Entropy-Based Decoding For Retrieval-Augmented Large Language Models

Zexuan Qiu[1]* **Zijing Ou**[2] **Bin Wu**[3] **Jingjing Li**[1] **Aiwei Liu**[4] **Irwin King**[1]

[1]The Chinese University of Hong Kong [2]Imperial College London
[3]University College London [4]Tsinghua University

Abstract

Augmenting Large Language Models (LLMs) with retrieved external knowledge has proven effective for improving the factual accuracy of generated responses. Despite their success, retrieval-augmented LLMs still face the distractibility issue, where the generated responses are negatively influenced by noise from both external and internal knowledge sources. In this paper, we introduce a novel, training-free decoding method guided by entropy considerations to mitigate this issue. Our approach utilizes entropy-based document-parallel ensemble decoding to prioritize low-entropy distributions from retrieved documents, thereby enhancing the extraction of relevant information of context. Additionally, it incorporates a contrastive decoding mechanism that contrasts the obtained low-entropy ensemble distribution with the high-entropy distribution derived from the model's internal knowledge across layers, which ensures a greater emphasis on reliable external information. Extensive experiments on open-domain question answering datasets demonstrate the superiority of our method.

1 Introduction

In recent years, Large language models (LLMs) have revolutionized natural language processing, showcasing remarkable performance across various downstream tasks [3, 26, 33]. However, they still struggle with hallucination due to the inaccuracy of parametric memory [4] and inherently tend to produce outdated information [13]. In contrast, explicitly augmenting LLMs with retrieved external knowledge from reliable datastores [17, 2] can enable LLMs to generate content that exhibits less deviation from the truth, and benefit downstream knowledge-intensive tasks [27].

Despite the success of retrieval-augmented LLMs, the augmented generation is still sub-optimal due to the *distractibility issue*, where the generated responses are easily negatively affected by noise from both external knowledge and intrinsic model knowledge. As for the input context, LLMs' understanding of context can be *explicitly distracted by irrelevant parts* within the retrieved context [30]. A typical illustrative case is the "lost in the middle" distraction phenomenon observed in the synthetic multi-document question-answering scenario [22, 28], where the oracle document containing the correct answer is encircled by numerous retrieved distracting documents. In this scenario, LLMs frequently fail to deliver the correct answer unless the oracle document is strategically placed at the very beginning or end of the context. With regard to intrinsic knowledge, LLMs are easily *implicitly distracted by the parametric knowledge* acquired during pre-training. This is in conflict with retrieval-augmented generation which is expected to generate responses based on reliable retrieved context. Particularly in the domain of question answering, previous works [23, 37] show that LLMs stubbornly adhere to their built-in knowledge even when it conflicts with external knowledge.

How to eliminate the impact of the above-mentioned distractibility issue, so as to extract useful knowledge from the retrieved context for the input query, is our research focus. Although existing

*Correspondence to: Zexuan Qiu (zxqiu22@cse.cuhk.edu.hk).

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

works strive to effectively leverage the retrieved context by directly fine-tuning retrieval-augmented LLMs [20, 39] or incorporating trainable encoder modules [9, 38], these approaches require additional training, rendering them potentially impractical in resource-constrained environments. In this paper, we propose a novel decoding method guided by entropy considerations to simultaneously mitigate the impact of noisy information from both the external context and parametric knowledge. The proposed method can be seamlessly integrated into LLMs without requiring additional tuning.

Specifically, to enhance LLMs' ability to extract useful information from multiple retrieved documents, we let LLMs process each retrieved document in parallel and ensemble of the output distributions from each document to determine the next-token distribution, with the ensemble weights adaptively assigned based on the uncertainty of each document-conditioned distribution. At each generation step, documents with lower uncertainty (*i.e.*, lower entropy) in the LLM output are given more attention during decoding. Ultimately, we obtain a low-entropy distribution aggregated among documents. Furthermore, to alleviate the potential distraction from parametric knowledge, we refine the next-token distribution by contrasting the obtained low-entropy distribution when feeding the retrieved documents, against the distribution without context. Here, we propose to use the distribution from the layer exhibiting the highest entropy without context for contrast, in order to highlight the proportional changes in token probabilities after introducing external knowledge.

The proposed decoding method shows an impressive performance in the synthetic challenging multi-document scenario [22] where the negative impact of retrieved distractor documents is emphasized. We further conduct extensive experiments across four LLMs of varying sizes on four diverse open-domain question answering tasks including NQ [14], TriviaQA [11], WebQ [1] and PopQA [24]. Experimental results confirm the superiority of our methods and validate the effectiveness of each component.

2 Methodology

Given an input query x, RAG first retrieves top-K relevant documents $D := \{d_1, d_2, \ldots, d_K\}$ from the knowledge base via a retriever as external evidence, which is then incorporated with the query as the input to large language model parametrized by θ for generating a faithful response y. A common approach, termed as "NAIVE RAG", involves concatenating the query x, the previously generated response $y_{<t}$, and the retrieved documents D as the input sequence, resulting in the following decoding method $p_\theta(y_t|D, x, y_{<t}) = p(y_t|d_0 \circ \cdots \circ d_k \circ x \circ y_{<t})$, where $\circ$ denotes the concatenation operation. Although this method performs well, Liu et al. [22] point out its vulnerability to the "lost in the middle" phenomenon, where LLMs may miss the key document among others unless it's positioned at the start or end of the input. Moreover, this simple approach does not consider the potential negative effects of the underlying parametric knowledge of LLMs. Moreover, this simple approach does not consider the potential negative effects of the underlying parametric knowledge of LLMs. In the subsequent sections, we discuss mitigating these two issues by entropy considerations.

2.1 Entropy-Based Document Ensemble

Instead of naively concatenating the documents D, we propose to alleviate the "loss in the middle" issue using the product-of-experts[2] ensemble approach [8]. Specifically, we model the log probability of the next-token distribution as $\log p_\theta\left(y_t \mid D\right) \propto \sum_{j=1}^{K} w_{j,t} \log p_\theta\left(y_t \mid d_j \circ x \circ y_{<t}\right)$ where $\sum_j w_{j,t} = 1, \forall t$, $p_\theta(y_t|D, x, y_{<t})$ is denoted as $p_\theta(y_t|D)$ for short and $w_{j,t}$ denotes the weight of the j-th document on generating the token at the t-th time step. Each document in D is concatenated with the query and the previously generated response. This combined input is then individually fed into the LLM. The output logit scores are subsequently averaged using the weights $w_{j,t}$. This ensemble approach, which leverages parallel decoding, helps mitigate position bias and provides a more effective means of utilizing the retrieved documents.

There are multiple choices to compute the weights. A straightforward option is to use uniform weighting, *i.e.*, $w_{j,t} = \frac{1}{|D|}$. However, this method may fail to effectively extract valuable information when irrelevant documents are included among the top-K retrieved documents. Another option is to utilize a similarity score $s(d_j, x)$ between the query and the retrieved document, *e.g.*, BM25, as the time-independent ensemble weights $w_{j,t} \propto s(d_j, x)$ [17]. However, this retriever-based scoring approach can impair the LLM's ability to extract relevant information amid many distractors. We suggest that the uncertainty in the next-token distribution naturally indicates the informativeness of the

[2] Empirically, we find that product-of-experts and mixture-of-experts methods yield similar performance.

retrieved documents. Similar concepts have been employed in previous work to reduce hallucinations in LLMs [34, 35]. Consequently, we propose using an entropy-based score $w^{\mathrm{H}}_{j,t}$ as the preference weight for each document at each decoding step:

$$w^{\mathrm{H}}_{j,t} = \frac{\exp^{(-H_{j,t})/\tau}}{\sum_{d_k \in D} \exp^{(-H_{j,t})/\tau}}, \quad H_{j,t} = -\sum_{y_t \in \mathcal{V}} p_\theta(y_t|d_j) \log p_\theta(y_t|d_j), \tag{1}$$

where $p_\theta(y_t|d_j)$ denotes $p_\theta(y_t|d_j \circ x \circ y_{<t})$ for short, $\mathcal{V}$ represents the vocabulary set and τ is a hyper-parameter controlling the concentration level of distributions. The motivation behind Eq. (1) is that the LLM can autonomously evaluate the significance of each document during the generation process. Intuitively, it implies that those document-conditioned distributions with lower uncertainty will be assigned higher weights. Such a time-dependent approach can effectively capture useful information from the retrieved documents at each generation step, thereby influencing the generation process more significantly. We refer to this method as **LeEns** (**L**ow-**e**ntropy **Ens**emble).

2.2 Entropy-Based Contrastive Decoding

While **LeEns** can effectively help LLMs discern valuable evidence from external knowledge, the parametric knowledge of LLMs embedded during the pre-training phase might affect the answer generation, especially when these two types of knowledge conflict [23, 37]. In this section, we propose to address this issue via entropy-based contrastive decoding.

Contrastive Decoding with PMI. Inspired by the success in contrastive decoding to mitigate hallucination of LLMs [31, 19], we adjust the logit score z_t for the generated token $y_t \in \mathcal{V}$ at the t-th time step by incorporating the pointwise mutual information (PMI) between y_t and the document set D, given the query x:

$$z_t = \log p^{\mathrm{H}}_\theta(y_t|D) + \beta \overbrace{\log \frac{p^{\mathrm{H}}_\theta(y_t|D)}{p_\theta(y_t|x, y_{<t})}}^{\text{PMI}} = (1+\beta) \log p^{\mathrm{H}}_\theta(y_t|D) - \beta \log p_\theta(y_t|x, y_{<t}), \tag{2}$$

where β is a positive coefficient controlling the contrast intensity, and $p^{\mathrm{H}}_\theta(y_t|D)$ denotes the previously proposed entropy-based document ensemble distribution. Intuitively, PMI serves as a measurement of information gains. The model tends to generate tokens with a high probability of $p^{\mathrm{H}}_\theta(y_t|D)$ and a low probability of $p_\theta(y_t|x, y_{<t})$. These tokens provide greater information gain for the next token generation. Therefore, incorporating PMI can enhance the model's reliance on external knowledge.

Layer-wise Contrast with High Entropy. To perform contrastive decoding, it is necessary to compute $p_\theta(y_t|x, y_{<t})$. This can be achieved by taking the hidden states from the last layer of LLMs and passing them through the classification head. However, the distribution derived from the last layer may exhibit overconfidence, characterized by extremely low probabilities for most words and disproportionately high probabilities for a few. Such overconfidence can erroneously amplify external knowledge when conducting contrasting, potentially leading to false positive failures, as illustrated in Figure 4 in the Appendix. To address this issue, we propose selecting the layer that contains the most "ambiguous" parametric knowledge among the layers as a proper reference for contrast. This allows the model to more effectively leverage external knowledge, reducing overconfidence and improving the accuracy of the generated outputs. Specifically, for a LLM consisting of a total of L layers, we denote the probability for $y_t \in \mathcal{V}$ in the l-th layer as $p^l_\theta(y_t|x, y_{<t}) = \mathrm{softmax}(\boldsymbol{W}_{\mathrm{LM}} h^l_{t-1})$, where h^l_{t-1} denotes the hidden state for layer l out of L, and $\boldsymbol{W}_{\mathrm{LM}}$ denotes the linear classification head. At each decoding step, we select the layer with the maximum uncertainty for contrast:

$$l^* = \arg\max_{l \in \mathcal{L}} H^l_t, \quad H^l_t = -\sum_{y_t \in \mathcal{V}} p^l_\theta(y_t|x, y_{<t}) \log p^l_\theta(y_t|x, y_{<t}), \tag{3}$$

where $\mathcal{L}$ is the set of candidate layers, which are set as the last few layers of LLMs practically to ensure that each of them contains certain plausible information. The dynamic strategy in Eq. 3 resembles that of DoLa [5], with the key distinction being the adoption of uncertainty rather than distribution differences as the selection criterion. Combining Eq. (1) and Eq. (3), the adjusted next-token distribution is formulated as:

$$y_t \sim \mathrm{softmax}\Big[(1+\beta) \sum_{d_j \in D} w^{\mathrm{H}}_{j,t} \log p_\theta(y_t|d_j, x, y_{<t}) - \beta \log p^{l^*}_\theta(y_t|x, y_{<t})\Big]. \tag{4}$$

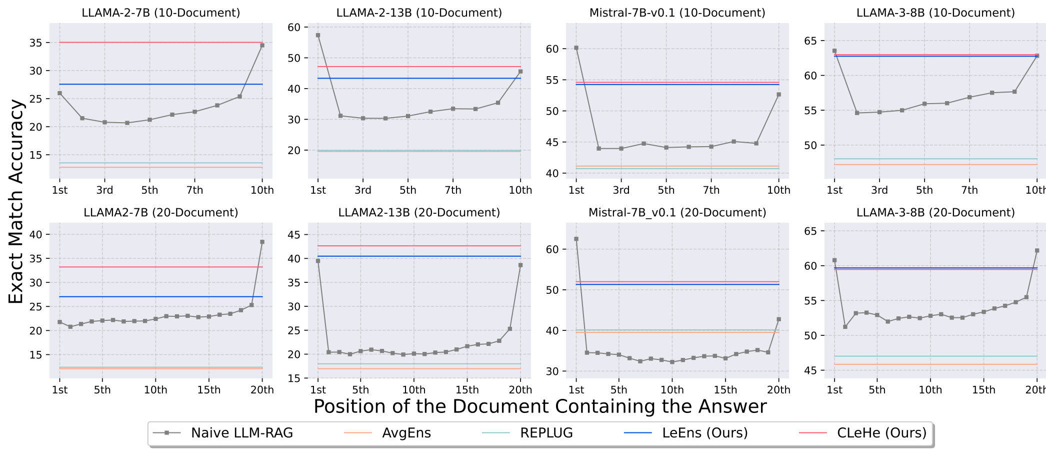
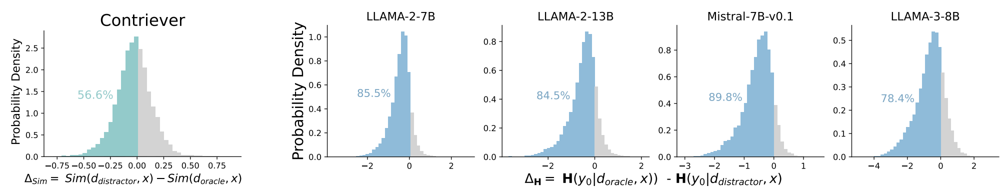


Figure 1: Impact of positioning the oracle document on multi-document question answering performance. A 10-document context typically uses less than 2K tokens; a 20-document context usually uses less than 4K tokens.



(a) (b)

Figure 2: (a) The distribution of the similarity difference between the query and the oracle document versus the query and distractor documents. (b) The distribution of the difference in entropy of the first token generated by LLMs when given the oracle document versus distractor documents.

Here, β represents the amplification intensity of external knowledge. The ultimate distribution in Eq. (4) can be interpreted as a two-stage ensemble process. Firstly, it ensembles the retrieved documents with uncertainty to generate a low-entropy distribution that more effectively captures the external knowledge within these documents. Secondly, it performs a contrastive ensemble by differentiating the logits of this low-entropy distribution from the high-entropy distribution of parametric knowledge selected across different layers, thereby prioritizing factual information from external sources.Therefore, we term the method in Eq. (4) as **CLeHe** (**C**ontrasting **L**ow-**e**ntropy distribution with **H**igh-**e**ntropy distribution).

3 Analyzing the Distraction Phenomenon in Retrieved Context

Experiment Settings. We are particularly interested in a challenging QA scenario proposed by [22], in which the oracle document is surrounded by numerous semantically similar distractor documents. Specifically, following [22], given a query from NaturalQuestions-Open [14], we select a Wikipedia paragraph containing the answer from the NaturalQuestions annotations as the oracle document. Then, Contriever [6] is employed to extract K-1 additional paragraphs from the Wikipedia corpus that are highly relevant to the query yet do not include the ground truth answer, functioning as distractor documents. The query, the oracle document, and K-1 distractor documents are subsequently processed by the LLM to generate an answer. Refer to Appendix B for more implementation details.

We compare three training-free baselines. NAIVE: this method concatenates all retrieved documents directly along with the question to form the prompt for LLMs; REPLUG [32]: it utilizes a normalized retriever weight to ensemble during the decoding process; AvgEns : it follows the formulation in similarly Eq.(1) but assigns the same weight to each document during each generation step.

Results. Figure 1 shows that the NAIVE method's performance, which concatenates all documents for context, depends heavily on the placement of the oracle document. Among the four evaluated LLMs, the performance of the NAIVE method significantly deteriorates when the oracle document is neither at the very beginning nor at the end. In contrast, since the proposed LeEns processes each document in parallel during decoding, its performance is naturally independent of the oracle document's position. In almost all positions, LeEns substantially surpasses the performance of the NAIVE method. Notably, in this challenging scenario, REPLUG which ensemble documents' distributions based on retriever weights perform exceedingly poorly, achieving results merely on par with AvgEns . Based on this observation, we further conduct the weight analysis in Figure 2. As depicted in Figure 2a, in only approximately 57% of instances, Contriever identifies the oracle document as more similar to the query than the distractor documents. Figure 2b shows the distribution of entropy differences for the first token when conditioned on oracle documents versus distractor documents. It indicates that the entropy is generally lower when the response to a query is based on the oracle document rather than the distractor documents.

4 Related Works

Retrieval-Augmented Language Models. Enhancing large language models (LLMs) with information retrieved from external knowledge bases has proven effective for various knowledge-intensive tasks. Initially, mainstream research in retrieval-augmented language models (RALM) focused on leveraging retrieved knowledge during the pre-training phase of LLMs [7, 10, 2]. To mitigate the computational costs, some studies have concentrated on lightweight fine-tuning methods to integrate retrieval capabilities into LLMs [17, 20, 39]. Notably, models like FiD [9] and CEPE [38] perform parallel encoding of multiple retrieved documents using a fine-tuned encoder, enabling decoder-only LLMs to more effectively capture and utilize external knowledge. Another approach leverages the in-context learning abilities of LLMs to incorporate external knowledge in a training-free manner [29, 32]. The work most closely related to ours is REPLUG[32], which utilizes the RAG-token model [17] to perform parallel retrieval augmentation based on retrieval scores. However, we empirically demonstrate that focusing on the inherent uncertainty within the LLM's output distribution, rather than relying solely on pre-existing retrieval scores, can significantly improve the factual accuracy of content generated from retrieved documents.

Contrastive Decoding. The idea of contrastive decoding (CD) has been previously applied in controllable text generation to produce non-toxic by DExperts [21]. Later, Li et al. [19] formalized CD as a method to enhance open-ended text generation without any additional training by maximizing the difference in log probabilities between an expert LLM and an amateur LLM. This approach has demonstrated strong performance in various domains, including reasoning [25] and neural machine translation [36]. CD can also be interpreted as maximizing pointwise mutual information (PMI), which has proven effective in other scenarios. For instance, Li et al. [18] uses a training objective that maximizes PMI to generate more diverse conversational responses, while CAD [31] employs a PMI-adjusting distribution to resolve the knowledge conflict. Chuang et al. [5] proposes a decoding strategy that contrasts different layers of the same LLM to more effectively highlight factual knowledge. Similar principles are also applied in visual LLMs, where Leng et al. [16] mitigates object hallucination by contrasting distributions derived from original and distorted visual inputs. Alternatively, our proposed CLeHe leverages layer-wise entropy-based contrastive decoding to prioritize external knowledge over the parametric knowledge inherent in the LLM itself.

5 Conclusions

In this paper, we proposed a novel decoding method that is guided by entropy considerations to mitigate the distractibility issue from both external retrieved documents and parametric knowledge. First, we conducted parallel retrieval augmentation with entropy-based ensemble weight to obtain the low-entropy distribution of context. Furthermore, we contrasted this distribution against the highest-entropy distribution among layers when without context to amplify the external knowledge preserved in context. Extensive experiments showed the proposed method's effectiveness in retrieval-augmented open-domain question answering.

References

[1] J. Berant, A. Chou, R. Frostig, and P. Liang. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544, 2013.

[2] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR, 2022.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[5] Y.-S. Chuang, Y. Xie, H. Luo, Y. Kim, J. R. Glass, and P. He. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*, 2023.

[6] I. Gautier, C. Mathilde, H. Lucas, R. Sebastian, B. Piotr, J. Armand, and G. Edouard. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022.

[7] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.

[8] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[9] G. Izacard and E. Grave. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*, 2020.

[10] G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, and E. Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43, 2023.

[11] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL (1)*, pages 1601–1611. Association for Computational Linguistics, 2017.

[12] V. Karpukhin, B. Oguz, S. Min, P. S. H. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781. Association for Computational Linguistics, 2020.

[13] J. Kasai, K. Sakaguchi, R. Le Bras, A. Asai, X. Yu, D. Radev, N. A. Smith, Y. Choi, K. Inui, et al. Realtime qa: What's the answer right now? *Advances in Neural Information Processing Systems*, 36, 2024.

[14] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

[15] K. Lee, M. Chang, and K. Toutanova. Latent retrieval for weakly supervised open domain question answering. In *ACL (1)*, pages 6086–6096. Association for Computational Linguistics, 2019.

[16] S. Leng, H. Zhang, G. Chen, X. Li, S. Lu, C. Miao, and L. Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*, 2023.

[17] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[18] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. In *HLT-NAACL*, pages 110–119. The Association for Computational Linguistics, 2016.

[19] X. L. Li, A. Holtzman, D. Fried, P. Liang, J. Eisner, T. Hashimoto, L. Zettlemoyer, and M. Lewis. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*, 2022.

[20] X. V. Lin, X. Chen, M. Chen, W. Shi, M. Lomeli, R. James, P. Rodriguez, J. Kahn, G. Szilvasy, M. Lewis, et al. Ra-dit: Retrieval-augmented dual instruction tuning. *arXiv preprint arXiv:2310.01352*, 2023.

[21] A. Liu, M. Sap, X. Lu, S. Swayamdipta, C. Bhagavatula, N. A. Smith, and Y. Choi. Dexperts: Decoding-time controlled text generation with experts and anti-experts. In *ACL/IJCNLP (1)*, pages 6691–6706. Association for Computational Linguistics, 2021.

[22] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

[23] S. Longpre, K. Perisetla, A. Chen, N. Ramesh, C. DuBois, and S. Singh. Entity-based knowledge conflicts in question answering. *arXiv preprint arXiv:2109.05052*, 2021.

[24] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, and H. Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *ACL (1)*, pages 9802–9822. Association for Computational Linguistics, 2023.

[25] S. O'Brien and M. Lewis. Contrastive decoding improves reasoning in large language models. *arXiv preprint arXiv:2309.09117*, 2023.

[26] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[27] F. Petroni, A. Piktus, A. Fan, P. Lewis, M. Yazdani, N. De Cao, J. Thorne, Y. Jernite, V. Karpukhin, J. Maillard, et al. Kilt: a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252*, 2020.

[28] Z. Qiu, J. Li, S. Huang, W. Zhong, and I. King. Clongeval: A chinese benchmark for evaluating long-context large language models. *arXiv preprint arXiv:2403.03514*, 2024.

[29] O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, and Y. Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023.

[30] F. Shi, X. Chen, K. Misra, N. Scales, D. Dohan, E. H. Chi, N. Schärli, and D. Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR, 2023.

[31] W. Shi, X. Han, M. Lewis, Y. Tsvetkov, L. Zettlemoyer, and S. W.-t. Yih. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*, 2023.

[32] W. Shi, S. Min, M. Yasunaga, M. Seo, R. James, M. Lewis, L. Zettlemoyer, and W.-t. Yih. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*, 2023.

[33] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[34] L. Van der Poel, R. Cotterell, and C. Meister. Mutual information alleviates hallucinations in abstractive summarization. *arXiv preprint arXiv:2210.13210*, 2022.

[35] N. Varshney, W. Yao, H. Zhang, J. Chen, and D. Yu. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*, 2023.

[36] J. Waldendorf, B. Haddow, and A. Birch. Contrastive decoding reduces hallucinations in large multilingual machine translation models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2539, 2024.

[37] J. Xie, K. Zhang, J. Chen, R. Lou, and Y. Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*, 2023.

[38] H. Yen, T. Gao, and D. Chen. Long-context language modeling with parallel context encoding. *arXiv preprint arXiv:2402.16617*, 2024.

[39] T. Zhang, S. G. Patil, N. Jain, S. Shen, M. Zaharia, I. Stoica, and J. E. Gonzalez. Raft: Adapting language model to domain specific rag. *arXiv preprint arXiv:2403.10131*, 2024.

A Methodology Overview

Figure 3 illustrates the overview of the decoding process of CLeHe. Figure 4 illustrates a false positive scenario. In this example, the retrieval-augmented model assigns high probabilities to the words "Washington", "New York", and "Columbia" as candidate positives. However, in the low-entropy output of a specific layer (typically the last layer) without context, the probability assigned to "Columbia" is notably low. If contrastive decoding is applied, it would mistakenly increase "Columbia's" probability, leading to an incorrect prediction.

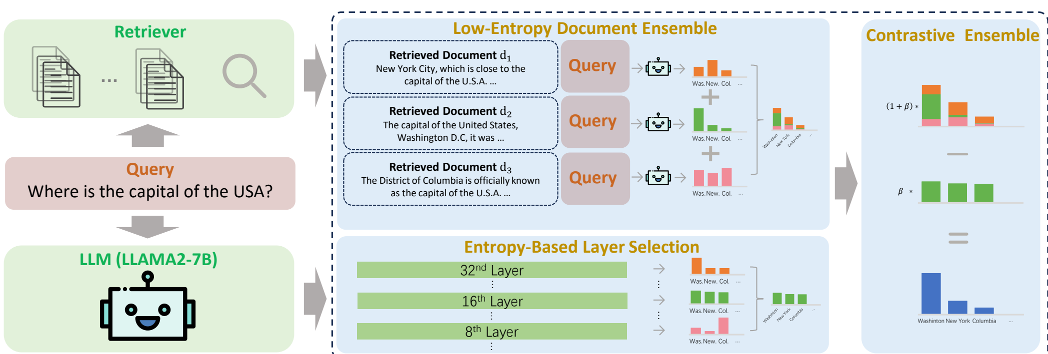


Figure 3: Overview of the decoding process of CLeHe.

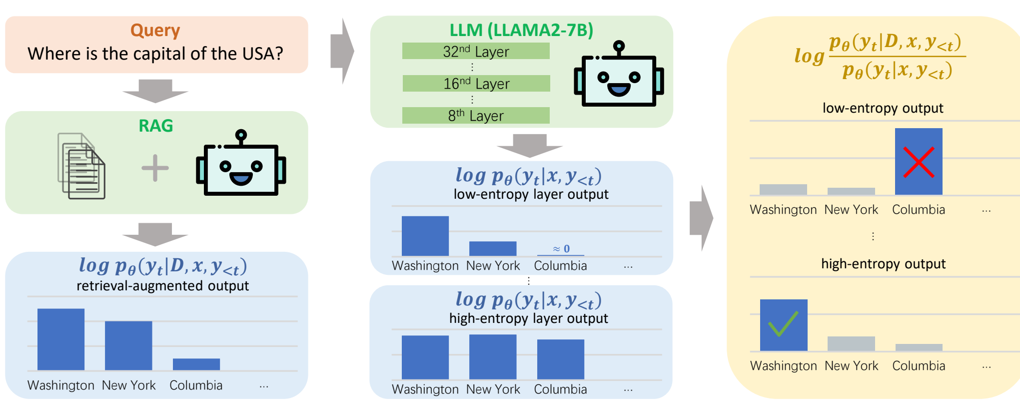


Figure 4: Illustration of the false positive case.

Dataset	NQ			TQA			WebQ			PopQA			Gain
# of Docs	5	10	20	5	10	20	5	10	20	5	10	20	
LLAMA-2-7B													
NAIVE	19.56	26.76	26.37	52.53	59.05	66.00	16.04	18.06	18.35	23.64	25.50	26.10	0.00
AvgEns	14.93	13.51	12.78	51.04	49.84	48.87	12.50	12.06	11.35	14.03	12.49	11.74	-9.40
REPLUG	23.85	23.87	23.92	55.56	55.37	55.31	16.63	16.39	16.77	23.85	23.80	23.91	-1.56
LeEns	25.48	25.79	25.90	61.87	62.72	63.16	20.28	21.65	21.66	27.84	26.57	25.07	+2.50
CLeHe	**37.62**	**36.29**	**35.48**	**69.56**	**69.92**	**69.72**	**36.12**	**36.22**	**35.77**	**32.12**	**31.11**	**28.89**	**+11.74**
LLAMA-2-13B													
NAIVE	**37.98**	**39.67**	29.07	66.82	68.15	69.66	29.83	37.02	27.58	**34.77**	**35.14**	31.62	0.00
AvgEns	23.52	20.19	18.71	66.92	65.48	63.86	25.89	24.16	22.64	24.28	22.58	22.02	-8.92
REPLUG	34.12	33.82	34.01	67.83	67.77	67.60	31.25	31.30	30.88	30.79	30.95	31.00	-1.33
LeEns	36.54	34.65	33.63	71.87	72.07	72.16	36.07	35.63	35.42	34.63	33.31	31.72	+1.70
CLeHe	37.31	36.01	**34.95**	**72.24**	**72.82**	**72.55**	**38.19**	**37.45**	**36.66**	34.22	33.16	**31.79**	**+3.18**
Mistral-7B-v0.1													
NAIVE	46.20	44.43	42.20	73.53	70.31	73.89	47.69	45.03	40.61	40.23	37.72	38.94	0.00
AvgEns	40.91	39.36	38.22	76.28	75.73	74.91	47.98	48.13	47.83	37.11	34.61	33.44	-0.52
REPLUG	44.35	44.44	44.58	74.62	74.80	74.60	47.59	47.49	47.21	37.76	37.75	37.78	+1.02
LeEns	**46.40**	**46.45**	**44.65**	**78.26**	**78.97**	**79.25**	**49.21**	49.70	50.32	**42.12**	43.31	43.87	**+4.32**
CLeHe	46.32	46.07	44.64	78.19	78.88	79.14	49.06	**49.76**	**50.37**	42.11	**43.34**	**43.89**	+4.25
LLAMA-3-8B													
NAIVE	**52.35**	**51.69**	**52.33**	79.92	81.11	82.10	50.49	50.15	50.12	40.92	42.24	42.95	0.00
AvgEns	47.12	45.70	44.51	81.31	80.73	79.78	51.82	51.24	51.01	38.95	36.49	35.07	-2.72
REPLUG	50.39	50.33	50.50	79.07	79.16	78.76	50.20	50.79	50.27	39.08	39.24	39.04	-1.63
LeEns	51.74	50.53	49.47	**81.80**	82.68	83.02	**52.17**	50.98	51.80	43.63	**44.92**	**45.62**	+1.00
CLeHe	52.02	50.78	49.67	81.78	**82.84**	**83.14**	51.67	**51.62**	**52.39**	**43.86**	44.84	45.57	**+1.17**

Table 1: Performance (%) comparison of different ensemble-based methods on benchmark datasets. "Gain" refers to the average absolute improvement (%) across all datasets and different numbers of retrieved documents when compared to the naive baseline.

B More Experiment on Open-Domain QA

Implementation Details. Our method introduces two hyperparameters: τ to control the relative importance of different documents during decoding; and β balancing contextual and parametric knowledge. We extract a subset from the WebQ training set for validation to determine the hyperparameter value for each LLM. Ultimately, τ is set as 0.25 for LLAMA-3-8B and as 0.1 for the other three models. For β, 5.0 is chosen for LLAMA-2-7B, while the other models are assigned a value of 0.25. During the inference, greedy decoding is utilized for reproducibility. When looking for the layer with the highest entropy, we focus our search exclusively on the candidate layers. In our preliminary experiments, this approach improves computational efficiency and slightly enhances model performance. For LLAMA-2-7B, Mistral-7B-v0.1, and LLAMA-3-8B with 32 hidden layers, the candidate layers are set to $\{17, \ldots, 32\}$, and only even-numbered layers will be searched. For LLAMA-2-13B with 40 hidden layers, the candidate layers are set to $\{31, \ldots, 40\}$. All experiments are conducted on a single A100 80GB GPU.

Datasets and Metrics. We evaluate our proposed method using four open-domain QA datasets. Natural Questions [14], TriviaQA [11], WebQ [1] and PopQA [24]. Natural Questions includes real anonymized queries from Google's search engine. We utilize a filtered test set [15] of 3,610 samples with answers limited to no more than five tokens. TriviaQA comprises trivia question-answer pairs that were scraped from the web. We evaluate its development set containing 7,993 samples. WebQ consists of questions generated through the Google Suggest API, with answers that are entities in Freebase. We use its test set of 2,032 samples for evaluation. PopQA is a novel entity-centric open-domain QA dataset that spans a wide range of entity popularity, emphasizing long-tail knowledge. We utilize its test set which includes 14,267 samples for evaluation. For each dataset, we retain only the questions and their corresponding answers. DPR [12] is employed to retrieve the top-k passages

	LLAMA-2-7B			LLAMA-2-13B			Mistral-7B-v0.1			LLAMA-3-8B		
	NQ	TQA	WebQ	NQ	TQA	WebQ	NQ	TQA	WebQ	NQ	TQA	WebQ
NAIVE	19.56	52.53	16.04	37.98	66.82	29.83	46.20	73.53	**47.69**	**52.35**	79.92	**50.49**
w/ JSD (DoLa)	38.72	68.15	35.38	27.34	45.15	21.90	46.20	74.42	46.80	52.30	81.36	49.54
w/ Last_Layer (CAD)	38.92	65.92	30.77	**41.52**	**68.96**	**33.76**	44.32	70.27	41.49	51.80	79.01	46.26
w/ Entropy	**41.36**	**70.32**	**37.30**	39.91	67.38	32.73	**46.30**	**74.59**	47.05	52.28	**81.60**	49.61
LeEns	25.48	61.87	20.28	36.54	71.87	36.07	**46.40**	78.26	49.21	51.74	**81.80**	**52.17**
w/ JSD (DoLa)	35.84	67.11	33.75	17.22	35.17	11.02	46.29	78.19	49.16	52.04	81.73	51.82
w/ Last_Layer	30.74	62.85	23.08	**38.19**	71.64	38.09	45.57	75.54	48.12	**52.60**	81.36	52.01
w/ Entropy (CLeHe)	**37.62**	**69.56**	**36.12**	37.31	**72.24**	**38.19**	46.32	**78.26**	**49.76**	52.02	81.78	51.62

Table 2: Performance on combining different external and parametric knowledge modeling methods. Experiments are conducted under the top-5 document setting.

from the Wikipedia corpus (Dec. 20, 2018) via as evidence documents for each question. Specifically, we report the performance of different decoding methods when retrieving the top-5, top-10, and top-20 documents. Following [22], exact match accuracy is utilized for performance evaluation.

Overall Performance. Table 1 presents the overall performance comparison between our proposed method and existing baselines on public benchmark datasets. The results show that when compared to the NAIVE method, our entropy-ensemble-based LeEns demonstrates significant average performance improvements across various LLMs, indicating its superior ability to extract useful information from the context. Moreover, LeEns outperforms REPLUG and AvgEns in almost all settings, indicating that using the uncertainty of LLM output distributions for document scoring more effectively facilitates generating answers than static retriever similarity and unweighted averaging. Comparing LeEns with CLeHe, we observe that further contrasting the ensemble-based low-entropy contextual distribution with the high-entropy distribution of the parametric knowledge leads to performance improvements, particularly noticeable in LLAMA2-7B and LLAMA2-13B. These observations substantiate that the proposed entropy-based decoding mechanism markedly augments the extraction and utilization of contextual information. Further, on Mistral-7b-v0.1 and LLAMA-3-8B, CLeHe performs similarly to LeEns, indicating no significant enhancement from the contrastive ensemble. We speculate that these two models are less distracted by parametric knowledge when generating answers.

Ablation Study. Within the contrastive decoding framework, we investigate the compositional effects on performances by combining different modeling techniques for external and parametric knowledge. To extract knowledge from retrieved external documents, we explore two modeling approaches.: NAIVE RAG and our entropy-based document ensemble modeling (LeEns). Additionally, we explore three layer-based strategies to derive parametric knowledge: (i) Last-Layer strategy. It defines parametric knowledge using the distribution from the last layer of LLMs when without retrieved context. CAD [31] utilizes this strategy, *i.e.*, contrasting the distribution derived from NAIVE RAG against the last-layer context-free distribution. (ii) JSD-based strategy. It first calculates the Jensen-Shannon Divergence (JSD) between the RAG-derived distribution and the distribution of each layer when without retrieved documents, then selects the layer with the highest JSD for contrast. (iii) Our proposed entropy-based strategy. It directly selects the layer with the highest entropy as the proxy of intrinsic knowledge. As shown in Table 2, compared to the other two layer selection strategies, the proposed entropy-based strategy consistently and significantly enhances model performance in both external knowledge modeling ways of NAIVE RAG and our LeEns.

Hyper-Parameter and Latency Analysis We study the influence of the introduced hyperparameters: τ and β. As shown in Figure 5, a small value of τ (*e.g.*, 0.1 or 0.25) typically results in better performance; as τ increases, the performance gradually declines. Ideally, when $\tau \to \infty$, the

# of Docs	**5**	**10**	**20**
NAIVE	27.51 (×1.00)	29.49 (×1.00)	32.56 (×1.00)
LeEns	30.43 (×1.11)	32.90 (×1.12)	37.76 (×1.16)
CLeHe	31.88 (×1.16)	34.10 (×1.16)	38.49 (×1.18)

Table 3: Decoding latency (ms/token) of LLAMA-2-7B based on the number of retrieved documents as context.

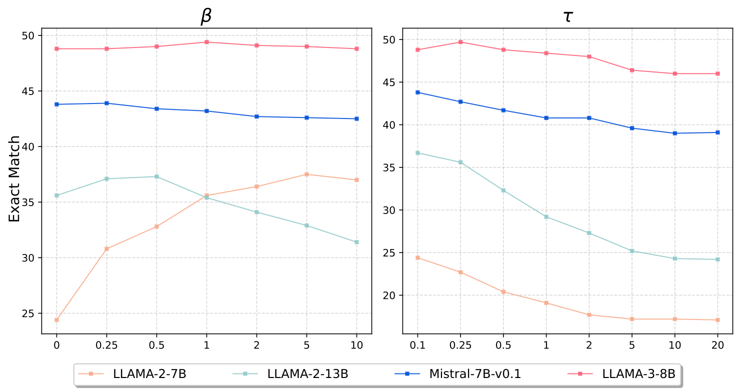


Figure 5: Hyper-parameter analysis using 1K evaluation samples of NQ under the top-5 document setting.

performance of the proposed LeEns will match that of AvgEns. Regarding β, it's observed that for LLAMA-2-7B, a high β (e.g., 5) enables it to effectively contrast the differences between external and parametric knowledge for improved performance. For other evaluated LLMs, we suggest setting it to a small value, saying [0.25, 0.5].

As for decoding latency, Table 3 shows that compared to the NAIVE method, our LeEns and CLeHE increase the decoding time by factors of less than 1.18, indicating that they can be applied at a reasonable cost.

C Limitations

One limitation of our study is that we only validated the effectiveness of our method on question answering datasets, without testing it on other knowledge-intensive tasks such as fact verification. Extending the method proposed in this paper to other retrieval-augmented scenarios will be a future research direction.

Additionally, due to computational power constraints, we only tested the effectiveness of the proposed method on models with fewer than 13B parameters. However, whether the method proposed in this paper is applicable to LLMs with more parameters (*e.g.*, 70B or more) remains to be explored in future research.