81	65.65	73.91	79.56	88.01
TriviaQA	49.60	71.32	76.72	80.39	85.71
PopQA	41.83	61.54	68.63	74.55	83.95

Table 5: Recall@100 performance for CONTRIEVER-MSMARCO

method effectively overrides model priors conflicting with provided context, offering significant performance enhancements in tasks requiring resolution of knowledge conflicts. MICD further enhances context grounded generation by integrating contrastive decoding with adversarial irrelevant passages. From a computational time perspective, MICD requires three times more than conventional greedy decoding, while CAD and ACD require twice as much.

MICD proposes two usage directions, referred to as MICD$_F$ and MICD$_D$, which offer fixed and dynamic α, respectively. MICD$_D$ determines α in use by comparing the highest token probabilities with and without given context. Throughout the experiments, fixed value of α is set to the value used in Zhao et al. (2024), 0.5 and 1.0 for CAD and MICD$_F$, respectively.

A.4 Retriever Performance

To assess performance in the RAG framework, the top-1 context from top-100 contexts retrieved by CONTRIEVER-MSMARCO (Izacard et al., 2022) is utilized. Recall@100 is reported for each dataset in Table 5.

A.5 Knowledge Conflict

For the NQ-swap dataset, we utilize the questions and entity-swapped contexts provided in Hong et al. (2024), which includes 3,650 samples. This total excludes 5 few-shot samples and those with contexts presented in a tabular format due to the limited context length. In the case of NQ-swap, each data point has a given context. Since it is a task that does not use a retriever, for MICD, we use the fixed negative context taken from the MICD as an adversarial context. MICD reports that the performance difference between fixed negative and the most distant context is negligible.

	NQ	TriviaQA	PopQA
LLAMA2-7B			
α_{ACD}	32.91	64.85	41.29
α_{oracle}	35.35 (+2.44)	65.31 (+0.46)	44.10 (+2.81)
LLAMA2-13B			
α_{ACD}	36.12	67.37	43.35
α_{oracle}	38.75 (+2.63)	68.19 (+0.82)	47.01 (+3.66)
LLAMA3 8B			
α_{ACD}	35.48	66.32	43.25
α_{oracle}	36.98 (+1.50)	66.10 (-0.22)	46.47 (+3.22)
MISTRAL 7B			
α_{ACD}	35.37	67.82	41.47
α_{oracle}	38.37 (+3.00)	67.29 (-0.53)	44.53 (+3.06)

Table 6: EM score comparison between ACD (α_{ACD}) and ACD with oracle alpha value (α_{oracle}).

B Results

B.1 Results on *Known-noisy* and *Unknown-gold*

For *Known-noisy* and *Unknown-gold*, the exact values of EM accuracy on each case are reported in Table 8 and Table 9, respectively.

B.2 AUROC between Adaptive Weight and Context Noisiness

AUROC of ACD and MICD$_D$ for three models not reported in Table 2 is reported in Table 10.

C Additional Analysis

C.1 Upper-bound of Alpha

In our approach, the parameter α is expected to be close to 1 when the retrieved context contains information that helps answer the given question, and close to 0 otherwise. To evaluate the upper-bound performance of ACD, we assume that we have prior knowledge of whether the context in use is gold or noisy. Under this assumption, we fix the α value to 1.0 if the context is gold and to 0.0 if the context is noisy.

For TriviaQA dataset, the performance of ACD is comparable to α_{oracle}, with less than 1 point difference (Table 6). NQ and PopQA show a difference of approximately 2-3 points, indicating that the method for calculating the α weight could be further enhanced in future research.

C.2 Case Study

We conduct the case study on α_{ACD}, examining its value in cases of *Known-noisy* and *Unknown-gold*. Table 7 shows the generations from LLAMA2

Case	Sample	Reg_{Cls}		Reg_{Opn}		ACD	
		Generation	$H(Y_t)$	Generation	$H(Y_t^c)$	Generation	α_{ACD}
Known-noisy	Question: who does the voice of nala in the lion king? Gold answer: Moira Kelly	Moira Kelly	2.9160	Whoopi Goldberg	5.4562	Moira Kelly	0.3483
Unknown-gold	Question: who was the actor that played ben stone on law and order? Gold answer: Michael Moriarty	Michael Tucker	6.6748	Michael Moriarty	1.5628	Michael Moriarty	0.8103

Table 7: Case study on the value of α_{ACD} for *Known-noisy* and *Unknown-gold* cases in LLAMA2 7B. Each value of entropy without context ($H(Y_t)$), entropy with context ($H(Y_t^c)$), and α_{ACD} is extracted at the first decoding step ($t = 0$).

7B and how the values of entropy from closed-book generation (Reg_{Cls}) and open-book generation (Reg_{Opn}) affect α_{ACD} at the first decoding time step.

In the case of *Known-noisy*, when the model generates the answer correctly even without the given context, the retrieved noisy context yields relatively higher entropy, resulting in α_{ACD} value of 0.3483. Conversely, in the case of *Unknown-gold*, the model's generated answer is incorrect, aligning with a relatively high entropy value of 6.6748. In this scenario, the retrieved gold context guides the model to correctly answer the question, which is reflected in a relatively lower entropy value of 1.5628. Thus, the value of α_{ACD}, adjusted with these entropy values, yields a relatively higher weight on the context at 0.8103.

	NQ	TriviaQA	PopQA
LLAMA2-7B			
Reg_{Opn}	45.13	68.12	33.47
CAD	29.22	48.91	25.97
MICD_F	51.07	72.37	36.81
MICD_D	72.92	86.33	**56.04**
ACD	**76.72**	**88.79**	54.58
LLAMA2-13B			
Reg_{Opn}	47.18	69.77	32.53
CAD	32.04	52.48	22.66
MICD_F	54.17	75.05	38.55
MICD_D	**76.31**	88.24	**59.38**
ACD	75.15	**88.78**	56.11
LLAMA3-8B			
Reg_{Opn}	46.20	68.50	33.39
CAD	32.91	50.51	23.00
MICD_F	43.25	70.67	39.07
MICD_D	61.18	83.70	**59.80**
ACD	**64.14**	**86.59**	56.87
MISTRAL-7B			
Reg_{Opn}	41.04	64.57	31.03
CAD	19.17	42.63	20.80
MICD_F	48.12	71.99	36.48
MICD_D	69.58	86.84	**57.14**
ACD	**70.62**	**89.36**	53.55

Table 8: EM accuracy of *Known-noisy* case.

	NQ	TriviaQA	PopQA
LLAMA2-7B			
Reg_{Opn}	**47.78**	**68.18**	**74.42**
CAD	43.90	62.22	66.12
MICD_F	40.47	61.51	64.63
MICD_D	29.82	50.43	65.17
ACD	36.03	57.10	73.41
LLAMA2-13B			
Reg_{Opn}	**46.52**	**65.09**	**75.04**
CAD	45.77	61.07	69.43
MICD_F	41.79	62.03	65.85
MICD_D	30.72	47.77	64.70
ACD	36.19	53.98	72.38
LLAMA3-8B			
Reg_{Opn}	**48.12**	**68.47**	**74.10**
CAD	48.00	60.52	70.41
MICD_F	38.15	61.40	65.55
MICD_D	33.33	48.45	64.81
ACD	41.67	61.24	72.52
MISTRAL-7B			
Reg_{Opn}	**49.57**	**64.82**	**73.09**
CAD	45.38	56.02	67.81
MICD_F	43.28	63.59	66.58
MICD_D	32.06	54.97	66.37
ACD	37.73	57.70	73.03

Table 9: EM accuracy of *Unknown-gold* case.

	α	NQ	TriviaQA	PopQA
LLAMA2 13B				
Max	MICD_D	52.77	60.09	61.84
	ACD	**69.24**	**75.31**	**74.12**
Avg.	MICD_D	57.86	62.00	71.79
	ACD	**71.61**	**73.41**	**77.92**
First	MICD_D	54.80	46.13	68.44
	ACD	**73.07**	**77.96**	**80.51**
LLAMA3 8B				
Max	MICD_D	50.75	52.59	63.72
	ACD	**63.12**	**57.82**	**75.00**
Avg.	MICD_D	51.80	52.83	67.99
	ACD	**64.08**	**59.67**	**75.90**
First	MICD_D	45.70	39.07	69.21
	ACD	**67.48**	**75.45**	**80.31**
MISTRAL 7B				
Max	MICD_D	56.98	64.95	61.93
	ACD	**71.27**	**77.46**	**74.11**
Avg.	MICD_D	63.66	69.27	73.82
	ACD	**76.02**	**78.20**	**79.08**
First	MICD_D	56.84	68.98	71.73
	ACD	**75.75**	**84.11**	**82.07**

Table 10: AUROC between α used in each method and the noisiness of the retrieved context. The best AUROC is in bold.