Enhancing Contextual Understanding in Large Language Models through Contrastive Decoding

Zheng Zhao[1,*] **Emilio Monti**[2] **Jens Lehmann**[2] **Haytham Assem**[2]
[1]School of Informatics, University of Edinburgh
[2]Amazon
zheng.zhao@ed.ac.uk, {monti,jlehmnn,hithsala}@amazon.com

Abstract

Large language models (LLMs) tend to inadequately integrate input context during text generation, relying excessively on encoded prior knowledge in model parameters, potentially resulting in generated text with factual inconsistencies or contextually unfaithful content. LLMs utilize two primary knowledge sources: 1) prior (parametric) knowledge from pretraining, and 2) contextual (non-parametric) knowledge from input prompts. The study addresses the open question of how LLMs effectively balance these knowledge sources during the generation process, specifically in the context of open-domain question answering. To address this issue, we introduce a novel approach integrating contrastive decoding with adversarial irrelevant passages as negative samples to enhance robust context grounding during generation. Notably, our method operates at inference time without requiring further training. We conduct comprehensive experiments to demonstrate its applicability and effectiveness, providing empirical evidence showcasing its superiority over existing methodologies.

1 Introduction

Improving large language models (LLMs) has been a primary focus in natural language processing research. Recent strides have incorporated retrieval mechanisms to enhance LLMs (Lewis et al., 2020; Guu et al., 2020; Izacard and Grave, 2021; Izacard et al., 2023), augmenting their ability to produce contextually relevant and precise responses (Min et al., 2023; Mallen et al., 2023). Retrieval-augmented LLMs, which leverage both *parametric* knowledge acquired during training and *non-parametric* knowledge retrieved during inference, exhibit potential in addressing challenges such as limited memorization (Kandpal et al., 2023), knowledge conflicts (Longpre et al., 2021), and outdated information (Kasai et al., 2022).

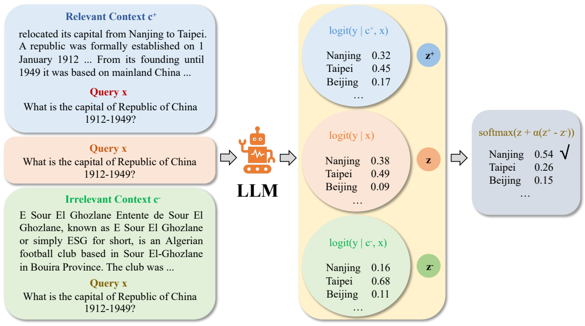

Figure 1: An illustration of our proposed decoding method. Despite the relevant context suggesting the answer as "Nanjing", it contradicts the LLM's prior knowledge. After reconciling different knowledge sources, the model correctly predicted the answer by boosting Nanjing's plausibility and reducing Taipei's likelihood. This decision was based on considering Nanjing to be less likely given the irrelevant context, while Taipei is deemed more probable.

An ongoing question pertains to how LLMs ought to balance these two knowledge sources during generation. Previous research suggests that LLMs can falter in adequately attending to newly introduced information within the contextual knowledge. To tackle this issue, context-aware decoding (CAD; Shi et al., 2023a) has been proposed. By employing a contrastive output distribution, CAD highlights discrepancies in output probabilities when the model operates with and without context. Their experiments illustrate CAD's effectiveness in overriding the model's parametric knowledge in cases of conflict with provided context. However, while prior works often assert context as inherently reliable, our perspective argues that LLMs should possess the capacity to navigate and reconcile both parametric and non-parametric knowledge, ultimately refining their ability to strike a judicious balance. This paper undertakes the development and assessment of a novel decoding strategy tailored for retrieval-augmented LLMs, seeking equilibrium in utilizing parametric and

[*]Work done during an internship at Amazon.

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
June 16-21, 2024

non-parametric knowledge sources. The proposed method involves a contrastive decoding approach (Li et al., 2023), integrating both relevant and irrelevant contexts, wherein the irrelevant context can be adversarially crafted retrieval or bottom-ranked retrieved text. Notably, we emphasize the criticality of leveraging irrelevant contexts, a distinguishing feature of our approach, with the expectation that the model will diverge from incorrect responses.

We conducted extensive experiments on diverse datasets like Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and PopQA (Mallen et al., 2023). We employed a range of vanilla LLMs, including OPT (Zhang et al., 2022), Falcon (Almazrouei et al., 2023), LLaMA families (Touvron et al., 2023a,b), and instruction-tuned Flan-T5 (Chung et al., 2022). Through this comprehensive evaluation, we provide empirical evidence supporting the superiority of incorporating irrelevant contexts in assisting LLMs to manage knowledge conflicts and seamlessly integrate contexts for generating responses in open-domain question answering against conventional decoding approaches without necessitating further fine-tuning. We also explore the impact of different retrieval sources on the decoding strategy, emphasizing the importance of refining retrieval mechanisms for further enhancements in performance.

Additionally, the paper explores different facets of the proposed decoding approach, including the impact of various hyperparameters, the effect of scaling model sizes, and the selection of irrelevant contexts. This exploration provides deeper insights into leveraging parametric and non-parametric knowledge sources. We demonstrate that although our approach outperforms regular decoding across most model sizes, it particularly excels with larger models. Moreover, we show our method's effectiveness even with simple fixed irrelevant contexts. Additionally, our approach exhibits consistent performance improvements in answering questions with knowledge across varying levels of popularity. Beyond benchmarking against existing methods, this study also explores practical implications and constraints of the proposed decoding strategy, delineating pathways for future research in generative tasks beyond question answering.

2 Related Works

Retrieval-augmented LLMs While LLMs relying solely on their parameters can capture extensive world knowledge, they exhibit limited memorization for less frequent entities (Kandpal et al., 2023), susceptibility to hallucinations (Shuster et al., 2021), and temporal degradation (Luu et al., 2022; Jang et al., 2022). Furthermore, the acquired parametric knowledge swiftly becomes outdated (Kasai et al., 2022). Recent research emphasizes the enhancement of LLMs with non-parametric memories, referred to as retrieved text chunks, enabling smaller models to match the performance of larger counterparts (Izacard et al., 2023). Studies exploring the integration of retrieved non-parametric memories within intermediate states or output spaces have shown effectiveness in overcoming LLM limitations in memorization and knowledge updating (Zhong et al., 2022; Min et al., 2023). Mallen et al. (2023) extensively analyze the circumstances favoring the benefits of retrieval augmentation. They demonstrate its efficacy in less frequent occurrences but caution about potential misguidance for LLMs. Building upon these insights, they introduce adaptive retrieval and empirically showcase its promising effectiveness.

Knowledge Conflicts In cases of conflicting knowledge in updated documents, language models are expected to generate responses based on provided contexts rather than relying solely on outdated parametric knowledge. Retrieval-augmented LLMs (Min et al., 2023; Shi et al., 2023b; Izacard et al., 2023) particularly benefit from this scenario by employing externally retrieved documents to enrich their knowledge. However, the mere addition of documents doesn't consistently influence model predictions, as current LLMs often overlook contexts and heavily rely on prior parametric knowledge (Longpre et al., 2021; Chen et al., 2022). Zhou et al. (2023) aim to improve a model's fidelity to context using prompting-based method, but are constrained to large-scale instruction-finetuned LLMs like OpenAI's gpt-3.5-turbo-instruct. Zhang et al. (2023) address how to combine retrieved and parametric knowledge to get the best of both worlds for open-domain QA, but their method requires further training discriminators with silver labels. In contrast, our work investigates a decoding strategy applicable to any LLMs without any training.

Contrastive Decoding The exploration of contrastive decoding methods extensively addresses text generation. MMI-based decoding (Li et al., 2016) utilizes a contrastive formulation to enhance output diversity in dialog generation. DExperts

(Liu et al., 2021) dampens the output distribution of an anti-expert (e.g., exposed to toxic language) to guide generations away from undesired attributes. Contrastive decoding (Li et al., 2023) demotes an amateur model (e.g., models with minimal parameters) to distill expert knowledge from larger, competitive models. Pozzobon et al. (2023) introduce an innovative toxicity mitigation approach that contrasts and ensembles the next token probabilities obtained from a LLM using both toxic and non-toxic retrievals. Context-aware decoding (Shi et al., 2023a) emphasizes output probability differences using a contrastive ensemble between model predictions with and without non-parametric knowledge. It effectively overrides a model's parametric knowledge when it conflicts with the provided non-parametric information. While our work builds upon the concept of context-aware decoding, one key distinction lies in the integration of irrelevant context. Unlike Shi et al.'s approach, which focuses solely on relevant non-parametric knowledge, our method incorporates potentially irrelevant non-parametric knowledge into the inference process with the expectation that the model will deviate from incorrect responses.

3 Methodology

3.1 Problem Statement

We consider decoding approaches for open-domain question answering, where the large language model θ receives an input query $\boldsymbol{x}$ and aim to generate a faithful answer $\boldsymbol{y}$. During the generation of $\boldsymbol{y}_t$ at each time step t, the language model computes the logits $\mathbf{z}_t \in \mathbb{R}^{|V|}$ for the t-th token, where V represents the vocabulary. The probability distribution over the vocabulary is derived by normalizing and exponentiating $\boldsymbol{z}_t$ as follows:

$$p_\theta(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}) = \mathrm{softmax}(\mathbf{z}_t).$$

Prompting the model for its parametric knowledge involves sampling the response from the probability distribution conditioned on the query $\boldsymbol{x}$ and the previously generated response $\boldsymbol{y}_{<t}$:

$$y_t \sim p_\theta(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}).$$

Similarly, when incorporating additional context $\boldsymbol{c}$, containing external knowledge beyond the model's parametric knowledge, our model θ generates a response $\boldsymbol{y}$ considering the query, context, and the previously generated response:

$$y_t \sim p_\theta(y_t|\boldsymbol{c}, \boldsymbol{x}, \boldsymbol{y}_{<t}).$$

We observe two sources of knowledge (parametric vs. non-parametric) contributing to model responses, which may sometimes conflict (Longpre et al., 2021; Neeman et al., 2023). While some argue for prioritizing non-parametric knowledge over potentially outdated parametric knowledge (Shi et al., 2023a), we propose the importance of striking a balance between these sources as non-parametric knowledge, derived from external retrievers, may also contain inaccuracies.

3.2 Multi-Input Contrastive Decoding

Context can be both beneficial and problematic. Thus, we segregate context $\boldsymbol{c}$ into relevant $\boldsymbol{c}^+$ and irrelevant $\boldsymbol{c}^-$. At each decoding time step t, our approach combines the model's prediction based on its parametric knowledge ($\mathbf{z}_t$) with predictions utilizing relevant ($\mathbf{z}_t^+$) and irrelevant ($\mathbf{z}_t^-$) contexts:

$$y_t \sim \mathrm{softmax}(\mathbf{z}_t + \alpha(\mathbf{z}_t^+ - \mathbf{z}_t^-)),$$

where α is a hyperparameter that governs the extent of modification to the parametric answer ($\mathbf{z}_t$). Equivalently,

$$\begin{aligned} y_t &\sim \tilde{p}_\theta(y_t|\boldsymbol{c}^+, \boldsymbol{c}^-, \boldsymbol{x}, \boldsymbol{y}_{<t}) \\ &\propto p_\theta(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}) \left(\frac{p_\theta(y_t|\boldsymbol{c}^+, \boldsymbol{x}, \boldsymbol{y}_{<t})}{p_\theta(y_t|\boldsymbol{c}^-, \boldsymbol{x}, \boldsymbol{y}_{<t})}\right)^\alpha. \end{aligned}$$

In essence, a response will exhibit high probability only if it holds high likelihood under both learned parametric knowledge and relevant non-parametric knowledge, while demonstrating low probability under irrelevant non-parametric knowledge. The ratio $\frac{p_\theta(y_t|\boldsymbol{c}^+, \boldsymbol{x}, \boldsymbol{y}_{<t})}{p_\theta(y_t|\boldsymbol{c}^-, \boldsymbol{x}, \boldsymbol{y}_{<t})}$ functions as a scaling factor used to modify the parametric answer for the given input query. A larger α implies a greater modification, with $\alpha = 0$ resulting in no modification, indicating regular decoding using solely parametric knowledge without additional context.

Fundamentally, our proposed decoding operates as an ensemble involving the logits $\mathbf{z}_t$, $\mathbf{z}_t^+$, and $\mathbf{z}_t^-$. A similar ensemble approach has been explored in Liu et al. (2021) and Li et al. (2023) for controllable and open-ended text generation, though their ensembles are based on predictions from different models. Another similar work to ours is CAD (Shi et al., 2023a), which examines scenarios where the model's parametric knowledge contradicts non-parametric knowledge. CAD essentially constitutes a contrastive ensemble between $\mathbf{z}_t$ and $\mathbf{z}_t^+$. In this study, we concentrate on the general case of open-domain question answering, proposing a dynamic adjustment of α, controlling the degree of modification without treating it as a fixed hyperparameter. We provide an illustration of our method in Figure 1.

Dynamic α In prior logit adjustment methods (Liu et al., 2021; Malkin et al., 2022; O'Neill et al., 2023; Shi et al., 2023a; Pozzobon et al., 2023), α remains a fixed hyperparameter, requiring exhaustive search within the parameter space. Our innovation lies in dynamically setting α at each time step t without supervision, enabling fine-grained token-level adjustments. We estimate LLM confidence following Jiang et al. (2021) by computing the highest probability from the normalized predicted token probabilities at each step:

$$C = \max_{y' \in V} P_\theta(y'|\boldsymbol{x}, \boldsymbol{y}_{<t}).$$

Similarly, we estimate LLM confidence using relevant non-parametric knowledge $\boldsymbol{c}^+$:

$$C_R = \max_{y' \in V} P_\theta(y'|\boldsymbol{c}^+, \boldsymbol{x}, \boldsymbol{y}_{<t}).$$

At each time step, the value of α is determined as follows:

$$\alpha = \begin{cases} 1 - C, & \text{if } C > C_R, \\ C_R, & \text{otherwise.} \end{cases}$$

Our rationale is that higher LLM confidence in parametric knowledge warrants minor adjustments, while greater confidence in relevant non-parametric knowledge necessitates more substantial modifications to the parametric answer. Note that we use $1 - C$ instead of using $C - C_R$ to avoid the case where both C and C_R are low. In such case, a larger modification is still desired.

Selection of $\boldsymbol{c}^+$ and $\boldsymbol{c}^-$ Choosing relevant context $\boldsymbol{c}^+$ is straightforward and we follow the retrieval-augmented LLM literature where we use the top retrieved texts from a retrieval module by running our input query over an external knowledge base. However, selecting irrelevant context $\boldsymbol{c}^-$ is not trivial. Potential methods include using lower-ranked retrievals, random text, or even deliberately crafted adversarial text. The primary aim of $\boldsymbol{c}^-$ is to provide adversarial knowledge to elicit incorrect predictions that can be disregarded from the final token distribution. We explore various strategies for selecting $\boldsymbol{c}^-$ in Section 5.3.

4 Experimental Setup

The present study revolves around open-domain question answering, which involves tasking models to generate responses to factual questions in natural language. Specifically, we concentrate on the *open-book* QA setting (Roberts et al., 2020), where we harness non-parametric knowledge by supplying relevant contexts along with the question itself to the model during inference. Consistent with prior investigations, we utilize prompting techniques to assess the models' performance.

4.1 Datasets and Metrics

Datasets Our method undergoes evaluation using three popular QA benchmarks: TriviaQA (Joshi et al., 2017), Natural Questions (NQ; Kwiatkowski et al. 2019), and PopQA (Mallen et al., 2023). TriviaQA comprises trivia questions sourced from the Web, whereas NQ consists of questions derived from actual Google search queries, with answer spans located in Wikipedia articles identified by annotators. PopQA is a novel entity-centric open-domain QA dataset covering factual information about entities across a spectrum of popularity, including *long-tail* knowledge often overlooked in other popular QA datasets.

Metrics In line with prior research, our primary metric for evaluating performance is the exact match (EM), which determines whether the predicted sequence matches precisely with one of the correct answers provided within the dataset.

4.2 Baselines and Models

Baselines Baseline approaches include regular decoding with greedy decoding, following prior work (Izacard and Grave, 2021). We prompt the model for an answer by providing contextual information. While our primary focus remains on the *open-book* QA setting, we also present a baseline employing the *closed-book* QA setting, where the prompt consists solely of questions. This exploration aims to scrutinize the parametric knowledge of LLM. Additionally, we compare our method to CAD, which accentuates the difference in output probabilities when employing a model with and without context.

Models Our decoding method undergoes evaluation across models varying in scale: Flan-T5 (XL-3B, XXL-11B; Chung et al. 2022), Falcon (7B, 40B; Almazrouei et al. 2023), OPT (6.7B, 13B, 30B, 66B; Zhang et al. 2022), Llama (7B, 13B, 33B, 65B; Touvron et al. 2023a), and Llama 2 (7B, 13B, 70B; Touvron et al. 2023b), without additional fine-tuning.

Instructions We employ a straightforward template, i.e., "`Answer the following question. Question: <question> Answer:`", to format all questions for generative prediction in the closed-book setting. For the open-book setting, the template becomes "`Answer the question based on the context below. Context: <context> Question: <question> Answer:`". Although more sophisticated prompts were trialed in preliminary experiments, their marginal improvement over the simple template did not warrant their use, especially considering the risk of overfitting the model. In alignment with prior work (Chung et al., 2022), we employ 5-shot prompting for all models.

Retrieval models As previously mentioned, we explore a retrieval-augmented LLM approach in the open-book setting. This involves running an off-the-shelf retrieval system offline to obtain relevant context from Wikipedia for each query[1], which is then concatenated with the original query. We utilize two widely-used retrieval systems: BM25 (Robertson and Zaragoza, 2009) and Contriever (Izacard et al., 2022). BM25 operates as a static term-based retriever without training, while Contriever is pre-trained on extensive unlabeled corpora. In this study, we leverage Contriever-MS MARCO, a Contriever fine-tuned on MS MARCO (Bajaj et al., 2018). Consistent with Mallen et al. (2023), we utilize the top one retrieved paragraph. Additionally, TriviaQA and NQ datasets provide gold contexts, which we employ to measure the theoretical upper bound of our proposed decoding method. We also investigate the impact of using different retrieval methods in Section 5.2.

Setting alpha Our approach introduces a hyperparameter α to govern the degree of modification atop LLM's parametric knowledge. For CAD, after a grid search using the validation set, we set $\alpha = 0.5$. In fixed alpha experiments for our method, we set $\alpha = 1.0$. In dynamic alpha experiments, we do not have to set alpha values explicitly. We explore the effect of α on our method in Section 5.4.

[1]We utilize the Wikipedia dump from 2018.

Model	Decoding	NQ	TQA	PopQA
Flan-T5 11B	Reg.-Cl.	14.82	40.5	13.98
	Reg.-Op.	57.84	79.36	31.16
	CAD	47.56	66.08	26.28
	Ours-F	59.58	76.75	31.37
	Ours-D	**63.16**	**80.09**	**34.64**
Falcon 40B	Reg.-Cl.	28.56	71.74	28.79
	Reg.-Op.	53.32	72.05	39.16
	CAD	49.36	20.72	35.31
	Ours-F	**53.77**	79.56	**39.87**
	Ours-D	50.53	**80.73**	38.28
OPT 66B	Reg.-Cl.	13.71	39.65	15.62
	Reg.-Op.	48.73	62.38	34.77
	CAD	45.93	24.51	33.45
	Ours-F	**51.97**	**68.11**	**34.83**
	Ours-D	44.41	63.89	33.44
Llama 65B	Reg.-Cl.	34.13	75.72	35.9
	Reg.-Op.	55.32	74.76	40.31
	CAD	48.03	24.51	31.97
	Ours-F	**57.01**	76.61	39.9
	Ours-D	52.35	**80.28**	**40.58**
Llama-2 70B	Reg.-Cl.	37.87	79.69	40.98
	Reg.-Op.	56.07	76.07	42.7
	CAD	47.53	31.36	33.05
	Ours-F	**58.86**	78.38	42.59
	Ours-D	55.24	**81.7**	**44.3**

Table 1: Results of models using gold retrieval (NQ, TriviaQA), and Contriever retrieval (PopQA). Reg.-Cl. refers to regular decoding with *closed-book* setting (i.e. no retrieval). Reg.-Op. refers to regular decoding with *open-book* setting (i.e. with retrieval). Ours-F refers to our method utilizing a fixed alpha, while Ours-D designates our method incorporating a dynamic alpha.

5 Results

We present the results of models featuring the largest variants in Table 1. Notably, employing regular decoding within an open-book setting consistently outperforms the closed-book setting across most models. This inclination suggests that LLM systems require non-parametric knowledge to excel in tasks demanding substantial knowledge assimilation. Interestingly, the performance of Llama 65B and Llama 2 70B in the closed-book setting surpasses that in the open-book setting concerning TriviaQA, indicating these models' proficiency in factual knowledge retention without resorting to non-parametric knowledge. This finding possibly implies that TriviaQA, being the oldest dataset among the three, potentially overlaps with the train-

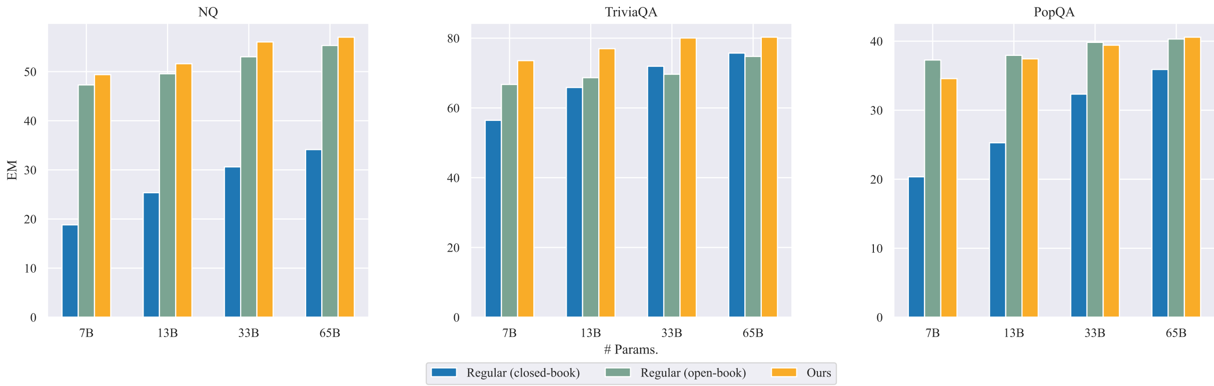

Figure 2: Performance comparison of our method against regular decoding across various sizes of Llama 1 models.

ing data of these LLMs.

Crucially, our proposed decoding approach demonstrates superior performance across all three datasets compared to both regular decoding and CAD.[2] Noteworthy variations exist in the efficacy of employing either the fixed alpha strategy or the dynamic alpha strategy; while in certain instances the fixed alpha approach exhibits better performance, the dynamic alpha approach outperforms in others. In subsequent references within this paper, when mentioning our method, we refer to the setting that delivers superior performance based on Table 1, without explicitly specifying whether it involves dynamic or fixed alpha.

5.1 Effect of Model Scaling

Thus far, our study has elucidated the efficacy of our proposed decoding approach across diverse model families. This segment aims to examine the impact of scaling the model's parameter count on our methods. The results pertaining to Llama variants–specifically, Llama 7B, 13B, 33B, and 65B–are illustrated in Figure 2. We provide the results of scaling for other models in Appendix A. An observable trend emerges wherein, with an increase in model size, the disparity between closed-book and open-book performance diminishes, indicating that larger models possess greater potential for assimilating parametric knowledge. Furthermore, our decoding method consistently outperforms regular decoding across all model sizes, except for a few instances in the case of PopQA with smaller model variants. We posit this discrepancy to the absence of gold context within the PopQA dataset, leading to reliance on Contriever's retrieval, which may occasionally introduce inaccuracies.

[2]The CAD results were based on our implementation, due to the unavailability of the original CAD implementation at the time of our study.

5.2 Using Different Retrievals

As previously highlighted, our investigation centers on retrieval-augmented LLMs, involving the implementation of retrieval modules over a knowledge base concerning a user query. Subsequently, the retrieved relevant passage supplements the prompt to facilitate the generation of answers by the LLM. In earlier experiments, we utilized the provided gold context by NQ and TriviaQA to establish the theoretical upper bound of our proposed decoding method. This segment aims to examine whether the utilization of off-the-shelf retrieval mechanisms would influence the efficacy of our proposed methods. In Figure 3, we present a comparative analysis between closed-book regular decoding and our decoding method, utilizing retrieval passages from BM25, Contriever, or the provided gold context.

It is pertinent to note that the PopQA dataset lacks gold context. The comparative analysis indicates that results derived from Contriever exhibit superiority over those derived from BM25. Moreover, a substantial disparity exists between outcomes obtained through retrieval and those derived from leveraging gold context. It is essential to underscore that while these observations do not negate the efficacy of our proposed decoding method, they do suggest that enhancements to the retrieval module could yield improved outcomes.

5.3 Selection of Irrelevant Context

An essential aspect of our decoding method involves the incorporation of the $\boldsymbol{c}^-$ irrelevant context. Here, we investigate various strategies for selecting $\boldsymbol{c}^-$ and its impact on our methods. Initially, we propose employing a random selection of $\boldsymbol{c}^-$ from the complete pool of available contexts

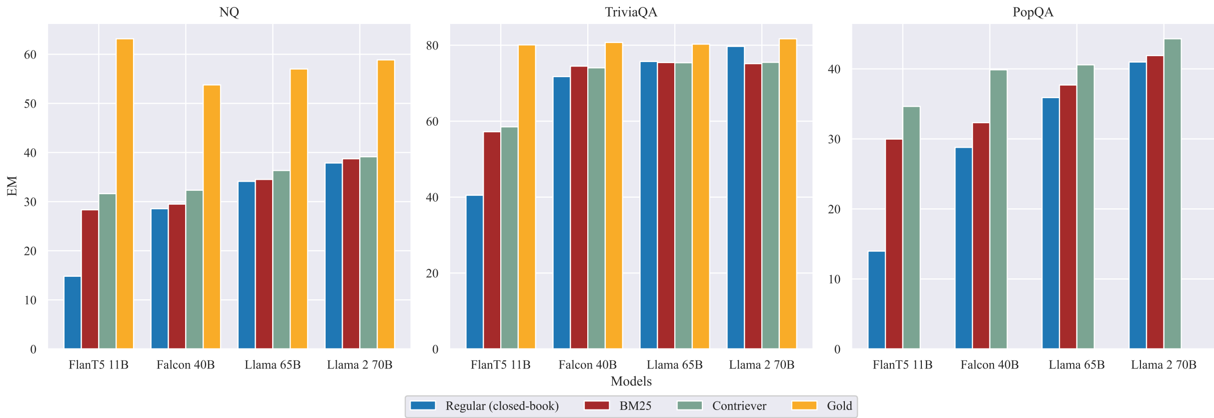

Figure 3: Performance comparison between regular decoding and our method using different sources of retrievals.

Irr. Passage	NQ	TQA	PopQA
Random	56.74	81.28	43.23
Fixed	57.95	80.84	43.82
Fixed (rand. perm.)	57.17	80.68	42.98
Most distant	**58.86**	**81.7**	**44.3**

Table 2: Comparison of performance on Llama 2 70B across various methods for selecting irrelevant $\boldsymbol{c}^-$: random selection, fixed adversarially constructed contexts, fixed context with random word permutation, and passages with the most distance from the relevant context.

(ensuring that the randomly selected $\boldsymbol{c}^-$ differs from $\boldsymbol{c}^+$). Subsequently, we manually construct an adversarial $\boldsymbol{c}^-$ devoid of meaningful or useful information, details of which are provided in Appendix B. Additionally, we experiment with shuffling the word order within this fixed $\boldsymbol{c}^-$. Another approach for determining $\boldsymbol{c}^-$ involves using lower-ranked retrievals. However, increasing the retrieval size arbitrarily is computationally inefficient, and even within the top-100 retrievals, relevant information can be present. Therefore, we approximate the bottom-ranked retrieval by selecting the $\boldsymbol{c}^-$ that exhibits the most distance from $\boldsymbol{c}^+$, based on the cosine distance of their embeddings in the retrieval module. The comparison results using Llama 2 70B are presented in Table 2. It is evident that $\boldsymbol{c}^-$ with the most distance yields the best performance. Throughout our experiments detailed in this study, if not explicitly specified, we employ the most distant option for selecting $\boldsymbol{c}^-$. However, if computing distance proves computationally expensive, the use of a fixed adversarial $\boldsymbol{c}^-$, as demonstrated in our results, remains a viable alternative.

5.4 Adjusting the Knowledge Modification

Our proposed decoding method introduces the hyperparameter α, regulating the degree of modification applied to the parametric answer for a given input query. A larger α signifies a more substantial modification, while $\alpha = 0$ denotes no alteration, thereby reducing decoding to a regular decoding scenario. Despite outlining a strategy to dynamically set this alpha value, we remain interested in assessing the impact of different alpha values on the efficacy of our method. We conducted experiments involving the adjustment of α levels and present the outcomes obtained from Llama models in Figure 4. Our analysis reveals that as the alpha values increase, the effectiveness of the method diminishes substantially. The model achieves optimal performance at $\alpha = 1.0$, outperforming all other alpha settings. Furthermore, setting $\alpha = 1.0$ yields consistent improvements over regular decoding on both the NQ and TriviaQA datasets. For PopQA, while a fixed α value offered no improvement over regular decoding, the dynamic setting we propose led to significant gains, as shown in Table 1.

5.5 Answering across Knowledge Popularity

The utility of retrieval mechanisms becomes evident in addressing less prevalent factual knowledge, an area where LLMs often exhibit limitations. Therefore, we conducted an analysis to evaluate the efficacy of our proposed decoding approach in facilitating LLMs to accurately respond to factual questions across a spectrum of popularity. Following Mallen et al. (2023), we utilized the popularity of entities gauged by Wikipedia's monthly page views as an indicator of their frequency in web discussions. Our findings, presented in Figure 5, juxtapose the performance of models employing reg-

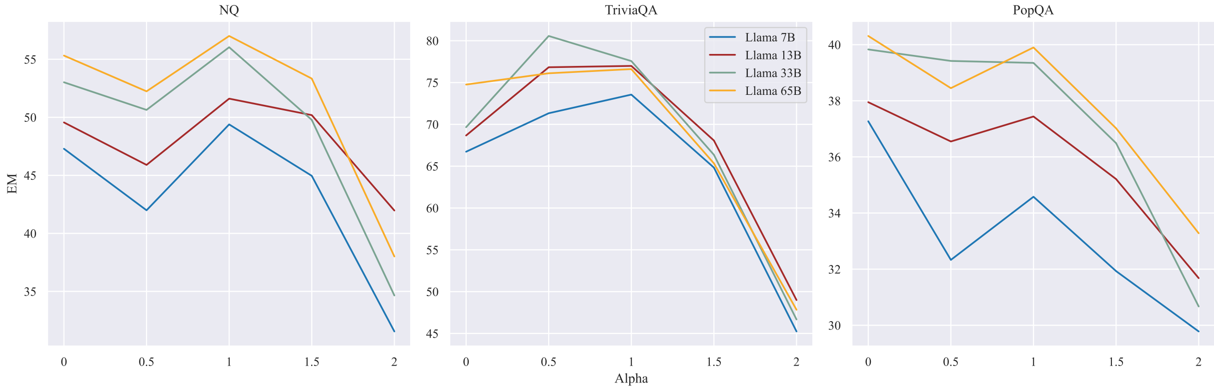

Figure 4: The impact of α on our decoding method across different sizes of Llama 1 models.

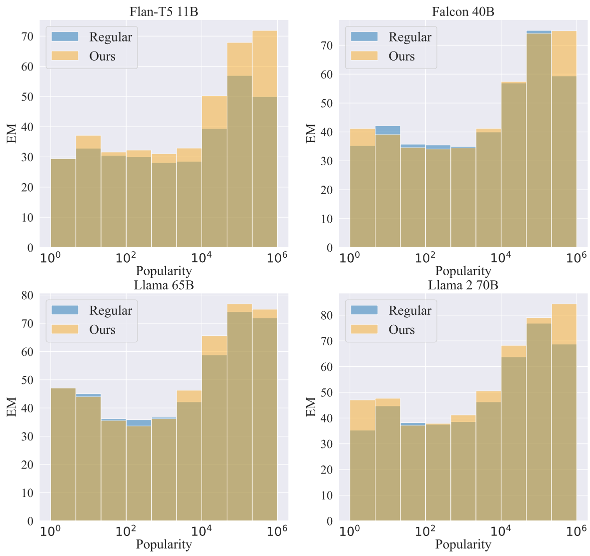

Figure 5: Comparison of performance between regular decoding (open-book) and our method on questions with varying levels of knowledge popularity.

Model	Decoding	NQ-SUB
Flan-T5 11B	Reg.-Cl.	0.19
	Reg.-Op.	56.4
	CAD	51.9
	Ours	**57.55**
Falcon 40B	Reg.-Cl.	0.13
	Reg.-Op.	46.78
	CAD	45.79
	Ours	**48.34**
Llama 65B	Reg.-Cl.	0.08
	Reg.-Op.	59.25
	CAD	60.41
	Ours	**61.65**
Llama-2 70B	Reg.-Cl.	0.02
	Reg.-Op.	57.63
	CAD	53.23
	Ours	**58.34**

Table 3: Comparison of decoding methods on the knowledge conflict dataset. Reg.-Cl. and Reg.-Op. denote regular decoding in closed-book and open-book settings.

ular decoding within an open-book setting against those employing our proposed method. The results manifest a consistent trend wherein our proposed method consistently outperforms regular decoding under an open-book setting across varying levels of popularity. This observation underscores the efficacy of our decoding strategy in assisting LLMs to generate more accurate responses to factual queries across a diverse range of entity popularities.

5.6 Resolving Knowledge Conflicts

As previously highlighted in the manuscript, tasks reliant on knowledge typically draw from two knowledge sources: parametric knowledge, acquired during training, and non-parametric knowledge, accessed via retrieval modules during inference. The issue of knowledge conflicts, wherein the contextual (non-parametric) information contradicts learned knowledge, has been formally addressed by Longpre et al. (2021) to understand how models utilize these dual sources of knowledge.

To generate question-answer pairs manifesting knowledge conflicts, we followed the methodology proposed by Longpre et al. (2021). Initially, we identified questions in the NQ dataset that contained named entity answers. Subsequently, we obtained the relevant context for each question and replaced the gold answer entity in the context with a random entity. In this setup, an accurate LLM should produce the substituted entity as the answer when provided with the question and the modified context, disregarding its pre-learned parametric answer. This resulting dataset, termed NQ-SUB, was created for assessing models in scenarios involving knowledge conflicts. The performance results on NQ-SUB are presented in Table 3. Remarkably, all models exhibited poor performance in the regular closed-book setting, given that the task requires the model to disregard its parametric knowledge. However, our proposed decoding method demonstrated superior performance compared to both regular decoding and CAD on this knowledge conflict task. The comparative results emphasize the effectiveness of our proposed decoding approach in addressing knowledge conflicts, particularly in scenarios where models encounter contradictions between their learned and contextual knowledge.

6 Conclusion

This study introduces a novel decoding strategy, employing contrastive decoding to incorporate relevant and irrelevant context. Through diverse experiments and analyses across datasets and model scales, our approach consistently outperforms regular decoding methods. Notably, it excels in managing knowledge conflicts, surpassing both regular decoding and CAD. Moreover, our exploration of retrieval sources underscores the need for refining these modules to enhance efficacy. The demonstration of the method's effectiveness in open-domain question answering also sets the stage for future research. The method's versatility suggests potential applications in various generative tasks, motivating our future exploration in tasks like summarization.

Limitations

Our study acknowledges several limitations that warrant consideration. First, we acknowledge the restriction imposed by employing a singular prompt template. The computational complexity inherent in our method limited the scope of experiments conducted within this framework. However, this constraint was pivotal in maintaining consistency across our comparisons, ensuring the reliability and robustness of the obtained results despite the limitation in the number of explored templates.

Secondly, while our decoding method was specifically showcased in the context of Question Answering (QA) using greedy decoding, another limitation of this study is that we haven't explored its application to other generative tasks. It's essential to note that our approach is designed as a general decoding framework applicable to various generative tasks. Thus, expanding this work to other domains such as summarization and mitigating language hallucination remains a promising avenue for future exploration.

Furthermore, it's imperative to recognize that the scalability and generalizability of our method across different problem domains and decoding strategies might present further challenges and considerations. Extending our investigation to encompass a broader array of prompt templates and decoding strategies (such as nucleus sampling) could potentially reveal nuanced insights into the adaptability and effectiveness of our proposed method.

Additionally, it is crucial to note that the decoding time required for our method is longer than regular decoding, approximately three times longer, owing to decoding using three logits distributions simultaneously. However, there exists potential for mitigating the time complexity by distributing the decoding of different distributions across multiple GPU machines, thereby enabling parallelization and potentially reducing the computational overhead. This approach might alleviate the time constraints associated with our decoding method, rendering it more feasible for applications requiring low decoding latency.

Acknowledgements

We would like to thank Christos Christodoulopoulos, Marco Damonte, and Clara Vania for their insightful discussions that contributed to this work. We also appreciate the assistance of Yash Malik, Miguel Angel Rubio, and Mohammad Shahdad in setting up the computing environment for our experiments. Finally, we thank the anonymous reviewers for their helpful feedback, which helped us improve the clarity and quality of the paper.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset.

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.

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.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

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

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.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. 2022. TemporalWiki: A lifelong benchmark for training and evaluating ever-evolving language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6237–6250, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

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.

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.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2022. Realtime qa: What's the answer right now?

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.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

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.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. 2022. Time waits for no one! analysis and challenges of temporal misalignment. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5944–5958, Seattle, United States. Association for Computational Linguistics.

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. 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.

Sewon Min, Weijia Shi, Mike Lewis, Xilun Chen, Wentau Yih, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2023. Nonparametric masked language modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2097–2118, Toronto, Canada. Association for Computational Linguistics.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics.

Charles O'Neill, Yuan-Sen Ting, Ioana Ciuca, Jack Miller, and Thang Bui. 2023. Steering language generation: Harnessing contrastive expert guidance and negative prompting for coherent and diverse synthetic data generation.

Luiza Pozzobon, Beyza Ermis, Patrick Lewis, and Sara Hooker. 2023. Goodtriever: Adaptive toxicity mitigation with retrieval-augmented models.

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

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

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

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

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

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. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pretrained transformer language models.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Merging generated and retrieved knowledge for open-domain QA. In *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing, pages 4710–4728, Singapore. Association for Computational Linguistics.

Zexuan Zhong, Tao Lei, and Danqi Chen. 2022. Training language models with memory augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5657–5673, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

A Additional Results on Scaling Experiments

We present additional scaling experiment results for different model variants. Specifically, we illustrate the outcomes for Flan-T5 variants, 3B and 11B, in Figure 6. The results for Falcon variants, particularly Falcon 7B and 40B, are depicted in Figure 7. Moreover, we showcase the results for OPT variants, encompassing OPT 6.7B, 13B, 30B, and 66B, in Figure 8. Additionally, the findings pertaining to Llama 2 variants, including Llama 2 7B, 13B, and 70B, are illustrated in Figure 9. We can see that our proposed decoding method outperforms regular decoding with open-book setting in most settings across different datasets and model sizes.

B Additional Details on Irrelevant Context

Here we provide the meticulously designed adversarial $\boldsymbol{c}^-$ irrelevant context that is used as the fixed $\boldsymbol{c}^-$ for every query:

"It was a pleasant weather day, with seasonally average temperatures. The local legislative and academic governing bodies held routine meetings regarding budgets and policies. Students focused on their studies while athletes practiced for upcoming competitions. Residents tended to their jobs and daily tasks around their neighborhood. Nothing particularly eventful occurred in the community. It was an ordinary midweek day. The weather was typical for the time of year without any extreme events. Overall it was an average day in the community with people pursuing their regular daily activities."

Here is the same fixed $\boldsymbol{c}^-$ but with word order permuted:

"an routine Overall was of community. average focused for The around tended upcoming their was policies. their budgets and Residents to eventful held competitions. It particularly extreme with academic temperatures. was day. weather local The their studies events. it meetings average pleasant typical Nothing ordinary time seasonally legislative people an the daily the Students in a neighborhood. activities. community pursuing weather and while in midweek regarding athletes occurred tasks the daily jobs It governing year bodies regular with their for day and practiced on day, was without any"

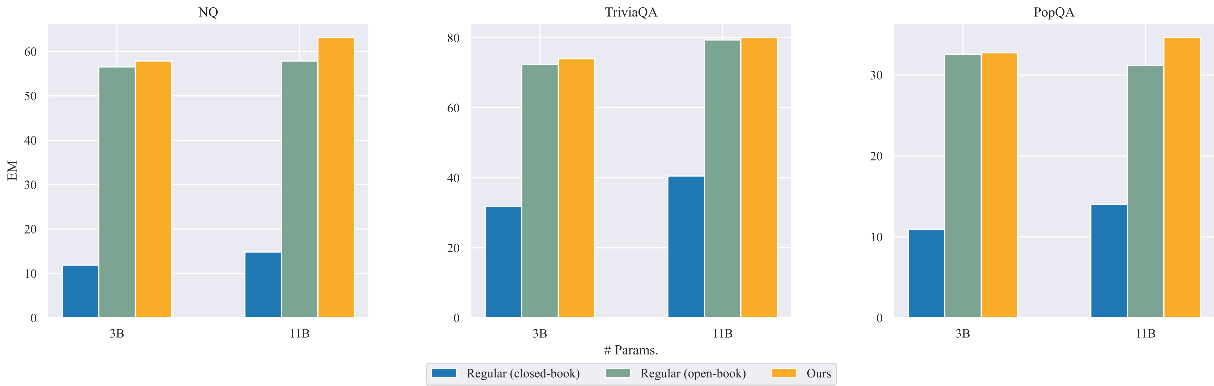

Figure 6: Performance comparison of our method against regular decoding across various sizes of Flan-T5 models.

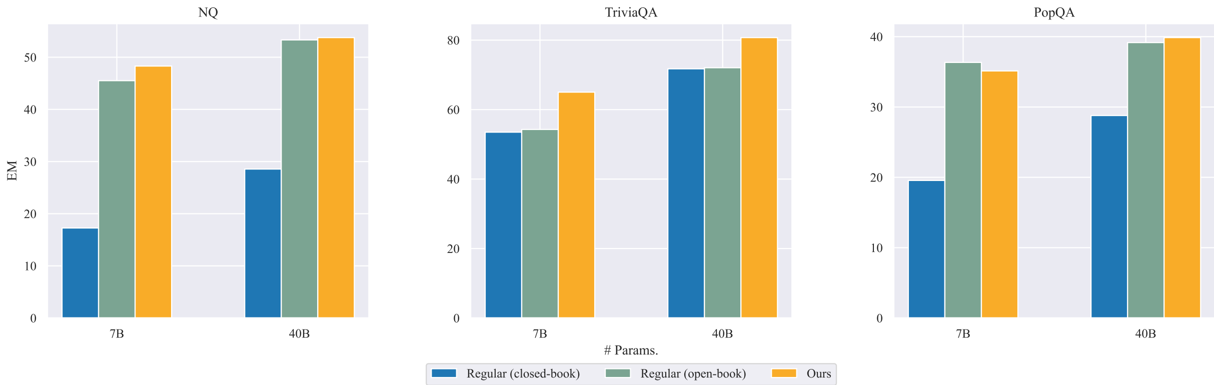

Figure 7: Performance comparison of our method against regular decoding across various sizes of Falcon models.

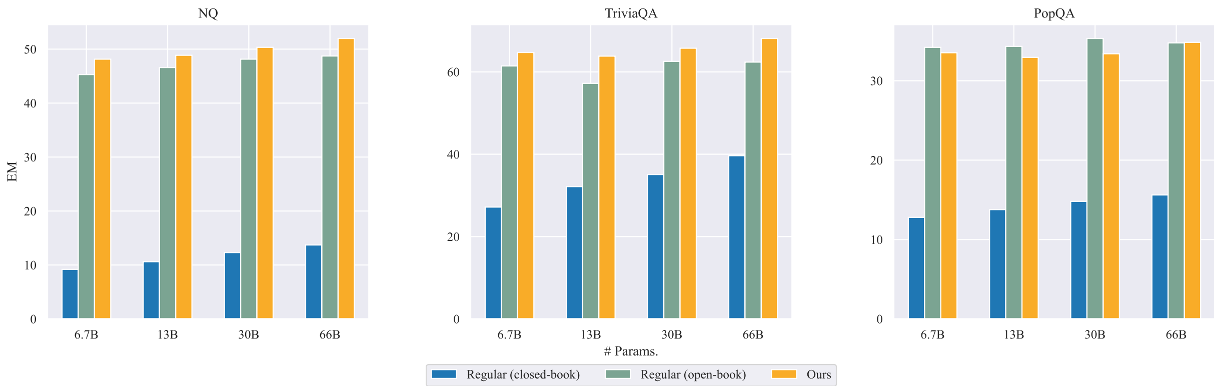

Figure 8: Performance comparison of our method against regular decoding across various sizes of OPT models.

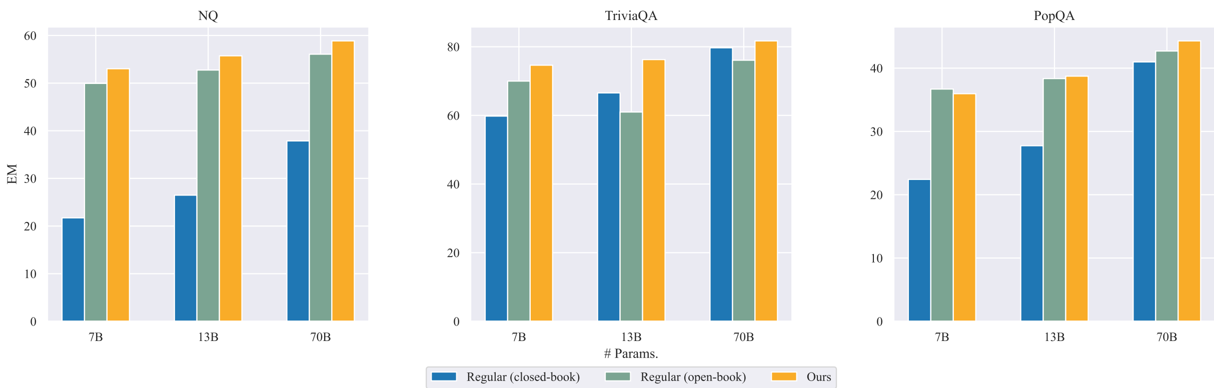

Figure 9: Performance comparison of our method against regular decoding across various sizes of Llama 2 models.