

BEE-RAG: Balanced Entropy Engineering for Retrieval-Augmented Generation

Yuhao Wang^{1*†} Ruiyang Ren^{1*} Yucheng Wang² Jing Liu^{2‡}
Wayne Xin Zhao^{1‡} Hua Wu³ Haifeng Wang²

¹Gaoling School of Artificial Intelligence, Renmin University of China

²Baidu Inc.

{yh.wang500, reyon_ren}@outlook.com, batmanfly@gmail.com

Abstract

With the rapid advancement of large language models (LLMs), retrieval-augmented generation (RAG) has emerged as a critical approach to supplement the inherent knowledge limitations of LLMs. However, due to the typically large volume of retrieved information, RAG tends to operate with extremely long context lengths. From the perspective of entropy engineering, we identify unconstrained entropy growth and attention dilution due to long retrieval context as significant factors affecting RAG performance. In this paper, we propose the balanced entropy-engineered RAG (BEE-RAG) framework, which improves the adaptability of RAG systems to varying context lengths through the principle of entropy invariance. By leveraging balanced context entropy to reformulate attention dynamics, BEE-RAG separates attention sensitivity from context length, ensuring a stable entropy level. Building upon this, we introduce a zero-shot inference strategy for multi-importance estimation and a parameter-efficient adaptive fine-tuning mechanism to obtain the optimal balancing factor for different settings. Extensive experiments across multiple RAG tasks demonstrate the effectiveness of BEE-RAG.

Introduction

Retrieval-augmented generation (RAG) has emerged as a transformative paradigm for augmenting large language models (LLMs) with dynamically integrated external knowledge (Gao et al. 2023; Lewis et al. 2020). While conventional RAG systems demonstrate remarkable capabilities in knowledge-intensive tasks, their performance exhibits critical fragility when processing extended contextual inputs (Ren et al. 2023). This vulnerability arises from the inherent requirement of RAG frameworks to incorporate external retrieved documents, which typically results in substantially longer input sequences compared to standard generation tasks (Zhang et al. 2024a). This limitation manifests acutely in scenarios demanding long-context comprehension, such as scientific literature analysis, multi-hop reasoning, and cross-document synthesis, where the growing complexity of modern knowledge systems necessitates robust processing of extended text spans (Liu et al. 2023).

* Equal contributions.

† The work was done during the internship at Baidu.

‡ Corresponding authors.

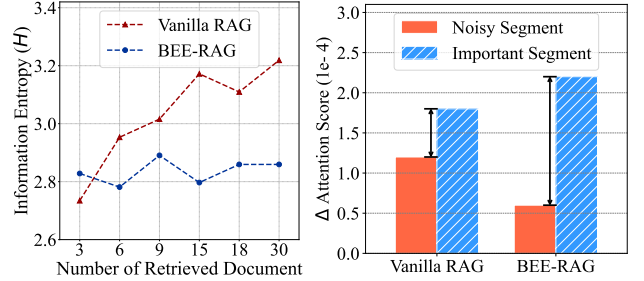


Figure 1: In vanilla RAG, increasing input context length raises the information entropy of attention scores (left), potentially harming performance, while the LLM is less focused on important segments (right). This paper proposes a balanced entropy engineering strategy, which maintains entropy stability for longer contexts and guide the LLM to focus on critical segments.

Typically, existing efforts to mitigate this challenge focus predominantly on heuristic document filtering, extrapolative training, or architectural modifications, introducing worth noting trade-offs. Threshold-based retrieval document truncation or selection reduces context length at the cost of discarding potentially relevant information (Jeong et al. 2024; Wang et al. 2024). While extended context training may enhance RAG performance (Lin et al. 2024; Zhang et al. 2024a), it substantially increases computational demands and risks compromising model generalization. The introduction of auxiliary modules, such as specialized encoders (Yen, Gao, and Chen 2024), further escalates the complexity of the system.

As shown in Figure 1, we identify a fundamental oversight in existing explorations: the failure to address the core tension between context entropy growth and attention allocation dynamics. As retrieved document length increases, unconstrained context entropy expansion progressively dilutes attention distributions, undermining LLMs’ capacity to prioritize semantically critical content. Such a phenomenon substantially degrades RAG’s ability to extract and utilize salient information.

In this work, we fundamentally rethink RAG’s context length adaptability through the perspective of entropy engineering, and propose Balanced Entropy-Engineered RAG (BEE-RAG), a principled framework that establishes

entropy-invariant dynamics to decouple attention sharpness from context length. First, we introduce balanced context entropy, a novel attention reformulation incorporating document-specific balancing entropy factor β to enforce entropy invariance across variable context lengths. Theoretical analysis demonstrates this mechanism maintains entropy levels within a stable regime for various context scales, contrasting with conventional entropy scaling. Building on this foundation, we develop intrinsic multi-importance inference, a zero-shot strategy that derives balancing entropy factors through context aggregation while preserving inter-document independence via parallel context scoring and generation. To ensure broad applicability, we further propose adaptive balancing factor learning, which is a parameter-efficient tuning method that learns task-specific balancing factor through lightweight linear projections for domain adaptation with minimal parameter updates (0.014% of total parameters).

Empirical analyses across multiple real-world benchmarks reveal that BEE-RAG fundamentally alters the entropy scaling laws governing conventional RAG systems, which provides a robust and scalable solution for RAG systems dealing with long-context scenarios. Our main contributions are summarized as follows:

- We introduce the concept of balanced context entropy, a novel attention reformulation that ensures entropy invariance across varying context lengths, and allocates attention to important segments. It addresses the critical challenge of context expansion in RAG.
- We propose a zero-shot strategy called *intrinsic multi-importance inference* and parametric-efficient fine-tuning strategy adaptive balancing entropy factor learning to obtain the balancing entropy factor in different scenarios while maintaining efficiency.
- Extensive empirical validation across multiple benchmarks demonstrates the effectiveness and efficiency of BEE-RAG.

Preliminaries

Task Formulation. The task of retrieval-augmented generation (RAG) involves the integration of retrieval-based techniques with generative models to enhance the accuracy and coherence of text generation tasks (Lewis et al. 2020). A typical RAG system consists of a corpus C , a retriever R , and a generative model LLM . Given a user query or input context q , the retriever R retrieves relevant documents D_q from the corpus. The generative model LLM uses both the query q and the retrieved documents to generate a response:

$$a = \arg \max_y P(y \mid LLM(q, D_q)). \quad (1)$$

We focus on how to enhance the multi-document utilization ability of the language model given D_q and q , with an emphasis on improving the LLM’s ability to leverage contextual knowledge, thereby improving its performance on factual generation.

Entropy Engineering. In this work, we reconsider RAG from an information-theoretic perspective and introduce the

concept of *Entropy Engineering*. In the standard Transformer (Waswani et al. 2017), the attention is computed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax} \left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}} \right) \mathbf{V}, \quad (2)$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{n \times d}$ represent the query, key, and value matrices, respectively, and the scaling factor \sqrt{d} stabilizes the dot-product magnitudes. This can be simplified by introducing a scaling parameter $\lambda = \frac{1}{\sqrt{d}}$:

$$a_{i,j} = \frac{e^{\lambda \mathbf{q}_i \cdot \mathbf{k}_j}}{\sum_{j=1}^n e^{\lambda \mathbf{q}_i \cdot \mathbf{k}_j}}. \quad (3)$$

We examine this formula from two perspectives: From the perspective of token t_j being attended to: a larger $a_{i,j}$ indicates that the LLM pays more attention to the information at position i . Therefore, the LLM should be guided to focus on relatively important information. From the perspective of token t_i receiving information, the values of $\{a_{i,j} \mid i \leq j\}$ across all positions collectively determine how much information t_i receives.

Based on this, inspired by existing analyses (Zhang et al. 2024c), we measure the amount of information a position receives using discrete entropy, as shown in the following equation:

$$H_i = - \sum_{j=1}^n a_{i,j} \log a_{i,j}, \quad (4)$$

which quantifies how much information t_i receives from the attention perspective. This insight suggests that LLMs struggle with longer sequences when not trained on them, likely due to the discrepancy in information received by tokens in longer contexts. Based on the previous analysis, the optimization of attention entropy should focus on two aspects:

- The information entropy at positions that are relatively important and likely contain key information should increase.
- The overall total information entropy of the context should ideally remain stable with respect to n , to prevent the model from being biased or misaligned when handling longer contexts.

Methodology

In this section, we first present a brief overview of our method, then we comprehensively elucidate the methodological details from the perspectives of entropy control and specific zero-shot and parametric-efficient fine-tuning strategies.

Overview

The performance of traditional RAG systems often degrades as the input context length increases, primarily due to the escalating context entropy that dilutes attention distributions across retrieved documents. To mitigate this issue, we propose **Balanced Entropy Engineering for RAG (BEE-RAG)**, stabilizing context entropy while adaptively optimizing attention allocation to enhance RAG performance. Our

method is driven by two key insights: (1) controlling context entropy growth is critical for maintaining robustness in RAG systems with extended contexts, and (2) document importance should dynamically guide attention allocation without costly computation or parametric updates.

The BEE-RAG framework introduces three principal innovations. First, we propose *balanced context entropy*, which incorporates an additive *balancing entropy factor* β into attention computation to maintain entropy invariance across varying context lengths. Theoretical analysis confirms that applying specific distributional constraints can effectively eliminate the impact of context length on information entropy. Second, we develop a *intrinsic multi-importance inference* strategy that computes the balancing entropy factor in a zero-shot manner, eliminating the need for auxiliary models or training data. Third, for scenarios requiring domain adaptation, we devise a parameter-efficient *adaptive balancing entropy factor learning* approach that achieves enhanced performance with minimal computational overhead.

Entropy Control

In this section, we first introduce the proposed balancing entropy factor to maintain the stability of context entropy as the input length increases, and then provide a theoretical analysis.

Balanced Context Entropy As outlined in Section , we aim to modify the attention mechanism to achieve two critical objectives: (1) maintaining the overall information entropy without a significant increase as a function of sequence length n , and (2) ensuring that significant positions receive higher attention scores. This entropy control is designed to enable LLM to maintain superior performance when processing extended contexts by dynamically allocating for attention weights.

Concretely, we propose a new conception balanced context entropy from the perspective of entropy engineering to satisfy the above goals by introducing an additive balancing entropy factor $\beta \in \mathbb{R}^n$ into the attention computation:

$$a_{i,j} = \frac{\exp\left(\frac{\mathbf{q}_i \cdot \mathbf{k}_j}{\sqrt{d}} + \beta_i\right)}{\sum_{l=1}^n \exp\left(\frac{\mathbf{q}_i \cdot \mathbf{k}_l}{\sqrt{d}} + \beta_i\right)}. \quad (5)$$

To focus on important documents, we set all tokens of the same chunk to share the same value. In this case, the balancing entropy factor β_i represents the importance of an individual retrieved document according to the query, with more important documents receiving higher scores. This effectively concentrates the LLM’s attention on important documents.

Building upon the proposed conception of BCE, we present two distinct approaches for deriving the balancing entropy factor β and systematically elaborate on their methodologies in Section . Before that, we conduct a rigorous theoretical analysis of BCE’s operational mechanisms to establish the theoretical rationale for its mathematical formulation.

Theoretical Analysis Here, we present a theoretical derivation to substantiate the rationale behind the proposed balanced context entropy and discuss how to set constraints on the balancing entropy factor for different context lengths.

We denote $\xi_{i,j} = \frac{\mathbf{q}_i \cdot \mathbf{k}_j}{\sqrt{d}}$, and combine Equation (3), (4), and (5) to obtain:

$$H_i = \log \sum_{j=1}^n e^{\xi_{i,j} + \beta_i} - \frac{\sum_{j=1}^n e^{\xi_{i,j} + \beta_i} (\xi_{i,j} + \beta_i)}{\sum_{j=1}^n e^{\xi_{i,j} + \beta_i}} \quad (6)$$

We assume that each row of the query matrix $\mathbf{Q} \in \mathbb{R}^{n \times d}$ and key matrix $\mathbf{K} \in \mathbb{R}^{n \times d}$ is an independent, *isotropic sub-Gaussian* vector with zero mean and identity covariance, i.e., $\mathbb{E}[\mathbf{q}_i] = \mathbb{E}[\mathbf{k}_i] = \mathbf{0}$. By the high-dimensional central limit theorem (Bolthausen and Wüthrich 2013), the scaled dot product $\xi_{i,j} = \frac{\mathbf{q}_i \cdot \mathbf{k}_j}{\sqrt{d}}$ converges in distribution to a standard Gaussian, i.e., $\xi \xrightarrow{d} \mathcal{N}(0, 1)$. We further assume that the document-level importance bias follows a Gaussian distribution $\beta \sim \mathcal{N}(\mu, \sigma^2)$ and is independent of ξ . This independence reflects the fact that the same token may appear in documents of varying importance, and thus the semantic similarity (captured by ξ) and document relevance (captured by β) should be modeled as separate random variables. Thus, we have:

$$H_i \approx \log n + \log \mathbb{E}(e^\xi) + \log \mathbb{E}(e^\beta) - \frac{\mathbb{E}(\xi e^\xi) + \mathbb{E}(\beta e^\beta)}{\mathbb{E}(e^\xi) \mathbb{E}(e^\beta)}, \quad (7)$$

After simplification, we obtain:

$$H_i \approx \log n + \frac{1}{2} + \mu + \frac{\sigma^2}{2} - \frac{e^{\frac{1}{2}} + \mu e^{\frac{1}{2}} + (\mu + \sigma^2) e^{\mu + \frac{\sigma^2}{2}}}{e^{\mu + \frac{\sigma^2}{2} + \frac{1}{2}}}. \quad (8)$$

We aim for entropy to remain invariant as the context length n increases, thereby ensuring that the performance of RAG remains unaffected (seen in Section). Thus, μ and σ should satisfy the following formulation:

$$-\log n = \mu + \frac{\sigma^2}{2} - \frac{e^{\frac{1}{2}} + \mu e^{\frac{1}{2}} + (\mu + \sigma^2) e^{\mu + \frac{\sigma^2}{2}}}{e^{\mu + \frac{\sigma^2}{2} + \frac{1}{2}}}. \quad (9)$$

The solution to Equation (9) is shown in Figure 2. When μ is set to a small value, σ approximates a linear relationship with n . Therefore, we need to constrain σ to increase as n grows, ensuring that the context entropy remains stable as the context length increases. Our experimental results confirm that the parameters derived from theory are consistent with those selected empirically.

BEE-RAG Framework

In the preceding section, we introduced balanced context entropy as a methodological enhancement to stabilize the performance of RAG systems. We hereby propose efficient methods to determine the balancing entropy factor for both zero-shot learning and parameter-efficient fine-tuning.

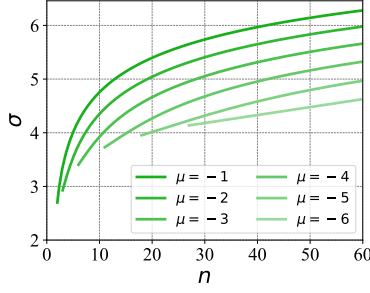


Figure 2: Based on Equation 9, this illustrates how to set μ and σ to balance the entropy increase caused by growing context length n .

Intrinsic Multi-Importance Inference In this part, we introduce the intrinsic multi-importance inference (IMI) approach for zero-shot reasoning in BEE-RAG. The core idea of IMI lies in leveraging the intrinsic parameters of LLM to assess the importance of individual documents within RAG inputs as the balancing entropy factor in BCE computation, thereby eliminating the need for external modules while maintaining computational efficiency and architectural simplicity.

Inspired by the success of prompt engineering for LLMs in retrieval and ranking tasks (Zhuang et al. 2024; Pradeep, Sharifmoghaddam, and Lin 2023), we use a prompt-induced attention calibration strategy to obtain importance scores for each document. Specifically, we append a prompt to each document for obtaining hidden state representations of importance. To extract importance scores from these representations, we apply an early-decoding strategy (Camarata et al. 2020), using `lm_head(·)` to map the generation probability of a critic token as the document’s importance score at each layer. For instance, the prompt can be “Does the passage support the answer to the question?”, the critic token is “yes”.

Since the characteristics of LLM, the derived importance scores show individual distributions in scales and dispersion. To enable stable entropy control across contexts and generalize to different context lengths, we constrain the distribution to be stable with consistent mean and variance, based on the formulation in Equation 9. Concretely, we normalize and rescale the importance scores to a target mean μ and variance σ^2 . This distributional alignment ensures comparability of importance while preserving the intrinsic ranking relationships between documents.

The cross-attention mechanisms employed in LLMs may inadvertently influence importance score estimation for individual documents, as it enables inter-document interference through shared attention parameters. Inspired by previous work (Ratner et al. 2023), we develop a parallel scoring mechanism in RAG architecture through attention mask modifications. This enables simultaneous computation of document importance and task outputs in a single forward pass. In this way, we create independent processing streams that prevent cross-document contamination, which is a critical foundation for stabilizing the context entropy.

Adaptive balancing entropy factor Learning To extend our entropy engineering framework to domain-specific scenarios, we introduce an adaptive balancing entropy factor learning strategy with improved training efficiency. We first introduce designed vector fine-tuning for sentence-level importance scoring. Then, we implement initialization constraints to ensure stable training convergence.

The core of our approach lies in the specially designed efficient vector fine-tuning method, which bridges token-level and sentence-level representations. Specifically, inspired by previous work (Wu et al. 2025), we transform token-level vectors into sentence-level vectors:

$$\mathbf{v}_{\text{sentence}} = \mathbf{v}_b + \mathbf{v}_e + \mathbf{R}^\top (\mathbf{R}\mathbf{v}_b - \mathbf{R}\mathbf{v}_e), \quad (10)$$

where \mathbf{v} represents either key or query vectors, and \mathbf{v}_b , \mathbf{v}_e correspond to the start and end token vectors of documents or queries, respectively. The $\mathbf{R} \in \mathbb{R}^{d \times d_r}$ a trainable projection matrix with $d_r \ll d$. With the sentence-level vectors, we can calculate the importance scores through query-key inner products:

$$\beta = \mathbf{q}_{\text{doc}} \cdot \mathbf{k}_{\text{query}}. \quad (11)$$

During implementation, we observe unconstrained values of β leading to gradient explosion. To address it, we employ orthogonal initialization (Saxe, McClelland, and Ganguli 2013) on \mathbf{R} and implement additional scaling on β when its values exceed a certain threshold.

Discussion

Novelty and Effectiveness. Our key innovation lies in introducing an additive balancing entropy factor into the attention mechanism, achieving two critical objectives: (1) maintaining stable context entropy across varying input lengths through mean and variance constraints, thereby enhancing performance in long-context scenarios; and (2) adaptively guiding the LLM to focus on crucial segments, improving its ability to effectively utilize knowledge. Unlike existing attention reweighting approaches that either uniformly reduce context entropy (Su 2021) (hindering important information extraction) or ignore attention entropy (Yan et al. 2024) (leading to unstable performance across varying context lengths). We further introduce novel adaptations for zero-shot and parameter-efficient fine-tuning scenarios: a prompt-induced importance scoring mechanism with parallel scoring-generation architecture, and an efficient vector fine-tuning strategy. These components collectively enhance BEE-RAG’s accuracy in question-answering tasks

Efficiency. We further analyze the efficiency of our EE-RAG framework. The parallel mechanism handles β calculation and answer generation simultaneously. Due to the quadratic time complexity in the transformer mechanism, our design effectively controls computational overhead, achieving performance comparable to vanilla RAG with improved scalability for longer context. These efficiency gains result from our optimized attention complexity handling.

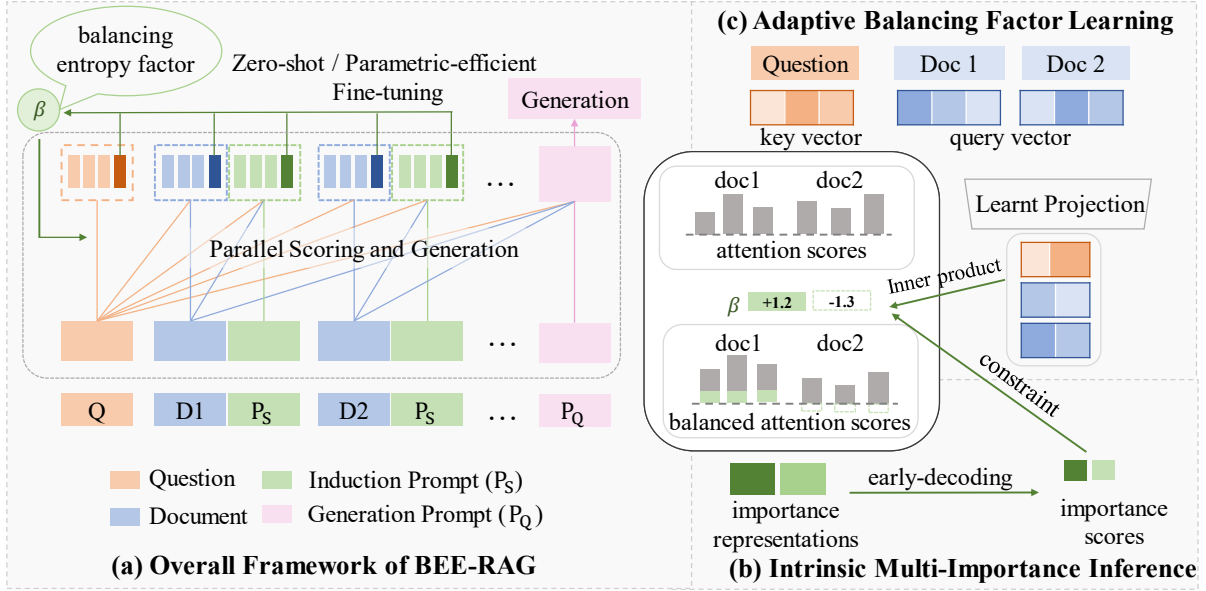


Figure 3: The overview of the proposed BEE-RAG. The left side shows the overall architecture design, while the right side illustrates the zero-shot strategy *intrinsic multi-importance inference* and the parametric-efficient fine-tuning strategy *adaptive balancing factor learning*.

Experiments

In this section, we detail the experimental setup and then report the evaluation results.

Experimental Settings

Datasets We conduct experiments on both single-hop and multi-hop question-answering (QA) tasks. Specifically, we evaluate our approach on the Natural Questions (NQ) (Kwiatkowski et al. 2019), TriviaQA (Joshi et al. 2017), HotpotQA (Yang et al. 2018), and 2WikiMulti-hopQA (2WikiQA) (Ho et al. 2020) datasets.

Metrics We use Exact Match (EM) (Lee, Chang, and Toutanova 2019) and F1 (Karpukhin et al. 2020) to assess the QA accuracy of retrieval-augmented LLMs, which are commonly adopted in QA evaluation. EM checks if predicted answers exactly match the ground truth, while F1 measures their overlap in precision and recall.

Implementation Details We select LLaMA-3-8B (Dubey et al. 2024) and Qwen-2.5-7B (Yang et al. 2024) as base LLM. For the zero-shot setting (Section), we derive the theoretical values for μ and σ^2 based on Equation 9, and perform a search around these values to determine their optimal configuration. For the adaptive balancing entropy factor learning, we set $d_r = 8$. We conduct the experiments on 8 NVIDIA A100 40GB GPUs, with a learning rate of $5e-4$, a cosine learning rate scheduler, and a batch size of 16. To demonstrate the robustness of our training approach, we train the LLMs for no more than 500 steps solely on the NQ dataset and subsequently validate its performance across multiple datasets.

Baselines We introduce two types of baselines for comparison with our approach, including zero-shot methods and lightweight fine-tuning methods.

- *Zero-shot methods.* We select several prompt-based strategies that enhance generation performance in RAG scenarios for comparison, including prompt-based methods (Chain-of-Thought (Wei et al. 2022), Chain-of-Note (Yu et al. 2023), Self-Critic (Asai et al. 2023)), architecture-driven methods (PCW (Ratner et al. 2023), Position-Engineering (He et al. 2024)), and attention reweighting method (Multiply-Attention (Chiang and Cholak 2022)).

- *Parametric-efficient fine-tuning methods:* We choose two commonly used PEFT approaches in RAG, including LoRA (Hu et al. 2022) and Prefix-Tuning (Li and Liang 2021) for comparison with our parameter-efficient training strategy.

Main Results

Table 1 presents the results of our approach and the baselines on four RAG tasks of open-domain QA. It can be observed that:

First, our proposed method outperforms all other baselines in terms of average QA performance. Whether in the zero-shot setting or with lightweight fine-tuning, BEE-RAG consistently achieves the best results. This highlights the effectiveness of our key motivation: the reduction of context entropy and dynamic attention allocation on documents, which proves to be applicable to both single-hop and multi-hop datasets.

Second, our method demonstrates its ability to handle more challenging datasets, assisting LLMs in effectively addressing complex RAG tasks. On relatively more diffi-

Settings	NQ		TriviaQA		HotpotQA		2WikiQA		Average	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
<i>Qwen-2.5-7B</i>										
Vanilla RAG	26.84	42.03	43.05	53.41	24.54	35.16	20.16	27.16	28.65	39.44
Chain-of-Thought	26.86	40.20	39.59	51.78	26.20	35.88	22.93	26.58	28.89	38.61
Chain-of-Note	26.50	41.75	41.20	52.03	26.74	35.95	19.93	26.58	28.59	39.08
PCW	23.87	38.21	40.88	51.50	24.72	34.63	23.11	27.19	28.15	37.88
Position-Engineering	25.76	38.61	41.38	52.30	27.77	36.56	22.69	26.96	29.40	38.60
Self-Critic	24.27	35.36	40.74	52.04	25.11	35.35	22.00	27.51	28.03	37.57
Multiply-Attention	26.18	41.43	42.92	53.62	27.13	34.94	18.87	24.02	28.78	38.50
Zero-BEE (ours)	<u>29.03</u>	<u>43.15</u>	<u>44.32</u>	<u>55.93</u>	<u>27.86</u>	<u>37.15</u>	<u>24.47</u>	<u>29.61</u>	<u>31.42</u> [†]	<u>41.46</u> [†]
LoRA (2.06M)	37.62	47.93	47.98	56.32	28.44	38.92	20.76	26.30	33.70	42.37
Prefix-Tuning (2.07M)	36.75	46.33	48.21	56.32	31.19	40.15	23.69	29.19	34.96	43.00
Light-BEE (ours, 2.06M)	38.71	50.11	49.39	56.73	35.95	44.81	28.32	31.73	38.09 [†]	45.84 [†]
<i>LLaMA-3-8B</i>										
Vanilla RAG	27.13	37.77	41.17	55.46	23.07	34.22	14.20	23.74	26.39	37.80
Chain-of-Thought	29.10	38.60	45.31	56.96	22.74	30.23	14.50	25.58	27.91	37.84
Chain-of-Note	28.46	37.83	41.49	51.22	21.34	28.99	19.87	29.50	27.79	36.89
PCW	25.71	35.47	38.72	48.99	17.14	25.92	17.88	26.65	24.86	34.26
Position-Engineering	25.98	35.72	41.01	54.21	23.24	30.62	16.08	26.61	26.58	36.79
Self-Critic	20.86	27.06	42.61	51.06	23.77	31.21	15.50	25.94	25.69	33.82
Multiply-Attention	27.08	36.33	43.40	55.05	21.27	28.24	11.87	23.74	25.91	35.84
Zero-BEE (ours)	<u>30.50</u>	<u>38.96</u>	<u>45.67</u>	<u>57.04</u>	<u>23.84</u>	<u>34.29</u>	<u>20.67</u>	<u>30.56</u>	<u>30.17</u> [†]	<u>40.21</u> [†]
LoRA (2.62M)	35.39	48.68	53.65	58.55	33.15	42.15	18.91	28.67	35.27	44.51
Prefix-Tuning (2.63M)	36.39	48.25	48.75	56.48	32.30	42.50	16.33	25.13	33.44	43.09
Light-BEE (ours, 2.62M)	38.42	49.23	53.96	58.67	38.06	44.72	24.70	31.89	38.79 [†]	46.13 [†]

Table 1: Main results of BEE-RAG on four RAG tasks with different LLM backbones for both zero-shot inference and lightweight fine-tuning settings. Zero-BEE denotes the zero-shot inference version of BEE-RAG, while Light-BEE denotes the parameter-efficient (lightweight) fine-tuning version. All results are averaged over 8 runs; † indicates statistical significance with $p < 0.05$.

cult multi-hop datasets, BEE-RAG shows significant performance improvements. Notably, on the latest 2WikiMultihopQA dataset, our approach achieves an approximately 5% improvement over vanilla RAG, further highlighting the effectiveness of our approach.

Third, our parametric-efficient fine-tuning method demonstrates better generalization performance with few parameters, emphasizing the effectiveness of the proposed adaptive balancing entropy factor learning approach. The result shows that whether on trained or zero-shot datasets, our method outperforms the others. Notably, the performance of the other two methods may degrade when transferred to multi-hop tasks, revealing their limitations in handling challenging RAG tasks.

Ablation Study

In this section, we conduct an ablation study to validate the effectiveness of the key strategies in BEE-RAG on both zero-shot and fine-tuning scenarios. As shown in Table 2, the left part of the table corresponds to the variants in the zero-shot setting, while the right part focuses on the variants with lightweight fine-tuning.

First, we observe that setting μ and σ to zero (*w/o μ* , *w/o σ^2*) results in performance degradation, demonstrating that

Zero-shot	EM	PEFT	EM
Zero-BEE	31.42	Light-BEE	38.09
w/o IMI	29.89	w/o IMI	36.58
w/o μ	27.47	w/o Residual	23.65
w/o σ^2	26.92	w/o Init constraint	3.23
w/o β	20.35	w/ FFN	36.41
w/ reranker	28.87	w/ reranker	35.82

Table 2: Ablation study on our BEE-RAG.

adjusting the mean (μ) and variance (σ^2) of the balancing entropy factor helps improve multi-document RAG performance. This observation also aligns well with our formulation in Equation (9).

Second, we validate the effectiveness of the specific strategies introduced in our lightweight training approach. We demonstrate the positive effect of our chunk-level adaptive balancing entropy factor learning formula by removing the residual and FFN module in our approach (*w/o Residual*, *w/ FFN*). Furthermore, we prove that adding constraints in the initialization of the balance factor (*w/o Init constraint*) prevents gradient explosion during training.

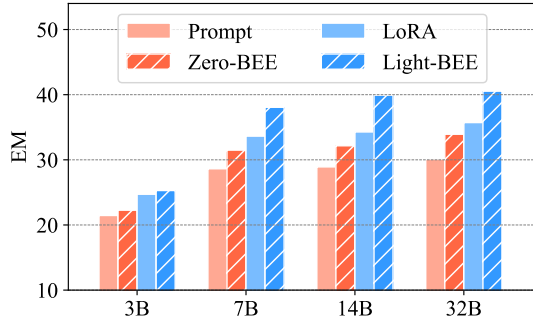


Figure 4: Effect of LLM capabilities with various Scales. Mean metrics are reported over four datasets.

Third, we observe that removing parallel multi-importance inference (*w/o IMI*) leads to a performance drop, which can be attributed to the interference from other chunks (documents) when the LLM generates balance factors required independent document information, causing a performance decline. In addition, we explore the impact of adding a reranker module to assess the relevance between the queries and documents as a substitute for the balancing entropy factor in BEE-RAG (*w/ reranker*). We observe a performance decline, proving our method avoids extra modules and costs while outperforming in both scenarios.

In-depth Analysis

In this section, we further present an analysis from two aspects: the scaling of LLMs and the quality of the retrieved documents.

Effect of Model Scaling In this part, we first explore the impact of different model scales on experimental performance. We test the Qwen-2.5 family LLMs, ranging from 3B, 7B, 14B to 32B, on their average generation results across four datasets. The results are shown in Figure 4. It is clear that our proposed BEE-RAG consistently improves RAG accuracy across different model sizes. This further demonstrates that our method enables LLMs to better handle generation tasks involving noisy contexts and exhibits strong generalization capabilities.

Effect of Retrieved Documents We consider both context length and retrieval quality to evaluate the effect of retrieved documents, showing the comprehensive priority of the proposed BEE-RAG method.

Context Length. We first examine the impact of the number of the input document. Figure 5 shows the average performance of our BEE-RAG and the baselines varying the document number. It is clear that our method overall outperforms the baselines. Notably, as the context length increases, the performance gap between BEE-RAG and the baselines becomes significantly larger. This highlights that our strategy is more effective in adapting to larger amounts of contextual information, allowing LLMs to efficiently extract important information through global context entropy reduction and dynamic document-level attention allocation.

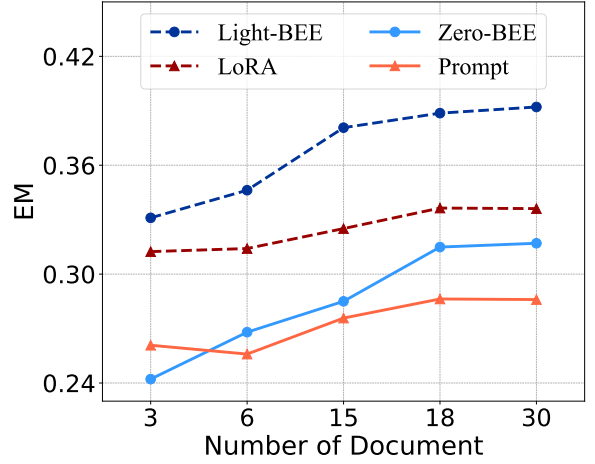


Figure 5: Effect of context length with various document numbers. Mean metrics are reported over four datasets.

Retriever	Prompt	Zero-BEE	LoRA	Light-BEE
BM25	23.82	26.98	28.22	35.03
Contriever	28.65	31.42	33.70	38.09
E5-Base	29.36	32.59	34.92	38.55
E5-large	29.77	34.02	34.02	38.84
BGE	30.51	33.91	36.72	39.05

Table 3: Effect of retrieval quality with various retrievers. Mean metrics are reported over four datasets.

Retrieval Quality. Moreover, we explore the impact of various retrievers with different retrieval document qualities for the input of BEE-RAG. Table 3 presents the QA performance of our approach and the baselines using various retrievers across four datasets, including BM25, Contriever, E5-base, E5-large, and BGE. From the table, it is evident that our method achieves the best performance across all retriever settings. Notably, our approach excels when retrieval quality is lower. This highlights that BEE-RAG is better at focusing on the critical parts of the retrieval results, thereby improving the overall accuracy of the generations.

Related Work

Recent efforts have aimed to improve RAG performance under long contexts (Laban et al. 2024). PCW (Ratner et al. 2023) restricts attention within fixed-size windows to reduce attention dilution. CEPE (Yen, Gao, and Chen 2024) uses a lightweight encoder to process inputs chunk by chunk, enabling better use of long contexts via cross-attention. RAFT (Zhang et al. 2024b) and Chain-of-Notes (Yu et al. 2023) extend the context capacity of models through instruction tuning on long-input data. However, these methods either yield limited performance gains or incur high training costs. Some works explore the effect of attention distribution (Wang et al. 2023). Su (2021) adjust attention scaling by context length to match short-context behavior, but this fails to highlight key segments and may miss important information.

Our BEE-RAG introduces an attention reallocation mechanism. It reduces attention entropy gap with short contexts and guides focus to key segments. This improves long-context RAG performance.

Conclusion

In this study, we identify and address the fundamental limitation of unconstrained context entropy in RAG systems through the lens of information entropy. We proposed BEE-RAG for the adaptability of RAG systems to various context lengths. By introducing balanced context entropy, we demonstrate that enforcing entropy invariance enables superior RAG performance with various context lengths without sacrificing computational efficiency. In order to extend applicability to a broader range of scenarios, we propose multi-importance inference for zero-shot reasoning, and also propose adaptive balancing factor learning for parameter-efficient fine-tuning. Empirical examination on four real-world knowledge-intensive tasks shows consistent performance gains of BEE-RAG.

References

- Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H. 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *The Twelfth International Conference on Learning Representations*.
- Bolthausen, E.; and Wüthrich, M. V. 2013. Bernoulli’s law of large numbers. *ASTIN Bulletin: The Journal of the IAA*, 43(2): 73–79.
- Cammarata, N.; Carter, S.; Goh, G.; Olah, C.; Petrov, M.; Schubert, L.; Voss, C.; Egan, B.; and Lim, S. K. 2020. Thread: Circuits. *Distill*. <https://distill.pub/2020/circuits>.
- Chiang, D.; and Cholak, P. 2022. Overcoming a Theoretical Limitation of Self-Attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7654–7664.
- Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; and Wang, H. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
- He, Z.; Jiang, H.; Wang, Z.; Yang, Y.; Qiu, L.; and Qiu, L. 2024. Position Engineering: Boosting Large Language Models through Positional Information Manipulation. *arXiv preprint arXiv:2404.11216*.
- Ho, X.; Nguyen, A.-K. D.; Sugawara, S.; and Aizawa, A. 2020. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, 6609–6625.
- Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.
- Jeong, S.; Baek, J.; Cho, S.; Hwang, S. J.; and Park, J. C. 2024. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 7029–7043.
- Joshi, M.; Choi, E.; Weld, D.; and Zettlemoyer, L. 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)*, 1601–1611.
- Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781.
- 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.
- Laban, P.; Fabbri, A. R.; Xiong, C.; and Wu, C.-S. 2024. Summary of a haystack: A challenge to long-context llms and rag systems. *arXiv preprint arXiv:2407.01370*.
- Lee, K.; Chang, M.-W.; and Toutanova, K. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6086–6096.
- 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.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. 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)*, 4582–4597.
- Lin, X. V.; Chen, X.; Chen, M.; Shi, W.; Lomeli, M.; James, R.; Rodriguez, P.; Kahn, J.; Szilvasy, G.; Lewis, M.; et al. 2024. RA-DIT: Retrieval-Augmented Dual Instruction Tuning. In *The Twelfth International Conference on Learning Representations*.
- Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.
- Pradeep, R.; Sharifmoghaddam, S.; and Lin, J. 2023. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *arXiv preprint arXiv:2309.15088*.
- Ratner, N.; Levine, Y.; Belinkov, Y.; Ram, O.; Magar, I.; Abend, O.; Karpas, E. D.; Shashua, A.; Leyton-Brown, K.; and Shoham, Y. 2023. Parallel Context Windows for Large

- Language Models. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.
- Ren, R.; Wang, Y.; Qu, Y.; Zhao, W. X.; Liu, J.; Tian, H.; Wu, H.; Wen, J.-R.; and Wang, H. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*.
- Saxe, A. M.; McClelland, J. L.; and Ganguli, S. 2013. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*.
- Su, J. 2021. On the Scale Operation in Attention from the Perspective of Entropy Invariance. <https://distill.pub/2020/circuits>.
- Wang, L.; Li, L.; Dai, D.; Chen, D.; Zhou, H.; Meng, F.; Zhou, J.; and Sun, X. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. *arXiv preprint arXiv:2305.14160*.
- Wang, Y.; Ren, R.; Li, J.; Zhao, X.; Liu, J.; and Wen, J.-R. 2024. REAR: A Relevance-Aware Retrieval-Augmented Framework for Open-Domain Question Answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 5613–5626.
- Waswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.
- Wu, Z.; Arora, A.; Wang, Z.; Geiger, A.; Jurafsky, D.; Manning, C. D.; and Potts, C. 2025. Reft: Representation fine-tuning for language models. *Advances in Neural Information Processing Systems*, 37: 63908–63962.
- Yan, S.-Q.; Gu, J.-C.; Zhu, Y.; and Ling, Z.-H. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.
- Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.
- Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2369–2380.
- Yen, H.; Gao, T.; and Chen, D. 2024. Long-Context Language Modeling with Parallel Context Encoding. In Ku, L.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, 2588–2610. Association for Computational Linguistics.
- Yu, W.; Zhang, H.; Pan, X.; Ma, K.; Wang, H.; and Yu, D. 2023. Chain-of-note: Enhancing robustness in retrieval-augmented language models. *arXiv preprint arXiv:2311.09210*.
- Zhang, P.; Liu, Z.; Xiao, S.; Shao, N.; Ye, Q.; and Dou, Z. 2024a. Soaring from 4k to 400k: Extending llm’s context with activation beacon. *arXiv preprint arXiv:2401.03462*, 2(3): 5.
- Zhang, T.; Patil, S. G.; Jain, N.; Shen, S.; Zaharia, M.; Stoica, I.; and Gonzalez, J. E. 2024b. Raft: Adapting language model to domain specific rag. *arXiv preprint arXiv:2403.10131*.
- Zhang, Z.; Wang, Y.; Huang, X.; Fang, T.; Zhang, H.; Deng, C.; Li, S.; and Yu, D. 2024c. Attention Entropy is a Key Factor: An Analysis of Parallel Context Encoding with Full-attention-based Pre-trained Language Models. *arXiv preprint arXiv:2412.16545*.
- Zhuang, S.; Zhuang, H.; Koopman, B.; and Zuccon, G. 2024. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 38–47.