ECD: Evidence-guided Contrastive Decoding in Retrieval-Augmented Generation with Accurate Knowledge Reference Adjustment

Yize Sui[1], Yan Xu[1], Kun Hu[1], Jing Ren[1*], Wenjing Yang[1],

[1]College of Computer Science and Technology,
National University of Defense Technology
{suiyize18, xuyan, khu, renjing, wenjing.yang}@nudt.edu.cn

Abstract

Retrieval-Augmented Generation (RAG) enhances the quality of question answering by integrating external knowledge with internal knowledge. A robust RAG system needs to precisely regulate the dependence of the response on the two types of knowledge. The recently proposed context-aware contrastive decoding (CCD) method attempts to achieve this goal by adjusting the knowledge reference weights by comparing the output distribution differences of LLMS when they rely on different knowledge sources. However, these methods are based on probabilistic knowledge reference adjustment strategies (such as the highest probability or entropy), only focus on the relative confidence of the output responses at each decoding step, without considering the absolute confidence of the responses, which may lead to misjudgment of the external knowledge and internal knowledge reference degree in the decoding process. To this end, we propose a novel decoding method, Evidence-guided Contrastive Decoding (ECD), which conducts evidence modeling by constructing the Dirichlet distribution and regards logits as evidence vectors, so as to regulate the reference degree of internal and external knowledge more accurately, and finally improve the quality of generated responses. Extensive evaluations across four public benchmark datasets on three mainstream LLMs have demonstrated the effectiveness and advantages of ECD.

Introduction

With the rapid development of large language models (LLMs), retrieval-augmented generation (RAG) has played an important role in various fields. It can effectively integrate external knowledge and parametric knowledge to enhance the accuracy and trustworthiness of LLMs in knowledge-intensive tasks. However, affected by the unavailable context and potential knowledge conflicts, how to effectively leverage the two distinct sources of knowledge to maximize performance of LLMs has become a major challenge.

For this problem, recent context-aware contrastive decoding (CCD) provides a training-free solution and inspires many subsequent works. Early research such as CAD (Shi et al. 2024) makes response generation promote greater attention to external knowledge rather than parameter knowledge by directly magnifying the retrieved context influence in the decoding step. COIECD (Yuan et al. 2024) categorizes instances into high and low conflict based on a complex information entropy constraint governed by tuned hyperparameters, and employs different decoding strategies for each. Given that the retrieved context information cannot be used as the gold reference, which may contain query-irrelevant noise, some recent studies attempt to adaptively weight the contrastive contextual influence on the parametric knowledge. For example, MICD (Zhao et al. 2024) combines adversarial independent negative samples and estimates LLM confidence by computing the highest probability from the normalized predicted token probabilities at each step, thus effectively balancing the two knowledge sources. ACD (Kim et al. 2024b) and DVD (Jin et al. 2024) use entropy to measure context relevance and dynamically adjust the weights of the context to adapt to the noisy context environment. CDA (Kim et al. 2024a) further introduces the abstention mechanism, allowing the model to choose not to answer when lacking relevant knowledge. These methods provide a new perspective for the application of contrastive decoding (CD) in the RAG system.

However, such methods typically rely on the probabilistic knowledge reference adjustment strategy (such as highest probability or entropy) in the decoding process, requiring confidence estimation of all possible candidate tokens for the next token. Due to the normalization of the probability distribution, this confidence estimate can only express "who is more likely to be correct", but not "how strong is the evidence supporting that the option is correct", which may lead to misjudgment in the knowledge reference adjustment strategy. Specifically, as shown in Figure 1(a), for the probability distribution of the output generated by RAG using highly relevant knowledge, CCD will increase the weight of its influence on the final probability distribution, otherwise it will decrease. When the answer is clear in the candidate tokens (Figure 1(a) **I**), its probability distribution shows a sharper density, so that the probabilistic uncertainty (such as entropy) will be low, and CCD will favor such knowledge at the current decoding step. When clear answers are lacking, the probability distribution will be flatter (Figure 1(a) **II**) and its uncertainty will be higher. At this point, CCD will reduce the reference degree to this knowledge source. In this case, the probabilistic knowledge reference adjustment strategy is reasonable. However, if there are multiple correct

*Jing Ren is the corresponding author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

candidate tokens, the probability distribution also tends to be flat (Figure 1(a) **III**), and the probabilistic knowledge reference adjustment strategy will produce misjudgment and reduce attention to this knowledge source. This misjudgment also occurs in situations where knowledge is outdated. That is, even if the answer is wrong, there will still be a sharp probability distribution, which leads to excessive attention to it (Figure 1(a) **IV**). Therefore, it is necessary to explore a more effective CCD method to adjust the reference degree of internal and external knowledge more accurately, thereby improving the quality and reliability of the responses generated by the model.

In this paper, we propose a novel decoding method, Evidence-guided Contrastive Decoding (ECD), which can more accurately balance the influence of internal and external knowledge on model prediction during the contrastive decoding process and effectively avoid the misjudgment of the knowledge reference adjustment strategy. Instead of calculating the entropy of the next token, ECD treats logits as parameters of the Dirichlet distribution for evidence modeling to represent uncertainty, which can represent both relative confidence and absolute confidence, so as to correctly guide the model to pay attention to the degree of two kinds of knowledge and improve the reliability of the model response. To verify the effectiveness of ECD, we conduct experiments on mainstream benchmark datasets such as TriviaQA, NQ, PopQA and HotpotQA. The experimental results show that ECD outperforms other baselines in improving the overall performance and can maintain good performance in different quality retrieval scenarios.

Our contributions can be summarized as follows:

- We propose Evidence-guided Contrastive Decoding (ECD), a novel decoding method for RAG that more accurately regulates the influence of internal and external knowledge during the decoding process, thereby improving the quality and reliability of the generated responses.
- ECD constructs a Dirichlet distribution based on logits for evidence modeling, which captures both relative and absolute confidence of the model's predictions. This enables a more nuanced understanding of the model's confidence in its predictions, which is crucial for making informed decisions in RAG.
- Extensive experiments on multiple knowledge-intensive benchmark datasets show that the proposed ECD achieves superior performance over the existing CCD-based methods, demonstrating its superiority.

Related Work

Retrieval-Augmented Generation

Despite significant progress, LLMs still face challenges such as outdated knowledge and hallucinations (Jiang et al. 2024). RAG overcomes the inherent limitations of model parameterized knowledge by integrating retrievable external knowledge, which as non-parametric memory and easily updatable, accommodates extensive long-tail knowledge, significantly improving response quality and reducing hallucinations, especially in knowledge-intensive tasks (Lewis et al. 2020; Asai et al. 2023). The naive RAG consists of two core stages: retrieval and generation. The workflow is that, given an input query, the retriever is responsible for identifying and retrieving relevant information from the data source, and then interacting the retrieved information with the generator to optimize the generation process. Many studies (Sui et al. 2024; Fan et al. 2024) have analyzed and improved these two stages respectively, while we mainly focus on the contrastive decoding (Li et al. 2022) method of the generation stage, which is a lightweight method without training and without inserting additional components, aiming to enable the LLMs to effectively utilize context and parameter knowledge.

Context-Aware Contrastive Decoding

In the field of text generation, contrastive decoding has attracted much attention. Its key advantage lies in being able to steer the model in the desired direction by contrasting different output distributions. Liu et al. (2021) introduced a group of "experts" and an "anti-expert" to guide the generation process, keeping it away from bad attributes. Later, Li et al. (2022) by maximizing the difference in log probabilities between expert and amateur models to enhance open-ended text generation. Recently, inspired by context-aware contrastive decoding (Shi et al. 2024), many studies (Qiu et al. 2024; Kim et al. 2024b) have been dedicated to enabling the model to effectively leverage parameter knowledge and context knowledge during the decoding process. The core idea is to evaluate the contributions of the two types of knowledge by using probabilistic uncertainty quantification methods such as entropy, and then introduce controllable weights to control the influence of different knowledge on output distributions. We continue this research direction, but abandon the existing probabilistic knowledge reference adjustment strategy and instead construct a contrastive decoding method through evidence modeling.

Methodology

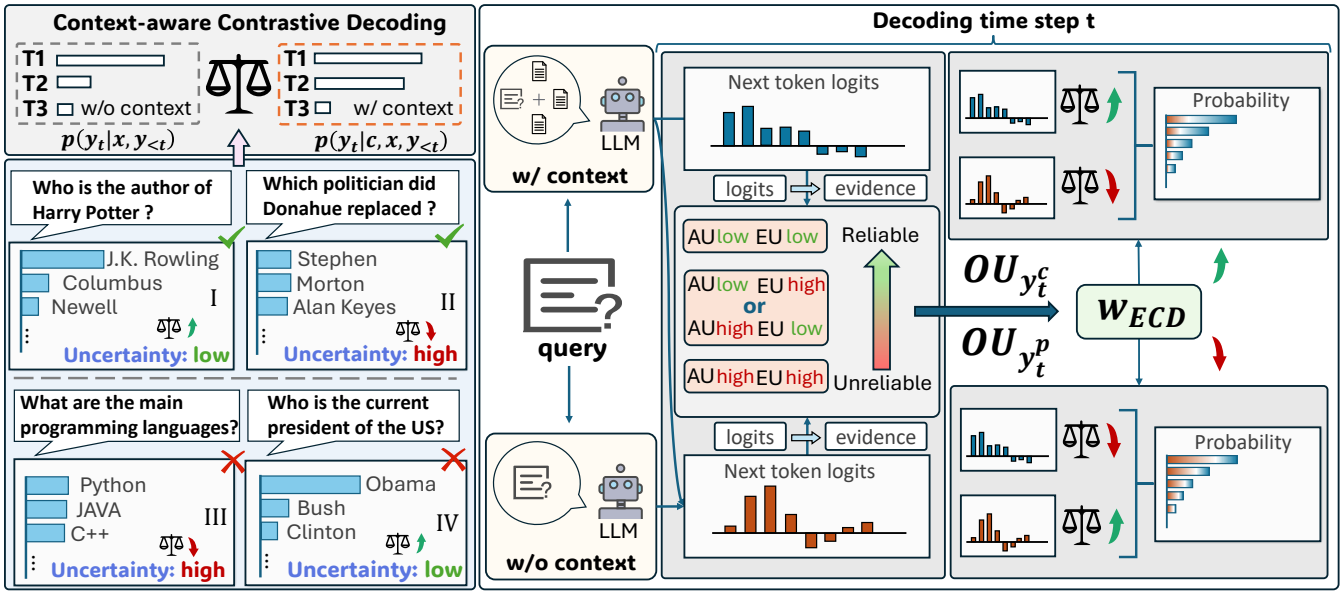
In this section, we introduce a novel contrastive decoding method ECD, which accurately regulates parameters and context knowledge dependence through evidence modeling. First, we describe the advantages of evidence modeling in capturing uncertainty and how to construct evidence modeling suitable for CCD. Further, we will elaborate on the specific process of evidence-guided contrastive decoding.

Problem Formulation

At present, an auto-regressive LLM generates the sequence y token-by-token. At decoding time step t, LLM θ calculates logits $\mathbf{z}_t \in \mathbb{R}^{|V|}$ for next token y_t based on the query x and the generated historical sequence $y_{<t} = \{y_1, y_2, ..., y_{t-1}\}$, where V is the vocabulary. The normalized probability distribution of the t-th token is calculated as follows:

$$p_\theta\left(y_t \mid x, y_{<t}\right) = \operatorname{softmax}\left(\mathbf{z}_t\right) \tag{1}$$

In the decoding stage, existing LLMs provide multiple strategies to select the next token $y_t \sim p_\theta\left(y_t \mid x, y_{<t}\right)$ optimally based on the probability distribution. This iterative process continues until the generated sequence triggers the



(a) Example of knowledge reference adjustment strategy (b) The overall process of Evidence-guided Contrastive Decoding (ECD) in LLM outputs

Figure 1: Overview of our ECD. **Figure 1(a)** shows an example of the knowledge reference adjustment strategy for the existing CCD method, which in some cases (**III**, **IV**) are not accurate. **Figure 1(b)** shows the overall process of ECD in LLM outputs.

termination marker or meets the preset stop criteria. Based on this mechanism, the output can be effectively controlled by adjusting the probability distribution of the next token during the decoding process.

Evidence Modeling via Logits

As shown in the example of Figure 1(a) III and Figure 1(a) IV, in the decoding stage of the CCD, estimating the uncertainty of the LLM through the probability distribution of candidate tokens at each decoding step is not always reliable. Multiple correct candidate tokens or outdated knowledge may both lead to misjudgment of the probabilistic knowledge reference adjustment strategy. In this work, we employ a more accurate uncertainty estimation method to guide the knowledge reference during the CCD process.

Recent Evidential Deep Learning (EDL) theory (Ulmer, Hardmeier, and Frellsen 2021; Ma et al. 2025) indicates that logits can be regarded as the evidence strength information accumulated by LLMs during the training process, and used this as parameters to construct the Dirichlet distribution for evidence modeling, so as to decouple the uncertainty into tokens relative aleatoric uncertainty (AU) and model inherent epistemic uncertainty (EU). Inspired by this, we construct a novel contrastive decoding method based on uncertainty estimation guided by evidence modeling. First, we use logits as parameters of the Dirichlet distribution for evidence modeling. Specifically, all non-negative logits are regarded as valid evidence, while negative values are considered as no evidence. The higher the strength of the evidence, the more frequently the model is exposed to similar scenarios during the pre-training or fine-tuning stage, so logits can effectively reflect the absolute confidence level of candidate tokens. Therefore, we select the logits of tokens with the top-K largest logits as concentration parameters (evidence vector) $\boldsymbol{\alpha} = \{\alpha_1, ..., \alpha_k, ..., \alpha_K\}$ of the Dirichlet distribution, and most candidate tokens with extremely low logits are discarded as noise:

$$\text{Dirichlet}(\boldsymbol{\alpha}), \quad \boldsymbol{\alpha} = sort_{desc}(ReLU(\mathbf{z}_t))\,[1:K]\,. \tag{2}$$

Here, we set logits with negative values to no evidence by applying the activation function $ReLU$.

Evidence-guided Contrastive Decoding

Uncertainty Estimation by Evidence. In fact, the probability distribution only captures the relative confidence of the LLMs in predicting the next token. For example, the logits vectors [10, 8, 5] and [82, 80, 77] have the same probability distribution [0.8756, 0.1185, 0.0059] when mapped by the softmax function, but according to EDL theory (Ulmer, Hardmeier, and Frellsen 2021), the evidence strength information contained in the two is actually different. The higher evidence (logits) indicates a higher absolute confidence in the prediction result, and the corresponding answer is also more likely to be the correct one. Therefore, knowledge with low relative confidence but high evidence strength information (such as Figure 1(a) III) should not be ignored in the contrastive decoding process. This indicates that uncertainty estimation needs to take into account not only the relative confidence of candidate tokens but also the absolute confidence of the candidate tokens. In order to solve this limitation, this work estimate the AU and EU respectively based on Dirichlet framework. Figure 2 shows the logits distribution under pairwise uncertainty condition combinations.

Aleatoric Uncertainty (Relative Confidence). The measurement of relative confidence is based on the expected entropy of the data distribution, which represents the "diver-

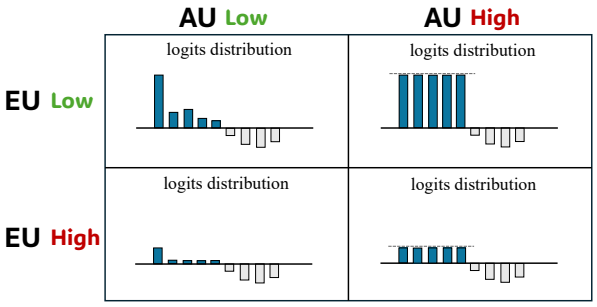


Figure 2: Illustration of logits distribution cases under different epistemic and aleatoric uncertainty conditions.

gence degree" of the model's prediction distribution for the next token. As the entropy captures the "peakiness" of the output distribution, a lower entropy (low AU) indicates that the model is concentrating most probability mass on a single class, while high entropy (high AU) represents a more uniform distribution, indicating that the model has less relative confidence in its predictions. For Dirichlet networks, this quantity has a closed-form solution (Ulmer, Hardmeier, and Frellsen 2021):

$$\mathrm{AU}_{y_t}(\boldsymbol{\alpha}) = -\sum_{k=1}^{K} \frac{\alpha_k}{\alpha_s} \left(\psi\left(\alpha_k + 1\right) - \psi\left(\alpha_s + 1\right)\right) \quad (3)$$

where $\alpha_k \in \boldsymbol{\alpha}$, $\alpha_s = \sum_{k=1}^{K} a_k$ is the total evidence, and ψ denotes the digamma function defined as $\frac{d}{dx} \log \Gamma(x)$.

Epistemic Uncertainty (Absolute Confidence). EU measures the "knowledge depth" of the candidate tokens. Its underlying intuition is that larger evidence produces a sharper density, which indicates increased absolute confidence in a prediction. The calculation of EU is based on the total evidence of the Dirichlet distribution. The stronger the total evidence is, the higher the absolute confidence of the candidate tokens will be. The definition of EU is as follows:

$$\mathrm{EU}_{y_t}(\boldsymbol{\alpha}) = K / \sum_{k=1}^{K} \left(\alpha_k + 1\right). \quad (4)$$

To integrate the uncertainty of the two dimensions, we combine AU and EU to represent the overall uncertainty. Its definition is as follows:

$$\mathrm{OU}_{y_t}(\boldsymbol{\alpha}) = \mathrm{AU}_{y_t}(\boldsymbol{\alpha}) \cdot \mathrm{EU}_{y_t}(\boldsymbol{\alpha}). \quad (5)$$

Knowledge Reference adjustment strategy of ECD. Since the quality of retrieval context in different RAG scenarios is difficult to guarantee, a robust CCD method needs to effectively control the preference for internal and external knowledge. At decoding step t, we distinguish between the logits distribution with parameter knowledge $\mathbf{z}_t^p$ and the logits distribution with context knowledge $\mathbf{z}_t^c$:

$$\mathbf{z}_t^p = \mathrm{logit}_\theta\left(y_t^p \mid x, y_{<t}^p\right) \quad (6)$$

$$\mathbf{z}_t^c = \mathrm{logit}_\theta\left(y_t^c \mid c, x, y_{<t}^c\right) \quad (7)$$

where the main difference between the latter and the former is that the retrieval context is incorporated into the input of the LLM. The evidence vectors corresponding to $\mathbf{z}_t^p$ and $\mathbf{z}_t^c$ are denoted as $\boldsymbol{\alpha}_t^p$ and $\boldsymbol{\alpha}_t^c$ respectively. Moreover, the generated historical sequence and the token at the t-th decoding step of the two cases are denoted as $y_{<t}^p$, y_t^p and $y_{<t}^c$, y_t^c respectively. The probability distribution is modified by weighted adjustment based on the difference between $\mathbf{z}_t^p$ and $\mathbf{z}_t^c$. Equation (1) becomes:

$$p_\theta\left(y_t \mid x, y_{<t}\right) = \mathrm{softmax}\left(\mathbf{z}_t^p + w_{ECD}\left(\mathbf{z}_t^c - \mathbf{z}_t^p\right)\right), \quad (8)$$

where w_{ECD} is a weight hyperparameter that controls the knowledge reference between parameters and context.

The basic motivation for the tunable weight w_{ECD} is that when the context makes no substantial contribution to solving the problem (increasing uncertainty), a lower weight should be assigned. Conversely, when the context can effectively assist in problem-solving (reducing uncertainty), a higher weight will be assigned. From Equation (5), the uncertainty estimation of decoding based on parameter knowledge and context knowledge at t-th step can be denoted as $\mathrm{OU}_{y_t^p}$ and $\mathrm{OU}_{y_t^c}$. Therefore, we set w_{ECD} as follows:

$$w_{ECD} = \frac{\mathrm{OU}_{y_t^p}(\boldsymbol{\alpha}_t^p)}{\mathrm{OU}_{y_t^p}(\boldsymbol{\alpha}_t^p) + \mathrm{OU}_{y_t^c}(\boldsymbol{\alpha}_t^c)}. \quad (9)$$

As the decoding steps proceed, the weight w_{ECD} continuously optimizes the probability distribution through dynamic adjustment, and the candidate token with the highest probability will be selected as the next token. This process is lightweight and automated, and the optimal decision can be made without human intervention.

Experiments

Experiment Setup

Datasets and Metrics. We conduct experiments on multiple open-domain QA datasets (Chen et al. 2017), including TriviaQA (Joshi et al. 2017), Natural Questions (Kwiatkowski et al. 2019), PopQA (Mallen et al. 2022) and HotpotQA (Yang et al. 2018).

TriviaQA comprises trivia questions sourced from the web, we sample 1K instances from the TriviaQA Wiki validation set for experiments. NQ consists of questions derived from actual Google search queries, and we test on 3231 instances from the NQ validation set. PopQA is a novel entity-centric open-domain QA dataset including long-tail knowledge often overlooked in other popular QA datasets, and we choose 1.6K instances from the PopQA test set. HotpotQA is a QA dataset that requires multi-hop reasoning, where the model needs to find and combine information from multiple sources to answer complex questions. We use the entire development set of HotpotQA, consisting of 7.4K instances. Additionally, we evaluate on knowledge conflict dataset, NQ-SWAP (Longpre et al. 2021), which is based on the NQ dataset and consists of synthetic conflicting data. Following previous work (Kim et al. 2024b), we use the Exact Match (EM) for evaluating the QA performance of LLMs.

Baselines. To verify the performance of ECD, we selected multiple baselines for comparison. First, we select the

Model	Method	TriviaQA ↑	NQ ↑	PopQA ↑	HotpotQA ↑	Average ↑
LLaMA-2-7B-Chat	Reg_{Opn}	60.21	44.14	37.85	37.91	45.03
	Reg_{Cls}	58.35	35.87	26.65	30.26	37.78
	CAD	56.82	43.38	33.74	32.18	41.53
	COIECD	59.63	36.56	40.82	39.25	44.07
	DVD_{Dyn}	61.45	<u>45.67</u>	**43.37**	40.68	47.79
	ACD	<u>63.32</u>	45.31	41.92	<u>41.52</u>	<u>48.02</u>
	CDA	62.19	44.45	39.61	39.31	46.39
	ECD	**65.26**	**47.78**	<u>42.93</u>	**43.38**	**49.84**
LLaMA-2-13B-Chat	Reg_{Opn}	63.22	46.33	39.95	38.29	46.95
	Reg_{Cls}	61.48	37.71	30.24	31.63	40.57
	CAD	60.73	45.85	35.48	34.51	44.14
	COIECD	62.59	46.32	40.83	40.55	47.57
	DVD_{Dyn}	65.35	<u>47.32</u>	42.03	40.92	48.91
	ACD	<u>66.51</u>	47.18	<u>42.72</u>	**43.02**	<u>49.86</u>
	CDA	65.27	45.29	40.35	41.27	48.05
	ECD	**68.78**	**49.52**	**44.08**	<u>42.91</u>	**51.32**
LLaMA-3-8B-Instruct	Reg_{Opn}	62.49	45.24	40.03	41.29	47.26
	Reg_{Cls}	61.56	36.54	30.36	35.94	41.10
	CAD	60.42	43.27	35.27	39.69	44.66
	COIECD	62.11	45.36	40.16	41.67	47.33
	DVD_{Dyn}	<u>66.23</u>	<u>46.67</u>	41.57	<u>43.86</u>	<u>49.58</u>
	ACD	66.12	46.28	<u>42.05</u>	43.29	49.44
	CDA	63.28	44.36	40.62	41.05	47.33
	ECD	**67.89**	**48.86**	**43.59**	**47.23**	**51.89**

Table 1: Experimental results of different methods on four datasets. The experiments were conducted respectively on three LLMs with different scales. For each dataset, the best method is highlighted in **bold**, and the second-best method is <u>underlined</u>.

regular decoding with greedy decoding as the fundamental baselines and set two situations: open book (Reg_{Opn}) and closed book (Reg_{Cls}). Furthermore, we compare our method against existing CCD-based methods, including Context Aware Decoding (CAD) (Shi et al. 2024), Contextual Information-Entropy Constraint Decoding (COIECD) (Yuan et al. 2024), Dynamic Contrastive Decoding (DVD) (Jin et al. 2024), Adaptive Contrastive Decoding (ACD) (Kim et al. 2024b) and Contrastive Decoding with Abstention (CDA) (Kim et al. 2024a). It is worth noting that we adopt the best hyperparameter for the previous work. For CAD, we set $\alpha = 1$, and for the COIECD, the values of λ and α are set to 0.25 and 1. For DVD, we consider the dynamic weight variation and denote it as DVD_{Dyn}.

Models. Our experiments evaluate the performance of ECD on three different scales of popular open source LLMs, LLaMA-2-7B-Chat, LLaMA-2-13B-Chat (Touvron et al. 2023) and LLaMA-3-8B-Instruct (Grattafiori et al. 2024). Specifically, we utilize Contriever-MSMARCO (Izacard et al. 2021) as a retriever, and the top-1 context is retrieved from the Wikipedia contexts [1]. Additionally, the TriviaQA, NQ and HotpotQA datasets all provide gold contexts, which we employ to measure the theoretical upper bound of our proposed decoding method. All experiments are conducted on a single NVIDIA A100 GPU.

[1]We utilize the Wikipedia dump from 2018.

Experimental Results and Analyses

Main Result. Table 1 presents the performance comparison of our proposed method with various CCD-based baseline methods on four public benchmark datasets, from which we can draw the following conclusions. Employing regular decoding within an open-book setting consistently outperforms the closed-book setting across most models. Since the Reg_{Cls} only relies on parameter knowledge to answer questions, the performance deficiency is more obvious on the PopQA dataset that focuses on long-tail knowledge. This inclination suggests that non-parametric knowledge is important for LLMs to complete knowledge-intensive tasks. In contrast, CAD emphasizes the focus on external knowledge, and its performance is also not ideal because the retrieved documents are not always gold contexts. Notably, CCD-based methods with dynamic weight adjustment, such as ACD, DVD_{Dyn}, CDA, and the ECD we proposed, outperform Reg_{Opn} and Reg_{Cls} in almost all settings, indicating that compared with static retrieval of external documents or direct decoding relying solely on parameter knowledge, the contrastive decoding method that flexibly utilizes internal and external knowledge can more effectively improve the quality of generated answers. By comparing ECD with DVD_{Dyn} and ACD, we can observe the performance differences between the evidence-based contrastive decoding method and the entropy-based contrastive decoding method. Overall, ECD performs better on most datasets compared to

Method	TriviaQA	NQ	HotpotQA
Gold Contexts			
Reg_{Opn}	84.25	**62.52**	**51.45**
CAD	73.27	57.83	41.42
COIECD	81.48	59.87	45.62
DVD_{Dyn}	84.64	61.92	49.26
ACD	84.38	62.04	48.37
CDA	83.21	62.11	47.52
ECD	**85.36**	62.20	50.82
Noisy Contexts			
Reg_{Opn}	37.55	22.42	25.54
CAD	25.73	24.53	32.26
COIECD	30.81	39.67	35.46
DVD_{Dyn}	55.72	41.61	38.44
ACD	56.38	42.91	37.69
CDA	52.31	42.42	36.55
ECD	**62.30**	**45.20**	**40.82**

Table 2: Experimental results of different CCD methods on the TriviaQA, NQ and HotpotQA datasets using LLaMA-3-8B-Instruct when providing **Gold** and **Noisy** contexts.

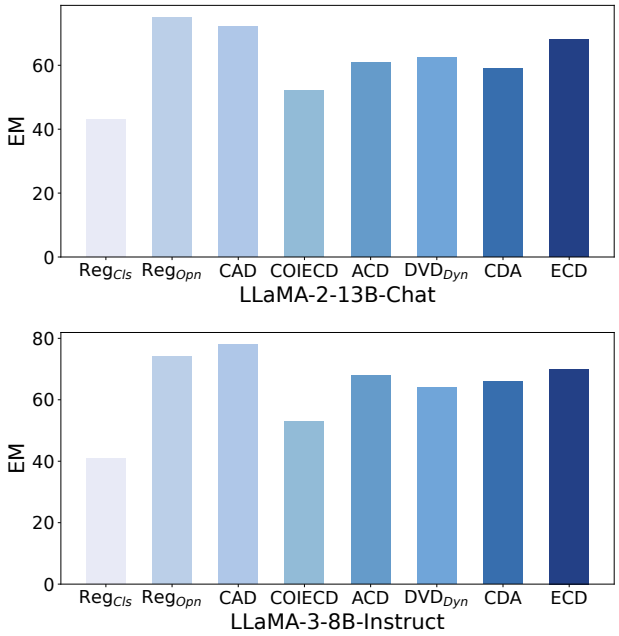


Figure 3: Performance evaluation of different CCD methods on knowledge conflict datasets NQ-SWAP.

DVD_{Dyn} and ACD, especially in the LAMa-3-8B-instruct, where its average performance on the four datasets TriviaQA, NQ, PopQA and HotpotQA is about 4.7% higher than the second-best method. Under the same conditions, compared with CDA, its average performance improvement reaches 9.6%. These findings confirm that the proposed ECD method can more effectively control knowledge dependence to utilize parameter knowledge and contexts for high-quality generation.

Performance in Different Quality Retrieval Scenarios. To verify that ECD can guide the knowledge reference in the decoding process more accurately, we analyzed the performance in three different quality retrieval scenarios. Specifically, in the experiment in Table 1, we retrieve the Wikipedia and select the top-1 relevant document as the context for performance evaluation to simulate the real RAG scenarios. For comparison, we consider the retrieved 10-th ranked document as an irrelevant noisy context, simulating the low-quality retrieval scenario. In addition, we use the gold contexts provided by the NQ, TriviaQA, and HotpotQA datasets as relevant contexts (PopQA does not provide gold contexts) to simulate the high-quality retrieval scenario. Here, we exclude the conflicting context from the noisy context and discuss it in the next section. The noisy context refers to the context that is irrelevant to the query. Methods that can guide knowledge reference more accurately should demonstrate better performance in various quality retrieval scenarios.

Since the change of context does not affect the performance of Reg_{Cls}, we only present the results of other methods here. The experimental results in high and low quality retrieval quality scenarios are shown in Table 2, while the performance in real retrieval scenarios is presented in Table 1. It can be seen that Reg_{Opn} significantly outperforms other baseline methods with gold context, but its performance is poor in low-quality retrieval scenarios. This confirms that the retrieval quality does indeed significantly affect the quality of the responses generated by RAG. In contrast, ECD performs well in different quality retrieval scenarios. Specifically, ECD performs well enough to be comparable to Reg_{Opn}in the high-quality retrieval scenario. For example, on the TriviaQA dataset, ECD is 1.3% higher than Reg_{Opn}, while on the NQ and PopQA, it is only 0.5% and 1.2% lower than Reg_{Opn} respectively. In low-quality retrieval scenarios, ECD is significantly superior to other baseline methods which outperforming the suboptimal method by 10.5%, 5.3%, and 6.2% on the three datasets, respectively. It can be seen that ECD can guide the knowledge reference during contrastive decoding more accurately.

Knowledge Conflict Analyses. As highlighted in the previous section, tasks reliant on knowledge typically draw from two knowledge sources: parametric knowledge within the LLMs and non-parametric knowledge obtained through retrieval during inference. The issue of knowledge conflicts, wherein the non-parametric knowledge contradicts the parametric knowledge, is worthy of in-depth exploration.

In order to evaluate the ability of ECD to handle knowledge conflicts, we evaluate on NQ-SWAP (Longpre et al. 2021) which introduces synthetic conflicts by substituting entities in the gold document with alternate entities to challenge the model's ability to manage conflicting information. We utilize the questions and entity-swapped contexts provided in (Hong et al. 2024), which includes 3,650 samples. In this setting, NQ-SWAP evaluates whether LLM produces the substituted entity as the answer when provided with the question and the modified context, disregarding its

pre-learned parametric answer. The performance results on NQ-SWAP are shown in Figure 3, where we find that CAD specifically designed for high knowledge conflict scenarios exhibits better performance than other methods, but this ability compromises its ability in low knowledge conflict scenarios (as shown in Table 1). It is worth noting that, due to the task requiring the model to generate the replaced entities, all models demonstrated poor performance in the regular closed-book setting. However, the ECD method we proposed effectively achieves performance balance in scenarios with both high and low knowledge conflicts. As shown in Figure 3, the performance of ECD on the two LLMs consistently exceeds that of other contrastive decoding methods except CAD and achieves results comparable to open-book regular decoding. The comparison results emphasize the effectiveness of the ECD in resolving knowledge conflicts.

w_{ECD}	TriviaQA	NQ	PopQA	HotpotQA
LLaMA-3-8B-Instruct				
0.0	61.56	36.54	30.36	35.94
0.2	61.95	38.26	33.84	37.19
0.4	<u>62.63</u>	42.41	38.41	40.37
0.6	61.40	44.59	<u>42.73</u>	<u>42.86</u>
0.8	62.03	44.85	40.07	41.55
1.0	62.49	<u>45.24</u>	40.03	41.29
ECD	**67.89**	**48.86**	**43.59**	**47.23**

Table 3: Experimental results for ablation studies of w_{ECD}.

Ablation Studies of w_{ECD}. To evaluate the impact of w_{ECD} on performance, we fix the value of w_{ECD} within a range [0, 1] and examine whether employing ECD is more effective than optimizing a fixed weight. The retrieval scenario of this experiment is consistent with that in Table 1. The experimental results are shown in Table 3, starting from $w_{ECD} = 0$, increasing the value of w_{ECD} makes the output distribution more influenced by contextual information. With $w_{ECD} = 0$ and $w_{ECD} = 1$, the ECD degenerates to Reg_{Cls} and Reg_{Opn}, respectively. We observe that increasing the value of w_{ECD} improves the EM score to some extent compared to Reg_{Cls}, but the performance with fixed w_{ECD} values is consistently inferior to that of the ECD method employing dynamic weight adjustment. This indicates that in scenarios with potential noisy context, a fixed w_{ECD} may not ensure optimal performance. Therefore, it is necessary to enable the contrastive decoding process to dynamically adjust its dependence on parameter knowledge and context knowledge during inference.

Ablation Studies of K. Since the number of candidate tokens generated by an LLM is significantly large (depending on the size of the vocabulary), with a considerable proportion of tokens having extremely low logits, ECD focuses on the distribution of the main candidate tokens with the top-K largest logits. To evaluate the impact of parameter K (different numbers of candidate tokens) on the performance of ECD, we set different K for ablation experiments. The results are shown in Figure 4 which presents the performance evaluation with K values ranging from 1 to 50 for different models on four datasets. It can be seen that for all models and datasets, the EM score increases rapidly as K increases from 1 to around 25. For example, on TriviaQA, the EM score increases from about 57 at K = 1 to approximately 65 at K = 25 on LLaMA-2-7B-Chat. Similarly, the EM score on NQ increases from around 43 at K = 1 to nearly 49 at K = 25 on LLaMA-2-13B-Chat. After K = 25, although there is still some improvement in EM score as K increases further, the rate of improvement slows down significantly. For instance, on PopQA, the EM score increases from roughly 43 at K = 25 to about 44 at K = 50 on LLaMA-3-8B-instruct. This pattern is consistent across all models and datasets. Therefore, considering the trade-off between performance gain and computational cost, we conclude that setting K = 25 is sufficient for achieving a good balance between performance and efficiency. All other experiments in this paper uniformly adopt K = 25 as the standard configuration.

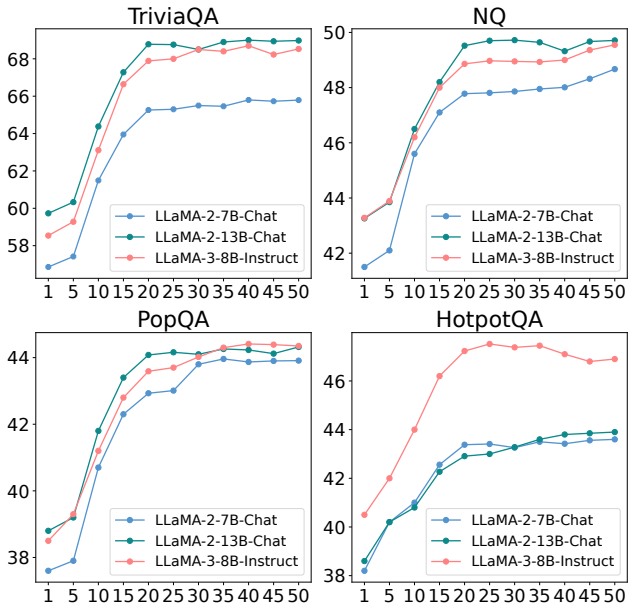


Figure 4: Performance evaluation (EM score) across K values ranges from 1 to 50.

Conclusion

In this paper, we propose ECD, a novel decoding method for RAG. By constructing a Dirichlet distribution based on logits for evidence modeling, ECD guides the model to dynamically balance knowledge dependencies more accurately during decoding. Our experiments across multiple datasets demonstrate that ECD significantly outperforms baseline methods, particularly in scenarios with varying retrieval quality. Results show that ECD can effectively improve the reliability and quality of the responses generated by the RAG, providing a robust solution for optimizing the collaboration between different knowledge sources in RAG.

Acknowledgments

This research is supported by the Natural Science Foundation of China (Nos. 62372459).

References

Asai, A.; Min, S.; Zhong, Z.; and Chen, D. 2023. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, 41–46.

Chen, D.; Fisch, A.; Weston, J.; and Bordes, A. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Fan, W.; Ding, Y.; Ning, L.; Wang, S.; Li, H.; Yin, D.; Chua, T.-S.; and Li, Q. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6491–6501.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hong, G.; Gema, A. P.; Saxena, R.; Du, X.; Nie, P.; Zhao, Y.; Perez-Beltrachini, L.; Ryabinin, M.; He, X.; Fourrier, C.; et al. 2024. The Hallucinations Leaderboard–An Open Effort to Measure Hallucinations in Large Language Models. *arXiv preprint arXiv:2404.05904*.

Izacard, G.; Caron, M.; Hosseini, L.; Riedel, S.; Bojanowski, P.; Joulin, A.; and Grave, E. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Jiang, X.; Tian, Y.; Hua, F.; Xu, C.; Wang, Y.; and Guo, J. 2024. A survey on large language model hallucination via a creativity perspective. *arXiv preprint arXiv:2402.06647*.

Jin, J.; Wang, H.; Zhang, H.; Li, X.; and Guo, Z. 2024. DVD: Dynamic Contrastive Decoding for Knowledge Amplification in Multi-Document Question Answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 4624–4637.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Kim, H. J.; Kim, Y.; Lee, S.-g.; and Kim, T. 2024a. When to Speak, When to Abstain: Contrastive Decoding with Abstention. *arXiv preprint arXiv:2412.12527*.

Kim, Y.; Kim, H. J.; Park, C.; Park, C.; Cho, H.; Kim, J.; Yoo, K. M.; Lee, S.-g.; and Kim, T. 2024b. Adaptive Contrastive Decoding in Retrieval-Augmented Generation for Handling Noisy Contexts. *arXiv preprint arXiv:2408.01084*.

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

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.

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

Liu, A.; Sap, M.; Lu, X.; Swayamdipta, S.; Bhagavatula, C.; Smith, N. A.; and Choi, Y. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. *arXiv preprint arXiv:2105.03023*.

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

Ma, H.; Chen, J.; Wang, G.; and Zhang, C. 2025. Estimating LLM Uncertainty with Logits. *arXiv preprint arXiv:2502.00290*.

Mallen, A.; Asai, A.; Zhong, V.; Das, R.; Khashabi, D.; and Hajishirzi, H. 2022. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*.

Qiu, Z.; Ou, Z.; Wu, B.; Li, J.; Liu, A.; and King, I. 2024. Entropy-based decoding for retrieval-augmented large language models. *arXiv preprint arXiv:2406.17519*.

Shi, W.; Han, X.; Lewis, M.; Tsvetkov, Y.; Zettlemoyer, L.; and Yih, W.-t. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, 783–791.

Sui, Y.; Ren, J.; Tan, H.; Chen, H.; Li, Z.; and Wang, J. 2024. Enhancing LLM's Reliability by Iterative Verification Attributions with Keyword Fronting. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 251–268. Springer.

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

Ulmer, D.; Hardmeier, C.; and Frellsen, J. 2021. Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. *arXiv preprint arXiv:2110.03051*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Yuan, X.; Yang, Z.; Wang, Y.; Liu, S.; Zhao, J.; and Liu, K. 2024. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. *arXiv preprint arXiv:2402.11893*.

Zhao, Z.; Monti, E.; Lehmann, J.; and Assem, H. 2024. Enhancing contextual understanding in large language models through contrastive decoding. *arXiv preprint arXiv:2405.02750*.