LFD: Layer Fused Decoding to Exploit External Knowledge in Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Retrieval-augmented generation (RAG) incorporates external knowledge into large language models (LLMs), improving their adaptability to downstream tasks and enabling information updates. Surprisingly, recent empirical evidence demonstrates that injecting noise into retrieved relevant documents paradoxically facilitates exploitation of external knowledge and improves generation quality. Although counterintuitive and challenging to apply in practice, this phenomenon enables granular control and rigorous analysis of how LLMs integrate external knowledge. Therefore, in this paper, we intervene on noise injection and establish a layer-specific functional demarcation within the LLM: shallow layers specialize in local context modeling, intermediate layers focus on integrating long-range external factual knowledge, and deeper layers primarily rely on parametric internal knowledge. Building on this insight, we propose Layer Fused Decoding (LFD), a simple decoding strategy that directly combines representations from an intermediate layer with final-layer decoding outputs to fully exploit the external factual knowledge. To identify the optimal intermediate layer, we introduce an internal knowledge score (IKS) criterion that selects the layer with the lowest IKS value in the latter half of layers. Experimental results across multiple benchmarks demonstrate that LFD helps RAG systems more effectively surface retrieved context knowledge with minimal cost. The code for replication is available at https://anonymous.4open.science/r/LFD/.

1 Introduction

Retrieval-Augmented Generation (RAG) empowers large language models (LLMs) by dynamically integrating external knowledge during inference, enabling precise adaptation to knowledge-intensive tasks and rapidly evolving domains (Borgeaud et al., 2022; Guu et al., 2020; Lewis et al., 2020a). As a cornerstone of context-aware generation, RAG has been widely deployed in real-world applications, including recommendation systems (Fan et al., 2024; Di Palma, 2023; Ji et al., 2024) and search engines (Salemi and Zamani, 2024; Xiong et al., 2024). The broad applicability has spurred extensive optimization efforts on dynamic knowledge integration, including reranking strategies (Yu et al., 2024; Dong et al., 2024) to prioritize relevance, adaptive retrieval mechanisms (Asai et al., 2024; Jiang et al., 2023c) to minimize redundancy, and graph-based architectures (He et al., 2024; Edge et al., 2024; Jiang et al., 2023b) to model inter-document semantic relationships.

Despite these advances, LLMs might underutilize accurate external contexts, disproportionately favoring internal parametric knowledge during generation (Sun et al., 2025b; Niu et al., 2024). This overreliance risks propagating outdated information or hallucinations, undermining the trustworthiness of RAG systems. Surprisingly, recent studies reveal a paradoxical phenomenon: **injecting noise—random documents or tokens—to retrieved contexts that already contain answer-relevant snippets can improve the generation accuracy** (Cuconasu et al., 2024; Sun et al., 2025a). While this noise-injection approach is simple and effective, its underlying influence on LLM remains unclear. Furthermore, long contexts containing noise documents create computational overhead. Therefore, it is important to design more principled strategies that can achieve similar benefits without incurring excessive cost.

This phenomenon enables more granular control and rigorous analysis of how LLMs integrate external knowledge. To investigate the underlying mechanisms, we study layer-wise external knowledge exploitation by measuring the divergence of ablating answer-determining context, i.e., the specific text segment within retrieved documents that directly supports the correct answer to a query. By intervening on injecting noise and measuring its

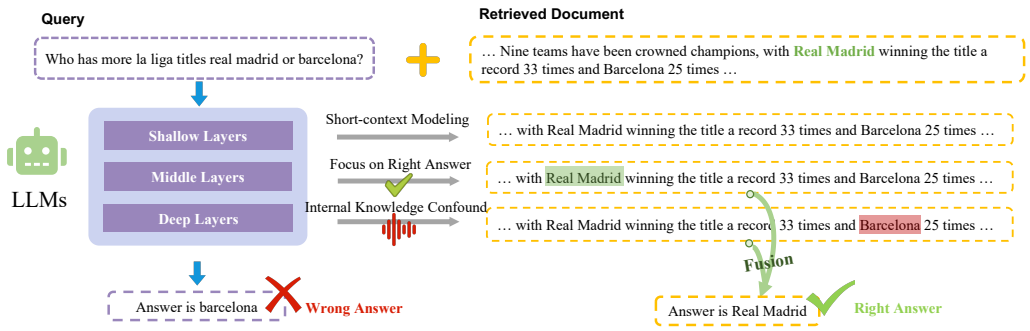


Figure 1: Illustration for layer-wise behavior in LLMs for RAG. Given a query and retrieved documents with the correct answer ("Real Madrid"), shallow layers capture local context, middle layers focus on answer-relevant content, while deep layers may over-rely on internal knowledge and hallucinate (e.g., "Barcelona"). Our proposal, LFD fuses middle-layer signals into the final output to preserve external knowledge and improve accuracy.

impact on divergence patterns, we identify the relative importance of different layers in exploiting external knowledge. Empirically, we find that noise amplifies the contribution of answer-determining context in middle layers, highlighting their critical role in integrating long-range external information. To further support this observation, we compare attention distributions across heads and layers with versus without answer-determining context. The analysis shows that attention differences peak in middle layers but decline in later layers, signaling a transition from external knowledge reliance to internal parametric knowledge utilization. Following these observations, we propose a functional categorization of LLM layers: (1) shallow layers for short-context modeling, (2) intermediate layers for external knowledge integration, and (3) deeper layers for internal knowledge transformation. Therefore, when retrieved context already contains the correct answer, excessive dependence on internal knowledge in later layers introduces confounding effects, reducing generation accuracy, as visualized in Figure 1 (left).

Based on this insight, we propose a simple decoding method, Layer Fusing Decoding (LFD), which enhances access to external factual knowledge without introducing additional noise overhead. The core idea is illustrated in the right panel of Figure 1. LFD fuses representations from the long-term context retrieval layer, where external knowledge is most effectively integrated, directly into the final decoding layer to maximize factual grounding. To identify the appropriate layer for fusion, we track the model's reliance on internal knowledge by measuring changes in hidden states across transformer feed-forward network (FFN) layers, where model knowledge is primarily stored (Geva et al., 2021; Dai et al., 2022). Specifically, we select the layer exhibiting minimal internal knowledge influence from the latter half of the model's layers. This criterion ensures that LFD captures the externally grounded signal before it is overridden by parametric knowledge in later layers. Importantly, LFD operates at inference time, requiring no post-hoc fine-tuning or architectural modifications, making it easily integrable into existing LLM pipelines. Finally, extensive empirical validation across diverse model architectures and datasets demonstrates that LFD delivers competitive performance relative to noise-based approaches, while incurring significantly lower computational overhead.

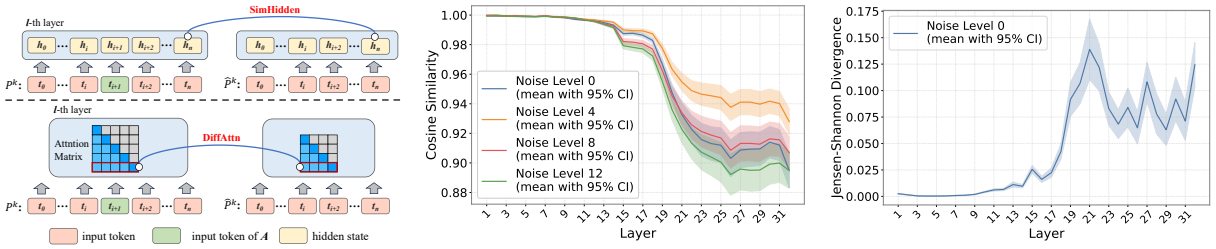
2 Preliminary

2.1 Formulation of RAG

In RAG systems, the generator $\mathcal{G}$ (typically a LLM) is expected to produce accurate and well-grounded responses based on retrieved documents $D = \{d_1, d_2, ..., d_\lambda\}$. To quantify this capability, we define $A = \{a_1, ..., a_\lambda\}$ as the set of key information extracted from D that is necessary for generating an accurate answer. The performance of the RAG system can be evaluated by measuring the inclusion rate of A in its output response r, which reflects the model's ability to fully utilize valuable documents. To produce the final response, the system first encodes the query q and documents D into a structured prompt through an instruction template $\mathcal{T}$, which instantiates the prompt $P = \mathcal{T}(q, D)$, then the generator processes this prompt to produce the final response r. To optimize the generation process, the generator $\mathcal{G}$ aims to ensure that all answers contained within A are included in the generator's output. The accurate answer generated by $\mathcal{G}$ can be formalized as:

$$r = \mathcal{G}(q, D), \text{ s.t. } \forall a_i \in A, \mathcal{I}(r, a_i) = \text{True}, \quad (1)$$

where $\mathcal{I}(r, a_i) = \text{True}$ means the answer a_i is included in r.



(a) Metric computation (b) SimHidden (Smaller is better). (c) DiffAttn (Larger is better).

Figure 2: (a) Calculation procedure for SimHidden and DiffAttn metrics; (b) Average SimHidden scores (with 95% confidence intervals) across layers under varying noise levels (0, 4, 8, 12); (c) Average DiffAttn scores (with 95% confidence intervals) across layers when noise level = 0.

2.2 External Knowledge Intervention in RAG

The counterintuitive effectiveness of noise in RAG systems motivates a deeper investigation into layer-wise behavior in LLMs. To this end, we conduct an empirical study that contrasts layer-wise representation dynamics under two controlled interventions: (1) ablation of the answer-determining context, and (2) injection of varying levels of noise into the retrieved documents. This differential analysis reveals how noise modulates the model's internal information flow, amplifying the influence of external knowledge in middle layers while mitigating the model's tendency to over-rely on internal parametric memory. Our findings highlight key transformation layers where noise injection helps reduce context-dependent fragility, providing a foundation for our proposed decoding strategy.

Experimental Setup We simulate noisy level by adding k irrelevant Wikipedia documents $N_k = \{n_1, \ldots, n_k\}$ to each prompt input (Cuconasu et al., 2024). To analyze how external knowledge flows in LLMs, we generate a modified document set $\hat{D}$, which deletes key information A from the original documents D. Without loss of generality, we assume the retrieval set D contains a single document, i.e., $D = \{d_1\}$. By varying noise levels k, we analyze how external knowledge impacts different layers of the model by comparing two prompts: the **original prompt** $P^k = \mathcal{T}(q, D, N_k)$ (containing key information A) and the **modified prompt** $\hat{P}^k = \mathcal{T}(q, \hat{D}, N_k)$, shown in Figure 2(a). This section focuses on Llama2-7B and the NQ dataset. Additional results are in Appendix A.

Quantify External Knowledge's Influence To measure how external knowledge influences LLMs, we compare intermediate representations before and after removing critical information A. For each layer $l \in \{1, \ldots, L\}$ in the model, we calculate the cosine similarity between the feed-forward network (FFN) outputs of the original input P^k and its modified version $\hat{P}^k$, focusing on the final prompt token as

$$\text{SimHidden}_l(P^k, \hat{P}^k) = \frac{\boldsymbol{h}_l(P^k) \cdot \boldsymbol{h}_l(\hat{P}^k)}{\|\boldsymbol{h}_l(P^k)\| \cdot \|\boldsymbol{h}_l(\hat{P}^k)\|}, \quad (2)$$

where $\boldsymbol{h}_l(P)$ denotes the intermediate representation of layer l under prompt P. This metric reveals how significantly removing A disrupts the model's contextual processing at each layer. A higher score indicates the model's understanding remains consistent even after removing A, while a lower score suggests removing A plays a critical role in shaping the layer's output. By analyzing this metric across varying levels of noise injection (k), we can assess how different noise perturbation intensities affect the model's reliance on external knowledge, providing insights into how contextual information is integrated across layers.

Additionally, since the divergence of $\boldsymbol{h}_l(P^k)$ and $\boldsymbol{h}_l(\hat{P}^k)$ tends to accumulate in deeper layers, we further analyze attention patterns before and after the removal of the answer-determining context A. For each transformer layer l, we compute the average Jensen-Shannon Divergence (JSD) (Lin, 2002) across all attention heads to quantify distributional shifts in attention:

$$\text{DiffAttn}_l(P^k, \hat{P}^k) = \frac{1}{M} \sum_{m=1}^{M} \text{JSD}(\boldsymbol{\alpha}_{l,m}(P^k) \,\|\, \hat{\boldsymbol{\alpha}}_{l,m}(\hat{P}^k)), \quad (3)$$

where $\boldsymbol{\alpha}_{l,m}(P^k)$ denotes the softmax-normalized attention distribution of head m in layer l at the final token position for prompt P^k. The variant $\hat{\boldsymbol{\alpha}}_{l,m}(\hat{P}^k)$ is computed by filling attention scores corresponding to answer-determining context A with $-\infty$ before softmax normalization, thereby preserving the relative distribution over remaining tokens in modified prompt $\hat{P}^k$. Larger divergence scores indicate that the external knowledge A significantly influences the model's attention. Note that we concentrate on intervention of noise-free models (where $k = 0$), we compare the original prompts P^0 with their intentionally altered counterparts $\hat{P}^0$.

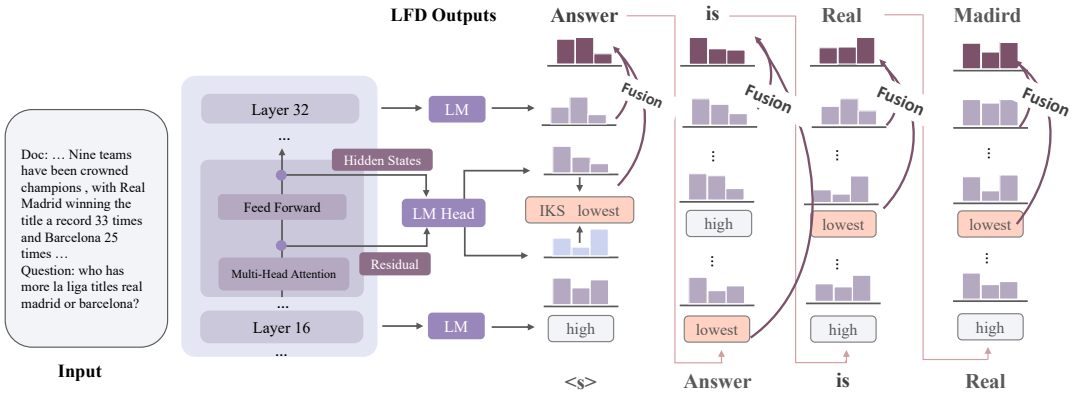


Figure 3: The proposed LFD includes two key components: (1) A dynamic layer selection method using IKS to pinpoint the most impactful layers for integrating external retrieval knowledge. (2) A knowledge fusion mechanism that merges external information with the model's predictions after adaptive filtering to ensure alignment with the model's reasoning.

Analysis and Conclusion The quantified impacts of external knowledge, as depicted in Figure 2(b-c), lead to the following observations and conclusions: **(1) The early layers (1-14) primarily perform short-context modeling**, which means they focus more on capturing local token relationships rather than integrating global contextual information. This manifests through two key observations: the hidden state similarity remains relativly high across all noise conditions and the attention divergence also stay constantly low compared to other layers. **(2) The middle layers (15-26) demonstrate long-term context retrieval capabilities**, as evidenced by two complementary patterns: a progressive decline in $\text{SimHidden}_l(P^k, \hat{P}^k)$ (Figure 2 (b)) and a corresponding increase in DiffAttn_l (Figure 2 (c)). This dual evidence indicates these layers' heightened sensitivity to the removal of A and their capacity for comprehensive global context integration. Meanwhile, an increasing noise level, especially when $k \geq 8$, leads to greater discrepancies in hidden state similarity, suggesting that a certain amount of noise can enhance the model's focus on A, thereby improving answer accuracy. **(3) The deeper layers (27-32) exhibit characteristics of parametric knowledge utilization.** As we can observe, hidden state similarity does not continue to decrease as the model depth increases, instead, it shows a rebound, with SimHidden increasing by a maximum of 0.1. Meanwhile, attention divergence, after peaking at layer 21, also exhibits a moderate decline, with DiffAttn decreasing by a maximum of approximately 0.06. This indicates that the role of internal knowledge may be enhanced, as the model may focus more on processing the already captured contextual information rather than continuing to attend to external knowledge. These observations motivate us to design our own methods to better leverage external knowledge with internal representations, thereby improving the model's performance in RAG systems.

3 Layer Fused Decoding

This section present LFD, a framework designed to improve how external knowledge is integrated into model predictions while retaining accuracy. Our approach has two core components: (1) A **dynamic external knowledge layer identification strategy**, which automatically selects the most impactful layer for integrating retrieved context. This selection is guided by Internal Knowledge Scores (IKS), which measure how strongly each layer reflects the model's parametric knowledge. (2) An **external knowledge fused decoding** mechanism, which merges external knowledge representations with the model's final output. An adaptive filtering step precedes fusion to ensure the incorporated information complements the model's reasoning. Figure 3 illustrates the complete workflow of our approach.

3.1 Dynamic External Knowledge Layer Identification

Effective integration of retrieval-based knowledge requires identifying layers that are responsive to external context while minimally influenced by internal parametric knowledge. To this end, we quantify the influence of internal knowledge across layers and select candidate layers based on the external information integration.

Internal Knowledge Score Recent advances in transformer interpretability reveal that FFN layers function as specialized knowledge repositories in LLMs (Geva et al., 2021; Dai et al., 2022). To quantify how different layers utilize this inter-

nal knowledge, we design the Internal Knowledge Score (IKS), a metric that captures layer-specific knowledge transformations. Given an input prompt P, let $\boldsymbol{h}_l^{\text{in}}(P) \in \mathbb{R}^d$ and $\boldsymbol{h}_l^{\text{out}}(P) \in \mathbb{R}^d$ denote the input and output activations of the l-th FFN layer. We project these vectors into the vocabulary space via LogitLens (Belrose et al., 2023) of LLMs, parameterized by $W_{\text{LM}} \in \mathbb{R}^{d\times|V|}$, as follows:

$$\begin{aligned} \boldsymbol{p}_l^{\text{in}} &= \text{softmax}(W_{\text{LM}}\boldsymbol{h}_l^{\text{in}}(P)), \\ \boldsymbol{p}_l^{\text{out}} &= \text{softmax}(W_{\text{LM}}\boldsymbol{h}_l^{\text{out}}(P)). \end{aligned} \tag{4}$$

The IKS for layer l is defined as the JSD divergence between these distributions as

$$\text{IKS}_l(P) = \text{JSD}(\boldsymbol{p}_l^{\text{in}} \parallel \boldsymbol{p}_l^{\text{out}}). \tag{5}$$

This divergence quantifies the **parametric knowledge impact** of FFN layers, where higher IKS indicates greater transformation in the residual stream and stronger reliance on internal knowledge.

External Knowledge Layer Selection To identify the optimal layer for leveraging context-derived factual knowledge during inference, we propose two principled criteria for layer selection: **(1) Integration of Late-Stage Layers**: We integrate the latter half of the LLM's layers for final decoding. This choice is motivated by the observation that early layers predominantly focus on short-context modeling, with limited capacity to capture contextual dependencies. The integration of external knowledge becomes progressively stronger in middle-to-late layers (as analyzed in Section 2.2). **(2) Lowest IKS Layer Selection**: Within the subset of late-stage layers, we select the layer exhibiting the lowest IKS score. This layer strikes a balance by retaining sufficient external contextual signals while substantially mitigating distortion of the model's inherent knowledge representations. By selecting these layers, we maximize the exploitation of retrieval context before the dominance of internal parametric knowledge obscures external signals. Empirical validation further confirms the efficacy of this strategy (detailed in Section 4.3).

3.2 External Knowledge Fused Decoding

We propose an intervention-aware fusion framework that dynamically integrates intermediate representations from layer i (identified via IKS scoring) with the final layer's predictions. To establish distributional coherence between these complementary knowledge sources, we first compute normalized log-probabilities through $\log\text{softmax}$ transformation:

$$\begin{aligned} \tilde{\boldsymbol{p}}_i^{\text{out}} &= \log\text{softmax}(W_{\text{LM}}\boldsymbol{h}_i^{\text{out}}(P)), \\ \tilde{\boldsymbol{p}}_L^{\text{out}} &= \log\text{softmax}(W_{\text{LM}}\boldsymbol{h}_L^{\text{final}}(P)) \end{aligned} \tag{6}$$

where $W_{\text{LM}}\boldsymbol{h}_i^{\text{out}}(P)$ and $W_{\text{LM}}\boldsymbol{h}_L^{\text{out}}(P)$ denote the raw logits from the intervention layer and final layer respectively. To mitigate noise amplification from early layer predictions while preserving critical external knowledge signals, we implement a dynamic gating mechanism inspired by (Li et al., 2022):

$$\tilde{\boldsymbol{f}}_i(t) = \begin{cases} \tilde{\boldsymbol{p}}_i^{\text{out}}(t) + \tilde{\boldsymbol{p}}_L^{\text{out}}(t), & \text{if } \tilde{\boldsymbol{p}}_i^{\text{out}}(t) \geq \Gamma \\ -\infty, & \text{otherwise,} \end{cases} \tag{7}$$

Within this gating mechanism, the threshold is formulated as $\Gamma = \min\{\tau \cdot \max(\tilde{\boldsymbol{p}}_L^{\text{out}}), \text{max-s}(\tilde{\boldsymbol{p}}_L^{\text{out}})\}$. Here, $\max(\cdot)$ and $\text{max-s}(\cdot)$ denote the maximum and s-th maximum values of the final output layer logits $\tilde{\boldsymbol{p}}_L^{\text{out}}$, respectively, and $\tau \in [0, 1]$ is a hyperparameter that controls the truncation aggressiveness to filter out low-probability tokens. The final decoding distribution, derived via normalized fusion $\boldsymbol{f}_i = \text{softmax}(\tilde{\boldsymbol{f}}_i)$, enables synergistic knowledge transfer between layers, preserving the final layer's discriminative capacity to balance the integration of external knowledge with the model's inherent confidence.

4 Experiments

4.1 Experiment Setup

Datasets and Metircs Building on related works (Cuconasu et al., 2024; Wu et al., 2024; Jeong et al., 2024), our evaluation encompasses both single-hop and multi-hop QA tasks. Single-hop performance is measured using the widely-used Natural Questions (NQ) (Kwiatkowski et al., 2019) and RGB (Chen et al., 2024a) benchmarks, which comprise real-world queries. Turning to multi-hop QA, we select HotpotQA (Yang et al., 2018) and 2WikiMultihopQA (Ho et al., 2020), enabling evaluation of complex reasoning types such as comparison and composition. Across all tasks, we report accuracy as the primary metric. More details of the datasets and metrics are in Appendix B.

Baselines To demonstrate its broad effectiveness, we evaluate LFD across four widely used LLMs: Llama2-7B (Touvron et al., 2023), Mistral-7B (Jiang et al., 2023a), DeepSeek-7B (Bi et al., 2024), and Qwen3-8B (Team, 2025). We further benchmark LFD against three established decoding strategies: **(1) Greedy Decoding (GD)** employs standard autoregressive decoding. We augment GD with varying numbers of noise documents, denoted as GD (0), GD (4), GD (8), and GD (12). **(2) Contrastive Search (CS)** (Su et al., 2022) balances response quality and diversity to encourage more coherent outputs. **(3) Decoding by Contrasting**

		NQ	RGB	HotpotQA			2WikiMultihopQA				
				Compare	Bridge	Total	Compare	Bridge	Inf	Compose	Total
Llama2-7B	GD (0)	0.5745	0.8900	0.4755	0.4108	0.4237	0.3095	0.2815	0.2607	0.2928	0.2904
	GD (4)	0.5559	0.8267	0.5084	0.4049	0.4257	0.2559	0.2564	0.2048	0.2405	0.2433
	GD (8)	0.5965	0.8300	0.5091	0.4096	0.4317	**0.3257**	0.2983	0.2789	0.2995	0.3030
	GD (12)	**0.6383**	0.8200	0.5286	0.4236	0.4447	0.3204	**0.3202**	0.2646	0.2943	0.3027
	CS	0.5775	0.8567	0.4284	0.4049	0.4096	0.2800	0.2462	0.2464	0.2866	0.2712
	DoLA	0.5809	0.8733	0.4438	0.3883	0.3995	0.2738	0.2411	0.2289	0.2592	0.2550
	LFD (Random)	0.5928	0.8900	0.4385	0.4295	0.4313	0.3065	0.2845	0.2744	0.3066	0.2978
	LFD	0.5949	**0.9067**	**0.5453**	**0.4295**	**0.4528**	0.3174	0.3005	**0.3043**	**0.3232**	**0.3144**
Mistral-7B	GD (0)	0.6130	0.8900	0.5864	0.5889	0.5884	0.5518	0.5204	0.5605	0.5337	0.5385
	GD (4)	0.5667	0.8033	0.5494	0.5348	0.5377	0.3681	0.3443	0.3173	0.3296	0.3406
	GD (8)	0.5678	0.7700	0.5326	0.5260	0.5273	0.2963	0.2874	0.2503	0.2582	0.2728
	GD (12)	0.5814	0.8267	0.5440	0.5461	0.5457	0.3098	0.2888	0.2503	0.2655	0.2795
	CS	0.5598	0.7433	0.4149	0.5117	0.4922	0.4346	0.3873	0.4486	0.4740	0.4423
	DoLA	0.6142	0.8900	0.5931	0.5870	0.5883	0.5366	0.5018	0.5572	0.5239	0.5262
	LFD (Random)	0.6270	0.9133	0.5985	0.6016	0.6009	0.5687	0.5317	0.5767	0.5506	0.5541
	LFD	**0.6357**	**0.9367**	**0.6026**	**0.6093**	**0.6079**	**0.5813**	**0.5434**	**0.5982**	**0.5645**	**0.5680**
DeepSeek-7B	GD (0)	0.5250	0.8233	0.3282	0.4033	0.3883	0.2625	0.2414	0.2663	0.2609	0.2609
	GD (4)	0.5216	0.8300	0.3968	0.4437	0.4343	0.2807	0.2939	0.2744	0.2730	0.2796
	GD (8)	0.5226	0.8800	0.4506	0.4579	0.4564	0.3578	0.3603	0.3407	0.3463	0.3515
	GD (12)	0.5204	**0.8933**	0.4573	**0.4618**	**0.4609**	0.3701	0.3705	0.3349	0.3614	0.3622
	CS	0.5322	0.8000	0.4028	0.4584	0.4472	0.3929	0.3833	0.3888	0.4077	0.3964
	DoLA	0.3755	0.5033	0.3490	0.3003	0.3100	0.3039	0.3118	0.2484	0.2594	0.2803
	LFD (Random)	0.4858	0.6700	0.3847	0.3734	0.3757	0.3466	0.3563	0.3277	0.3320	0.3403
	LFD	**0.5412**	0.8267	**0.4801**	0.4466	0.4533	**0.4492**	**0.4227**	**0.4194**	**0.4223**	**0.4285**
Qwen3-8B	GD (0)	0.7318	0.9571	0.6960	0.6708	0.6759	0.6637	0.6335	0.5897	0.6085	0.6250
	GD (4)	0.7213	0.9467	0.6658	0.6544	0.6567	0.6044	0.5729	0.4948	0.5268	0.5517
	GD (8)	0.7233	0.9500	0.6584	0.6568	0.6571	0.6117	0.5780	0.5078	0.5370	0.5605
	GD (12)	0.7133	0.9533	0.6530	0.6582	0.6571	0.6146	0.5722	0.5052	0.5462	0.5634
	CS	0.7204	0.9533	0.7081	0.6762	0.6826	0.6468	0.6204	0.5754	0.6015	0.6134
	DoLA	0.7168	0.9533	0.7014	0.6703	0.6766	0.6485	0.6058	0.5650	0.5929	0.6057
	LFD (Random)	0.7357	0.9567	0.7108	0.6935	0.6970	0.6627	0.6334	0.6014	0.6136	0.6289
	LFD	**0.7380**	**0.9600**	**0.7182**	**0.6974**	**0.7016**	**0.6663**	**0.6342**	**0.6034**	**0.6148**	**0.6301**

Table 1: The accuracy performance comparison of different methods on four datasets. Compare, Bridge, Inf, and Compose represent distinct reasoning question types (details in Appendix B). **Bold** values indicate the best performance, while underlined values represent the second-best.

Layers (DoLA) (Chuang et al., 2024) is a contrastive decoding approach that reduces hallucinations by comparing predictions across different model layers. **(4) LFD (Random)** is a variant of LFD that randomly selects an intermediate layer to fuse during the final decoding stage. Implementation details are provided in Appendix C.

4.2 Main Results

Table 1 summarizes the accuracy of RAG based on four different models across four QA datasets. From the table, we have the following observations:

(1) **LFD matches or exceeds the performance of noise-injection strategies.** We can see our method demonstrates consistent performance gains, ranging from minimal 0.29% (Qwen3-8B, RGB) to maximal 16.76% (DeepSeek-7B, 2WikiMultihopQA) improvement. Injecting noise, i.e., GD (12), shows performance gains on some datasets and models, with the highest improvement being 10.13% (DeepSeek-7B, 2WikiMultihopQA). However, it significantly degrades performance for Mistral-7B and Qwen3-8B across all datasets, leading to a notable reduction in accuracy (maximum $\Delta = -26.57\,\%$ on Mistral-7B). **Additionally, our method also achieves significantly lower computational overhead than noise-injection strategies (see Appendix D).**

(2) **LFD outperforms decoding strategies without noise injection.** Compared to alternative decoding methods, LFD consistently delivers superior performance. While methods like DoLA and CS show strong results in specific cases, e.g., achieving 1.94% and 13.55% gains on DeepSeek-7B with the 2WikiMultihopQA dataset, they occasionally underperform even relative to the greedy decoding baseline. In particular, DoLA shows the largest decline of 30% on DeepSeek-7B with the RGB dataset, while CS exhibits a maximum drop of 14.57% on Mistral-7B with the RGB dataset. These results indicate that both strategies are sensitive to specific model architectures or data characteristics. **Additional comparisons of these methods under different noise levels and the case study are provided in Appendix E and Appendix H, respectively.**

(3) **The fusion layer selection plays a critical role for RAG.** Our dynamic layer selection strategy consistently outperforms the random approach, with an average improvement of 3.11%, particularly achieving a 15.67% gain on DeepSeek-7B/RGB, demonstrating its efficacy.

4.3 Analysis Experiments

Effects of late-stage layer Integration We assess how different layer selection ranges affect performance using Llama2-7B and Mistral-7B across four benchmark datasets in Table 2. As Table 2 shows, our approach (selecting from layers 16–32) consistently achieves higher accuracy than select-

		NQ	RGB	HotpotQA			2WikiMultihopQA				
				Compare	Bridge	Total	Compare	Bridge	Inf	Compose	Total
Llama2-7B	GD (0)	0.5745	0.8900	0.4755	0.4108	0.4237	0.3095	0.2815	0.2607	0.2928	0.2904
	LFD[0, 16)	0.5758	0.8833	0.4371	0.4027	0.4096	0.2956	0.2582	0.2529	0.2739	0.2731
	LFD[0, 32)	**0.5957**	0.9067	0.4586	0.4208	0.4284	**0.3181**	**0.3038**	0.3004	**0.3237**	**0.3151**
	LFD[16, 32)	0.5949	**0.9067**	**0.5454**	**0.4295**	**0.4528**	0.3174	0.3005	**0.3043**	0.3232	0.3144
Mistral-7B	GD (0)	0.6130	0.8900	0.5864	0.5889	0.5884	0.5518	0.5204	0.5605	0.5337	0.5385
	LFD[0, 16)	0.6081	0.8867	0.5864	0.5919	0.5908	0.5551	0.5218	0.5650	0.5402	0.5429
	LFD[0, 32)	0.6344	0.9367	0.5938	0.5994	0.5982	**0.5842**	**0.5492**	0.5917	0.5639	**0.5690**
	LFD[16, 32)	**0.6357**	**0.9367**	**0.6026**	**0.6093**	**0.6079**	0.5813	0.5434	**0.5982**	**0.5645**	0.5680

Table 2: Comparison of accuracy between different layer selection ranges under dynamic layer selection strategy. **Bold** indicate the best performance, while underline represent the second-best. LFD[0, 16) and LFD[16, 32) mean selecting layers from the earlier and later half respectively.

ing from earlier layers (layers 0-16), with an average gain of 3.01%. Notably, our approach provides marginal accuracy gains ($\leq$0.1%) over full-range selection (layers 0-32), which indicates the advantage of the proposed layer selection strategy. These results demonstrate the greater efficacy of deeper layers for utilizing knowledge in RAG. Results for DeepSeek-7B and Qwen3-8B are provided in Appendix F.

Effects of the lowest IKS layer selection We evaluate the effectiveness of the lowest IKS layer selection on the NQ and HotpotQA datasets. Extended results across models and datasets appear in Appendix G. First, we compare our dynamic strategy with fixed-layer selection LFD (Fixed) in Figure 4(a-b). Second, we analyze the distribution of dynamical layer selections against the optimal fixed-layer baseline on NQ (Figure 4(c)).

Key findings emerge: **(1) Fixed-layer selection requires dataset-specific validation for optimality**. While fixed-layer achieves peak performance at layer $24-26$ (NQ) and layer 22 (HotpotQA), these layers differ across datasets, necessitating extra validating datasets. **(2) The lowest IKS tends to achieve near-optimal performance with small margins**. Compared to the best fixed-layer results, the lowest IKS exhibits performance gaps of only 1.8% (NQ) and 0.2% (HotpotQA), demonstrating robust generalization without dataset-specific tuning. **(3) Dynamic layer selection concentrates near the optimal fixed-layer.** For NQ dataset, the lowest IKS selections cluster around layer 26 (Figure 4(c)), aligning closely with the optimal fixed layer despite stochasticity.

Effects of Conflicting Documents Real-world retrieval often involves conflicting documents, which are semantically similar to the query but do not contain the correct answer (Cuconasu et al., 2024). Following (Chen et al., 2024a), we utilize the RGB dataset with five documents per prompt and vary the ratio of conflicting documents from 0 to 0.8. As shown in Table 3, LFD consistently outperforms GD (0), demonstrating its robustness when encountering misleading interference.

		0	0.2	0.4	0.6	0.8
Llama2-7B	GD (0)	0.9267	0.9167	0.9167	0.8933	0.8500
	LFD	**0.9267**	**0.9167**	**0.9167**	**0.8967**	**0.8833**
Mistral2-7b	GD (0)	0.8933	0.8933	0.8867	0.8867	0.8367
	LFD	**0.9567**	**0.9333**	**0.9233**	**0.9167**	**0.8800**

Table 3: Accuracy comparison under varying ratios of conflicting documents (from 0 to 0.8) on the RGB dataset. **Bold** values indicate the best performance.

		NQ	HotpotQA
Glm4-9B	GD (0)	0.7980	0.6620
	LFD	**0.8280**	**0.7340**
Llama2-13B	GD (0)	0.6500	0.5940
	LFD	**0.7280**	**0.5980**
Qwen3-32B	GD (0)	0.8440	0.7660
	LFD	**0.8480**	**0.7780**

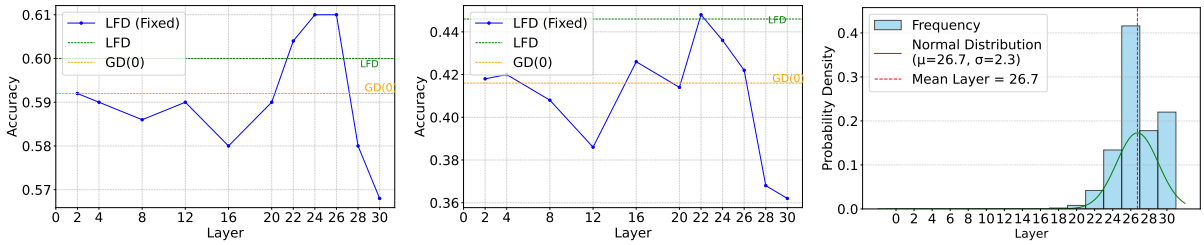
Table 4: Comparison of accuracy across different larger models on the NQ and HotpotQA datasets. **Bold** values indicate the best performance.

Results for Larger Models Following the same experimental setup (Section 4.1), We extend our evaluation to larger models on NQ and HotpotQA. As shown in Table 4, LFD consistently yields significant gains (e.g., +7.8% on Llama2-13B for NQ).

5 Related Works

5.1 Retrieval Augmented Generation

Retrieval-Augmented Generation (RAG) enhances model reasoning by integrating relevant external knowledge retrieved through user queries. Recent advances focus on three directions: refined retrieval mechanisms (Jeong et al., 2024; Chan et al., 2024), structured knowledge organization (Procko and Ochoa, 2024; Baek et al., 2023), and optimized context embedding (Hu et al., 2024; Sahoo et al., 2024). Self-RAG (Asai et al., 2024) and FLARE (Jiang et al., 2023c) achieve adaptive retrieval through self-evaluation and uncertainty prediction respectively, dynamically optimizing knowledge acquisition. GraphRAG (He et al., 2024; Edge et al.,



(a) LFD vs. LFD(Fixed) on NQ. (b) LFD vs. LFD(Fixed) on HotpotQA. (c) Histgram of layer selection in LFD.

Figure 4: Comparison between LFD and LFD (Fixed) on the NQ (a) and HotpotQA (b) datasets. (c) illustrates the layer selection distribution in LFD compared to the optimal fixed layer selection.

2024; Jiang et al., 2023b) advances reasoning capabilities by constructing document-derived knowledge graphs that capture semantic relationships for multi-hop inference. Embedding optimizations include noise injection (Cuconasu et al., 2024; Wu et al., 2024) to counter overfitting through strategic low-relevance document insertion, and context re-ranking (Yu et al., 2024; Dong et al., 2024) that prioritizes high-utility knowledge via learned document scoring. While these methods enhance knowledge orchestration through pipeline improvements, they systematically neglect the internal mechanisms through which LLMs process external information during generation. Our work bridges this fundamental gap by surfacing stratified knowledge integration patterns across LLM's layers through systematic layer-wise analysis.

5.2 Decoding Strategy in LLMs

Decoding strategies are pivotal in transforming raw model probabilities into coherent text outputs, critically influencing the quality and factual integrity of LLM generations (Shi et al., 2024). While traditional methods like greedy decoding and beam search (Meister et al., 2020; Xie et al., 2023) remain prevalent, recent work has introduced advanced techniques to address their limitations. Contrastive search (CS) (Su et al., 2022) balances diversity and coherence by selecting tokens through a weighted combination of probability and semantic dissimilarity to preceding context, mitigating repetition while preserving fluency. Simple Decoding (FSD) (Yang et al., 2024) suppresses redundant patterns by dynamically constructing an anti-language model to penalize overused token sequences. DoLa (Chuang et al., 2024) addresses hallucinations by contrasting later-layer logit distributions with earlier ones, prioritizing factually consistent predictions. Building on these advances, we propose a novel decoding strategy for RAG that dynamically balances external retrieved knowledge with the model's internal parametric knowledge, enhancing factual accuracy by mitigating interference from outdated or conflicting internal representations.

5.3 Hallucinations in LLMs

Hallucinations in LLMs refer to instances where the model generates false or unsupported information not grounded in its reference data (Perković et al., 2024). Existing mitigation strategies include multi-agent debating, where multiple LLM instances collaborate to detect inconsistencies through iterative debates (Chen et al., 2024b; Du et al., 2023); self-consistency verification, which aggregates and reconciles multiple reasoning paths to reduce individual errors (Wang et al., 2022); and model editing, which which directly modifies neural network weights to correct systematic factual errors (Zhang et al., 2024; Hartvigsen et al., 2023). While RAG systems aim to ground responses in retrieved external knowledge, recent studies show that they still exhibit hallucinations, especially those that contradict the retrieved content (Sun et al., 2025b). To address this limitation, our work conducts an empirical study analyzing how LLMs internally process external knowledge in RAG settings by controlling the noise from different granularity. Based on these findings, we propose a novel decoding method designed to improve answer accuracy and reduce hallucination by enhancing the integration of retrieved evidence.

6 Conclusion

By analyzing how noise injection amplifies external knowledge exploitation in LLMs, we establish a functional demarcation across LLMs' layers: shallow (local context), intermediate (external knowledge), and deep (internal parametric knowledge). Leveraging this, we propose LFD, a training-free decoding strategy that fuses intermediate-layer representations to enhance external knowledge integration in final outputs via a the lowest internal knowledge score to pinpoint the ideal fusion layer. Experiments across diverse benchmarks demonstrate that LFD enhances factual grounding in RAG systems while incurring minimal computational overhead.

Limitations

Our work has the following two main limitations.

First, LFD occasionally yields slightly lower gains than GD(12). However, it maintains consistent superiority over GD(0) in every experimental configuration. More importantly, LFD achieves these results with only 1/10 of the context tokens and 0.4x the latency of GD(12) (see AppendixD). This superior computational efficiency, combined with its stable performance, underscores the overall effectiveness of LFD.

Second, While noise-enhanced RAG has proven effective (Cuconasu et al., 2024; Sun et al., 2025a), a rigorous theoretical justification for its success remains unclear. Our study utilizes noise primarily as an intervention tool to analyze layer-specific effects, leading to the development of LFD to bypass costly augmentation. Theoretical exploration of such noise-driven improvements remains a promising direction for future research.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 78–106, Toronto, Canada. Association for Computational Linguistics.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *CoRR*, abs/2303.08112.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, and 1 others. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, and 9 others. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024a. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024b. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for RAG systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 719–729. ACM.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Dario Di Palma. 2023. Retrieval-augmented recommender system: Enhancing recommender systems with large language models. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1369–1373.

Jialin Dong, Bahare Fatemi, Bryan Perozzi, Lin F. Yang, and Anton Tsitsulin. 2024. Don't forget to connect! improving rag with graph-based reranking. *Preprint*, arXiv:2405.18414.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and

Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 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*, pages 6491–6501.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Tom Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2023. Aging with grace: Lifelong model editing with discrete key-value adaptors. *Advances in Neural Information Processing Systems*, 36:47934–47959.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *Advances in Neural Information Processing Systems*, 37:132876–132907.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhibo Hu, Chen Wang, Yanfeng Shu, Hye-Young Paik, and Liming Zhu. 2024. Prompt perturbation in retrieval-augmented generation based large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1119–1130.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 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)*, pages 7036–7050, Mexico City, Mexico. Association for Computational Linguistics.

Jianchao Ji, Zelong Li, Shuyuan Xu, Wenyue Hua, Yingqiang Ge, Juntao Tan, and Yongfeng Zhang. 2024. Genrec: Large language model for generative recommendation. In *European Conference on Information Retrieval*, pages 494–502. Springer.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *Preprint*, arXiv:2310.06825.

Jinhao Jiang, Kun Zhou, Xin Zhao, and Ji-Rong Wen. 2023b. UniKGQA: Unified retrieval and reasoning for solving multi-hop question answering over knowledge graph. In *The Eleventh International Conference on Learning Representations*.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023c. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2025. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4745–4759, Albuquerque, New Mexico. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*.

Jianhua Lin. 2002. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12.

Clara Meister, Tim Vieira, and Ryan Cotterell. 2020. Best-first beam search. *Transactions of the Association for Computational Linguistics*, 8:795–809.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878.

Chanhee Park, Hyeonseok Moon, Chanjun Park, and Heuiseok Lim. 2025. MIRAGE: A metric-intensive benchmark for retrieval-augmented generation evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2883–2900, Albuquerque, New Mexico. Association for Computational Linguistics.

Gabrijela Perković, Antun Drobnjak, and Ivica Botički. 2024. Hallucinations in llms: Understanding and addressing challenges. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, pages 2084–2088.

Tyler Thomas Procko and Omar Ochoa. 2024. Graph retrieval-augmented generation for large language models: A survey. In *2024 Conference on AI, Science, Engineering, and Technology (AIxSET)*, pages 166–169.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Alireza Salemi and Hamed Zamani. 2024. Towards a search engine for machines: Unified ranking for multiple retrieval-augmented large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 741–751.

Chufan Shi, Haoran Yang, Deng Cai, Zhisong Zhang, Yifan Wang, Yujiu Yang, and Wai Lam. 2024. A thorough examination of decoding methods in the era of llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8601–8629.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. *Advances in Neural Information Processing Systems*, 35:21548–21561.

Hao Sun, Chenming Tang, Gengyang Li, and Yunfang Wu. 2025a. Lost in the passage: Passage-level in-context learning does not necessarily need a "passage". *CoRR*, abs/2502.10634.

ZhongXiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2025b. RedeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *The Thirteenth International Conference on Learning Representations*.

Qwen Team. 2025. Qwen3.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jinyang Wu, Feihu Che, Chuyuan Zhang, Jianhua Tao, Shuai Zhang, and Pengpeng Shao. 2024. Pandora's box or aladdin's lamp: A comprehensive analysis revealing the role of rag noise in large language models. *arXiv preprint arXiv:2408.13533*.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, James Xu Zhao, Min-Yen Kan, Junxian He, and Michael Xie. 2023. Self-evaluation guided beam search for reasoning. *Advances in Neural Information Processing Systems*, 36:41618–41650.

Haoyi Xiong, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and Sumi Helal. 2024. When search engine services meet large language models: visions and challenges. *IEEE Transactions on Services Computing*.

Haoran Yang, Deng Cai, Huayang Li, Wei Bi, Wai Lam, and Shuming Shi. 2024. A frustratingly simple decoding method for neural text generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 536–557, Torino, Italia. ELRA and ICCL.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 121156–121184. Curran Associates, Inc.

Shaolei Zhang, Tian Yu, and Yang Feng. 2024. Truthx: Alleviating hallucinations by editing large language models in truthful space. *arXiv preprint arXiv:2402.17811*.

A Comprehensive Analysis of External Knowledge Intervention in RAG

In this section, we expand our analysis by incorporating a multi-hop question answering dataset HotpotQA to further quantify the impact of external knowledge. As shown in Figure 5, LLaMA2-7B exhibits a consistent three-stage pattern of knowledge utilization across layers, as reflected by the SimHidden scores: early layers (1–14), middle layers (15–26), and deeper layers (27–32). This stratification is further supported by the DiffAttn scores, which peak at layer 21 and remain lower in both ealier and latter layers, reinforcing the validity of the three-stage division.

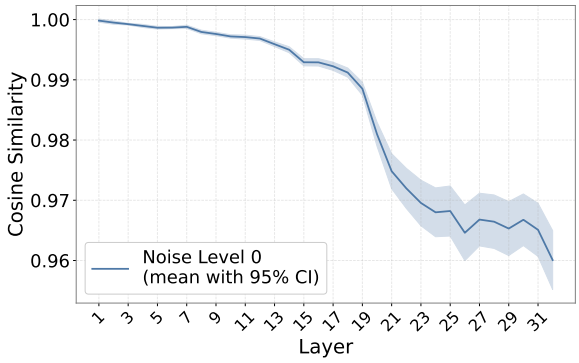
To assess the generality of this phenomenon, we evaluate a diverse set of models across various architectures and scales. Specifically, we include models with different architectural designs: Mistral-7B, DeepSeek-7B, and Qwen3-8B. Results are shown in Figures 6–8. Despite these architectural variations, all models exhibit a consistent three-phase trend in SimHidden scores and a middle-peak pattern in DiffAttn scores across both the NQ and HotpotQA datasets (e.g., for Mistral-7B on NQ, the three phases span layers 1–14, 15–29, and 30–32, with the DiffAttn peak occurring at layer 20). This consistency extends to larger scales, such as Llama2-13B and Qwen3-32B (Figures 9–10).

B Details About Datasets and Metrics

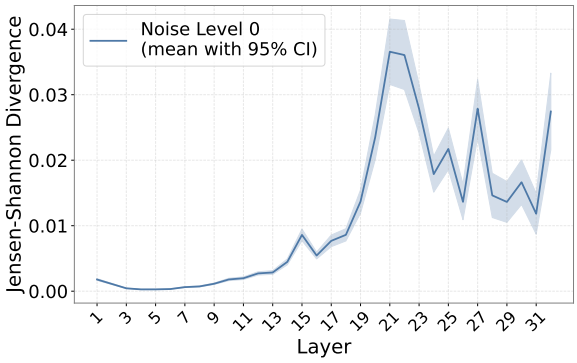
Following the experimental setups of (Cuconasu et al., 2024; Wu et al., 2024; Jeong et al., 2024), we evaluate our approach across following datasets: **(1) Natural Questions (NQ)** (Kwiatkowski et al., 2019), a large-scale QA dataset based on real Google search queries. **(2) RGB** (Chen et al., 2024a), a RAG benchmark that evaluates models' ability to utilize retrieved information, focusing on noise robustness, negative rejection, information integration, and counterfactual robustness. We use its English test set for evaluation. **(3) HotpotQA (HQA)** (Yang et al., 2018), which requires multi-hop reasoning over multiple documents, featuring both compare and bridge question types: compare questions involve contrasting information from multiple sources, while bridge questions require connecting intermediate facts, both necessitating the synthesis of external documents with the model's internal knowledge. We evaluate all methods on their dev set, reporting results under the categories Compare, Bridge, and Total in the main results table. **(4) 2WikiMultihopQA (2WQA)** (Ho et al., 2020), which presents a more challenging multi-hop QA scenario with four distinct task types: comparison (comparing information), bridge comparison (connecting intermediate facts for comparison), inference (deriving conclusions), and compositional (integrating multiple facts). We use its dev set for evaluation, reporting separate performance for each question type (denoted as Comapre, Bridge, Inf, Compose and Total) in the experimental results.

Regarding the experimental data, since each sample in the aforementioned benchmark datasets is accompanied by multiple retrieved documents, we construct the input context using these documents, supplemented with golden documents, i.e., the ground-truth passages that contain the information necessary to answer the question. This setup, following prior work (Lewis et al., 2020b; Cuconasu et al., 2024; Krishna et al., 2025; Park et al., 2025), allows us to evaluate the model's ability to effectively leverage external knowledge when it is explicitly provided in the input context. To construct the noise document set for each sample, we consider all documents within the corresponding dataset. We utilize the all-MiniLM-L6-v2 model to retrieve documents from this collection that exhibit the lowest relevance scores relative to the query of the given sample.

We use accuracy as our primary evaluation metric. For the NQ and RGB datasets, which include samples with multiple acceptable answer variants (e.g., alternative phrasings of the same concept), we follow the evaluation protocol established in



(a) SimHidden (Smaller is better) on HotpotQA dataset.



(b) DiffAttn (Larger is better) on HotpotQA dataset.

Figure 5: (a) Average SimHidden scores (with 95% confidence intervals) across layers when noise level = 0; (b) Average DiffAttn scores (with 95% confidence intervals) across layers when noise level = 0. Results are from Llama2-7B on HotpotQA dataset.

	Input Length	Latency (ms)	Throughput (tokens/s)	GPU Memory (MB)
GD (0)	239 (×1.00)	42.23 (×1.00)	24.29 (×1.00)	144.05 (×1.00)
GD (4)	958 (×4.01)	73.42 (×1.74)	14.95 (×0.62)	572.23 (×3.97)
GD (8)	1675 (×7.01)	97.99 (×2.32)	11.92 (×0.49)	997.81 (×6.93)
GD (12)	2396 (×10.02)	129.67 (×3.07)	9.31 (×0.38)	1426.87 (×9.91)
DoLA	239 (×1.00)	49.66 (×1.18)	20.58 (×0.85)	201.00 (×1.39)
LFD	239 (×1.00)	52.25 (×1.24)	19.56 (×0.81)	203.15 (×1.41)

Table 5: Input Length, Decoding Latency (ms), Throughput (tokens/s), and GPU Overhead (MB).

(Cuconasu et al., 2024; Kandpal et al., 2023; Liu et al., 2024). A model's response is marked correct if it matches any annotated ground-truth answer.

C Implementation details

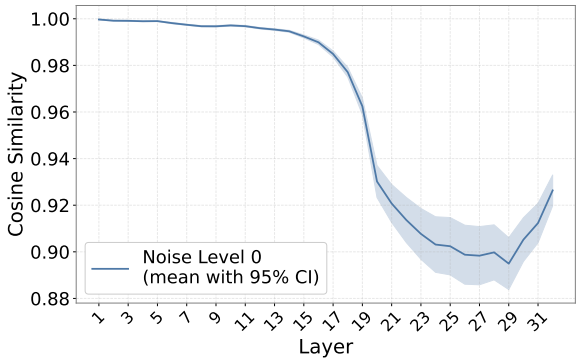
Both LFD and DoLA require the implementation of dynamic layer selection strategies. For LFD, we prioritize layers in the latter half of the architecture (e.g., layers 16–32 for 32-layer models like Llama2-7B and Mistral-7B). In contrast, DoLA selects layers across the entire depth (layers 0–32) based on divergence from the final layer's predictions. Similar configurations apply to DeepSeek-7B (30 layers: LFD uses 15–30; DoLA uses 0–30) and Qwen3-8B (36 layers: LFD uses 18–36; DoLA uses 0–36). Following DoLA's convention, we restrict candidates to even-numbered layers within these ranges for efficiency. Contrastive Search uses a degeneration penalty $\alpha = 0.6$ and top-k candidate size $k = 5$, adopting the parameters from (Su et al., 2022). Regarding the gating mechanism in LFD, we set the hyperparameter $\tau = 0.1$ following the configuration in (Li et al., 2022)

D Latency, Throughput & Memory Usage

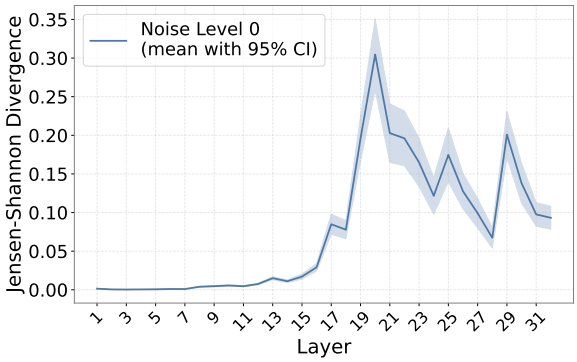
We compare the decoding latency, throughput, and GPU overhead between LFD and the greedy decoding method (with varying levels of noise), and the experimental results are illustrated in Table 5. The results demonstrate that, compared to the noise-injection baselines, LFD exhibits advantage in terms of decoding time and memory overhead. Furthermore, when compared to noise-free decoding baseline DoLA, our approach incurs just 1.05× the latency and 1.01× the memory usage, **keeping efficiency on par with state-of-the-art decoding methods**.

E Noise Injection Analysis for CS, DoLA, and LFD

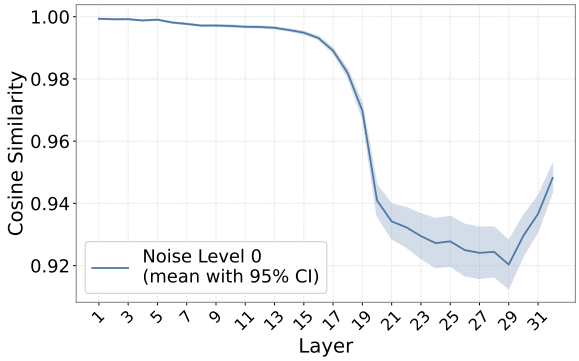
To investigate the performance of different decoding strategies under varying noise conditions, we introduce controlled noise levels (4, 8, and 12) to CS, DoLA, and LFD. Experiments are conducted using LLaMA2-7B and Mistral-7B on the NQ and HotpotQA datasets, respectively. As shown in Table 6 and Table 7, LFD generally achieves higher accuracy than other decoding methods under various noise levels, indicating comparatively stronger robustness to noise. Notably, a significant improvement in LFD's accuracy is observed under moderate noise conditions. For example, on the NQ dataset using LLaMA2-7B, applying noise level 12 yields a 6.7%accuracy gain compared to the noise-free setting. This suggests that controlled noise exposure can further improve the performance of LFD.



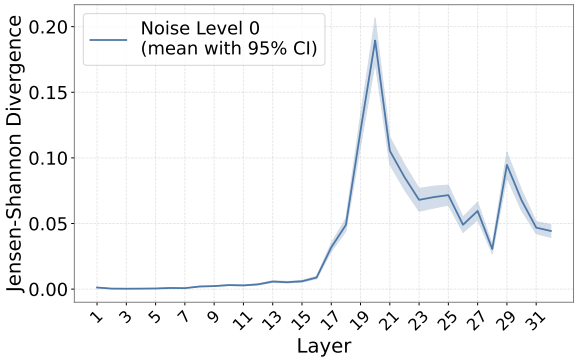
(a) SimHidden (Smaller is better) on NQ dataset.



(b) DiffAttn (Larger is better) on NQ dataset.



(c) SimHidden (Smaller is better) on HotpotQA dataset.



(d) DiffAttn (Larger is better) on HotpotQA dataset.

Figure 6: (a) Average SimHidden scores (with 95% confidence intervals) across layers when noise level = 0; (b) Average DiffAttn scores (with 95% confidence intervals) across layers when noise level = 0. Results are from Mistral-7B on NQ dataset and HotpotQA dataset.

F Ablation Study on Layer Selection Range with DeepSeek-7B and Qwen3-8B

To complement the layer selection range analysis in Section 4.3, we extend our experiments to include DeepSeek-7B and Qwen3-8B models across four datasets: NQ, RGB, HotpotQA, and 2WikiMultihopQA. As shown in Table 8, our findings remain consistent with those presented in Table 2, demonstrating that selecting layers from the latter half of the model consistently yields superior performance compared to earlier ranges.

G Evaluating IKS Score Effectiveness Across Different Models and Datasets

To further validate the effectiveness of the lowest IKS layer selection discussed in Section 4.3, we conduct additional evaluations using models: Mistral-7B, DeepSeek-7B, and Qwen3-8B. Our experiments encompass both a single-hop QA dataset (NQ) and a multi-hop QA dataset (HotpotQA). The complete results are presented in Figures 11–13.

H Case Studies

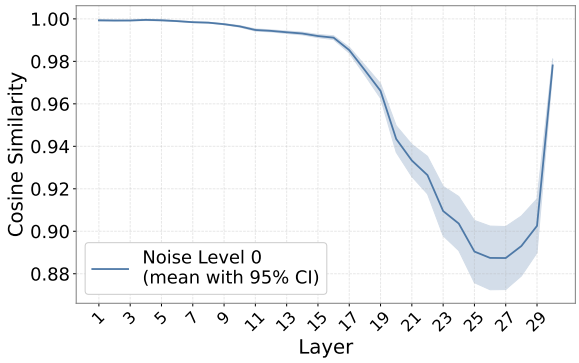
We provide some case studies using the Llama2-7B model across four benchmark datasets: NQ, RGB, HotpotQA, and 2WikiMultihopQA. As shown in Tables 9–12, our method enables the model to better adhere to the provided context and correctly identify answers within the given information.

I Computational Details

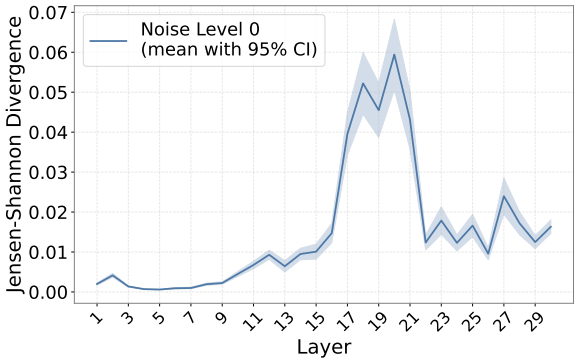
All experiments are performed on a GPU-accelerated computing system equipped with NVIDIA GeForce RTX 3090 graphics processors (24GB GDDR6X VRAM each), supported by dual Intel Xeon Gold 6271C CPUs (2.6GHz base frequency, 48 cores total) and 251GB of system memory.

J The License For Artifacts and Data Consent

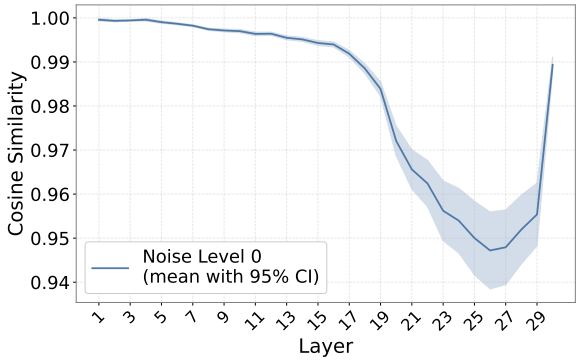
All models and datasets used in this work are publicly available artifacts released for academic research. Regarding model licenses, Llama-2



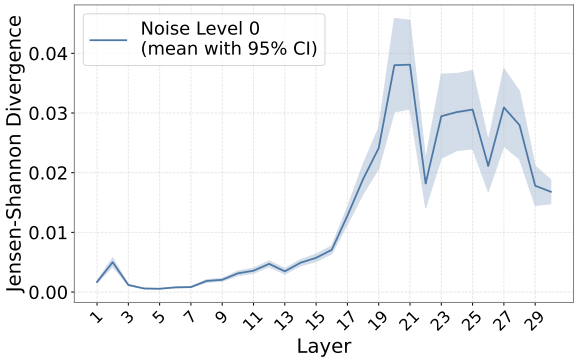
(a) SimHidden (Smaller is better) on NQ dataset.



(b) DiffAttn (Larger is better) on NQ dataset.



(c) SimHidden (Smaller is better) on HotpotQA dataset.



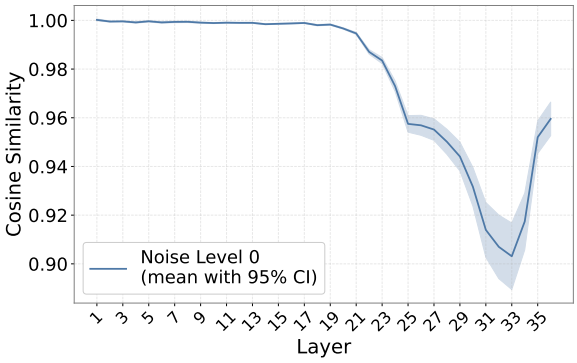
(d) DiffAttn (Larger is better) on HotpotQA dataset.

Figure 7: (a) Average SimHidden scores (with 95% confidence intervals) across layers when noise level = 0; (b) Average DiffAttn scores (with 95% confidence intervals) across layers when noise level = 0. Results are from DeepSeek-7B on NQ dataset and HotpotQA dataset.

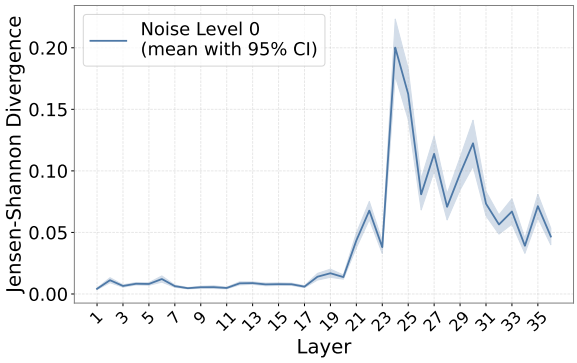
(7B/13B) is used under the Llama 2 Community License Agreement; Mistral-7B and Qwen-3 (8B/32B) are governed by the Apache License 2.0; DeepSeek-7B and GLM-4-9B are used under their respective proprietary model licenses (DeepSeek and GLM-4 licenses). For evaluation datasets, Natural Questions (NQ) and 2WikiMultihopQA are licensed under the Apache License 2.0, while HotpotQA and RGB are used under CC BY-SA 4.0 and CC BY-NC-SA 4.0, respectively. All datasets are pre-anonymized and filtered for offensive content by the original creators. Our use of these materials is strictly for academic purposes and is fully consistent with their specified intended use.

	NQ	HotpotQA		
		Compare	Bridge	Total
GD (0)	0.5745	0.4755	0.4108	0.4237
CS (0)	0.5775	0.4284	0.4049	0.4096
DoLA (0)	0.5809	0.4438	0.3883	0.3995
LFD (0)	**0.5949**	**0.5453**	**0.4295**	**0.4528**
GD (4)	0.5559	**0.5084**	0.4049	0.4257
CS (4)	**0.5657**	0.4546	0.4199	0.4269
DoLA (4)	0.5574	0.4775	0.4111	0.4244
LFD (4)	0.5637	0.5071	**0.4304**	**0.4458**
GD (8)	0.5965	0.5091	0.4096	0.4317
CS (8)	0.6054	0.4694	0.4236	0.4328
DoLA (8)	0.6023	0.4956	0.4199	0.4351
LFD (8)	**0.6216**	**0.5400**	**0.4400**	**0.4601**
GD (12)	0.6383	0.5286	0.4236	0.4447
CS (12)	0.6410	0.4781	0.4419	0.4492
DoLA (12)	0.6458	0.5158	0.4306	0.4477
LFD (12)	**0.6622**	**0.5521**	**0.4532**	**0.4731**

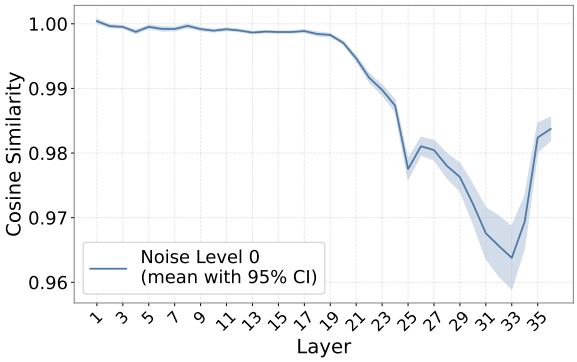
Table 6: Accuracy performance comparison of different decoding methods with varying levels of noise, evaluated on the NQ and HotpotQA datasets using the LLaMA2-7B. **Bold** values indicate the best performance, while underlined values represent the second-best.



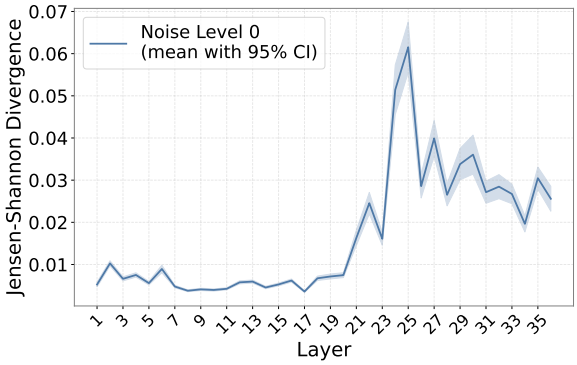
(a) SimHidden (Smaller is better) on NQ dataset.



(b) DiffAttn (Larger is better) on NQ dataset.



(c) SimHidden (Smaller is better) on HotpotQA dataset.

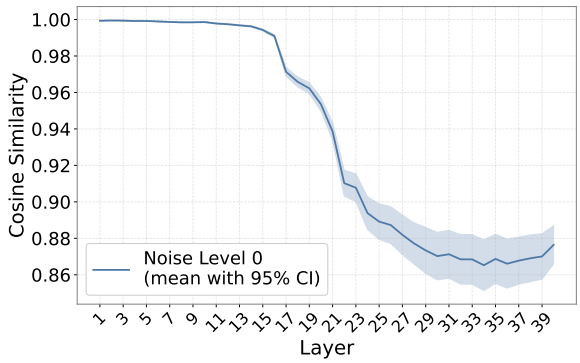


(d) DiffAttn (Larger is better) on HotpotQA dataset.

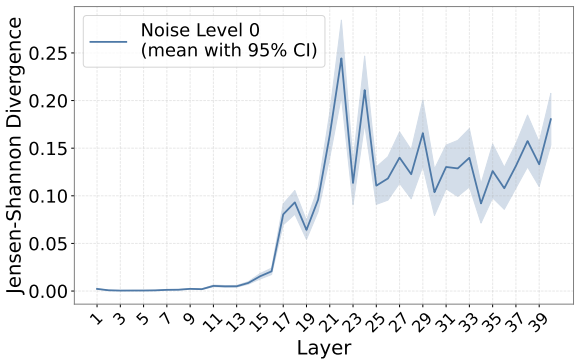
Figure 8: (a) Average SimHidden scores (with 95% confidence intervals) across layers when noise level = 0; (b) Average DiffAttn scores (with 95% confidence intervals) across layers when noise level = 0. Results are from Qwen3-8B on NQ dataset and HotpotQA dataset.

	NQ	HotpotQA		
		Compare	Bridge	Total
GD (0)	0.6130	0.5864	0.5889	0.5884
CS (0)	0.5598	0.4149	0.5117	0.4922
DoLA (0)	0.6142	0.5931	0.5870	0.5883
LFD (0)	**0.6357**	**0.6026**	**0.6093**	**0.6079**
GD (4)	0.5667	**0.5494**	0.5348	0.5377
CS (4)	0.5014	0.2065	0.2778	0.2635
DoLA (4)	0.5625	0.5373	0.5331	0.5340
LFD (4)	**0.6036**	0.5467	**0.5502**	**0.5495**
GD (8)	0.5678	**0.5326**	0.5260	0.5273
CS (8)	0.5201	0.1137	0.1926	0.1768
DoLA (8)	0.5659	0.5111	0.5252	0.5223
LFD (8)	**0.5966**	0.5259	**0.5446**	**0.5409**
GD (12)	0.5814	**0.5440**	0.5461	0.5457
CS (12)	0.5334	0.1432	0.2328	0.2149
DoLA (12)	0.5845	0.5293	0.5492	0.5452
LFD (12)	**0.5943**	0.5346	**0.5640**	**0.5581**

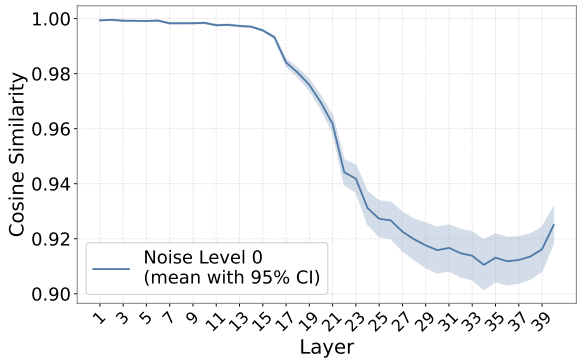
Table 7: Accuracy performance comparison of different decoding methods with varying levels of noise, evaluated on the NQ and HotpotQA datasets using the Mistral-7B. **Bold** values indicate the best performance, while underlined values represent the second-best.



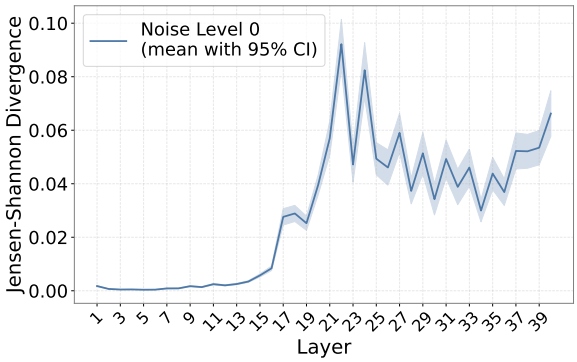
(a) SimHidden (Smaller is better) on NQ dataset.



(b) DiffAttn (Larger is better) on NQ dataset.



(c) SimHidden (Smaller is better) on HotpotQA dataset.



(d) DiffAttn (Larger is better) on HotpotQA dataset.

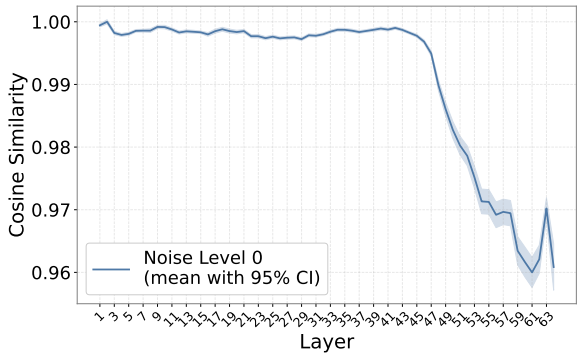
Figure 9: (a) Average SimHidden scores (with 95% confidence intervals) across layers when noise level = 0; (b) Average DiffAttn scores (with 95% confidence intervals) across layers when noise level = 0. Results are from Llama2-13B on NQ dataset and HotpotQA dataset.

		NQ	RGB	HotpotQA			2WikiMultihopQA				
				Compare	Bridge	Total	Compare	Bridge	Inf	Compose	Total
DeepSeek-7B	GD (0)	0.5250	0.8233	0.3282	0.4033	0.3883	0.2625	0.2414	0.2663	0.2609	0.2609
	LFD[0, 15)	0.3865	0.5000	0.2313	0.3366	0.3154	0.2155	0.2316	0.2334	0.2152	0.2211
	LFD[0, 30)	0.5389	0.8233	0.4768	0.4395	0.4470	0.4432	**0.4277**	0.4135	0.4159	0.4248
	LFD[15, 30)	**0.5412**	**0.8267**	**0.4801**	**0.4466**	**0.4533**	**0.4492**	0.4227	**0.4194**	**0.4223**	**0.4285**
Qwen3-8B	GD (0)	0.7318	0.9571	0.6960	0.6708	0.6759	0.6637	0.6335	0.5897	0.6085	0.6250
	LFD[0, 18)	0.7304	0.9567	0.7068	0.6845	0.6889	0.6640	0.6334	0.5903	0.6097	0.6256
	LFD[0, 36)	0.7372	0.9567	**0.7196**	0.6949	0.6999	0.6660	**0.6349**	0.6021	0.6146	0.6299
	LFD[18, 36)	**0.7380**	**0.9600**	0.7182	**0.6974**	**0.7016**	**0.6663**	0.6342	**0.6034**	**0.6148**	**0.6301**

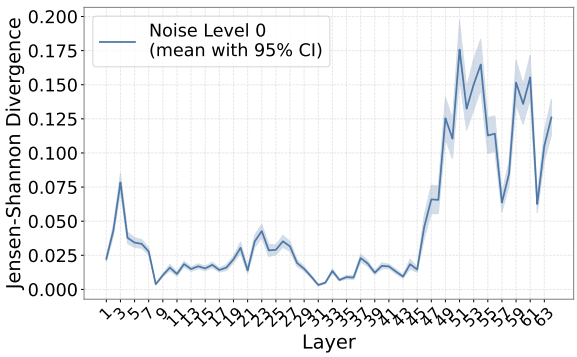
Table 8: Comparison of accuracy between different layer selection ranges under dynamic layer selection strategy. **Bold** indicate the best performance, while underline represent the second-best.

Method	GD(0)	CS	DoLA	LFD
Prompt	You are given a question and you MUST respond by EXTRACTING the answer from one of the provided documents. If none of the documents contain ... Document [20971901](Title: Pirates of the Caribbean (film series)) The film series started in 2003 with Pirates of the Caribbean : The Curse of the Black Pearl ... The franchise 's second film , subtitled Dead Man 's Chest , was released three years later in 2006 ... The third film in the series , subtitled At World 's End , ... and Disney released a fourth film , subtitled On Stranger Tides , in 2011 ... Question: All the names of the pirate of the caribbean movies. Answer:			
Ground Truth	[At World 's End, Dead Man 's Chest, On Stranger Tides]			
Answer	Captain Jack Sparrow, Will Turner, Elizabeth Swann, James Norrington, Barbossa, and Blackbeard.	Captain Jack Sparrow, Barbossa, Will Turner, Elizabeth Swann	Captain Jack Sparrow, Will Turner, Elizabeth Swann, James Norrington, Barbossa, and Blackbeard.	**Pirates of the Caribbean : The Curse of the Black Pearl, Dead Man's Chest, At World's End, and On Stranger Tides.**

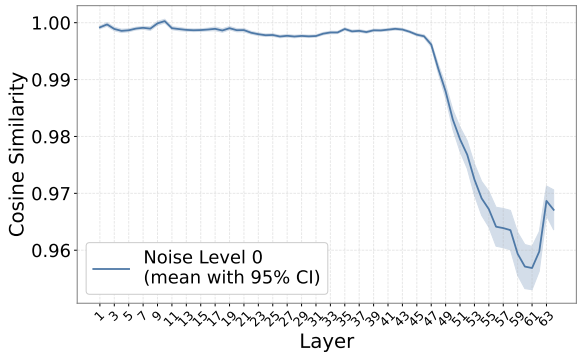
Table 9: Case study on GD(0), CS, DoLA, and LFD on the NQ dataset using LLaMA2-7B.



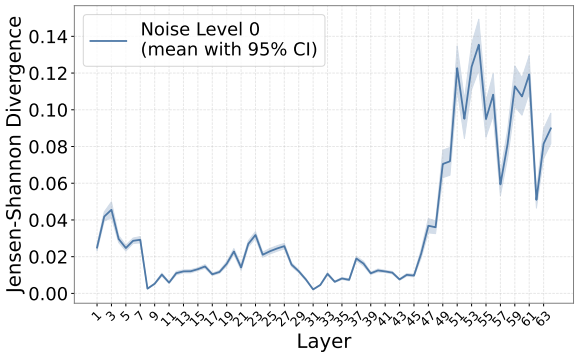
(a) SimHidden (Smaller is better) on NQ dataset.



(b) DiffAttn (Larger is better) on NQ dataset.



(c) SimHidden (Smaller is better) on HotpotQA dataset.



(d) DiffAttn (Larger is better) on HotpotQA dataset.

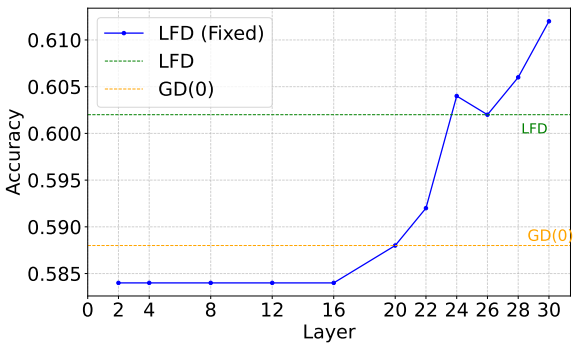
Figure 10: (a) Average SimHidden scores (with 95% confidence intervals) across layers when noise level = 0; (b) Average DiffAttn scores (with 95% confidence intervals) across layers when noise level = 0. Results are from Qwen3-32B on NQ dataset and HotpotQA dataset.

Method	GD(0)	CS	DoLA	LFD
Prompt	You are given a question and you MUST respond by EXTRACTING the answer from one of the provided documents. If none of the documents contain ... Document: Riot Games Singapore ... In October 2022, Riot acquired Wargaming Sydney—a subsidiary of Cyprus-based Wargaming that had originally developed the MMO middleware BigWorld—for an undisclosed amount, and renamed it Riot Sydney ... Question: What gaming software development studio did Riot Games acquire? Answer:			
Ground Truth	Wargaming Sydney			
Answer	Riot Sydney	RIOT SYDNEY	Riot Sydney	**Riot Games acquired Wargaming Sydney—a subsidiary of Cyprus-based Wargaming that had originally developed the MMO middleware BigWorld—for an undisclosed amount, and renamed it Riot Sydney.**

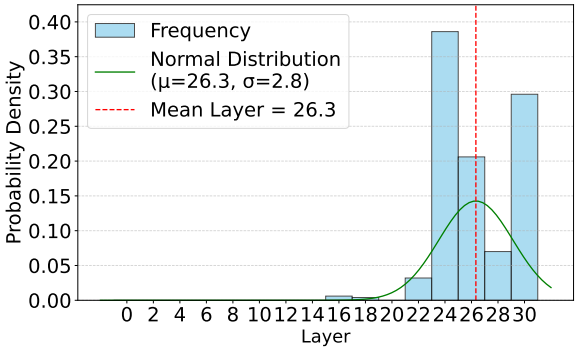
Table 10: Case study on GD(0), CS, DoLA, and LFD on the RGB dataset using LLaMA2-7B.

Method	GD(0)	CS	DoLA	LFD
Prompt	You are given a question and you MUST respond by EXTRACTING or DERIVING the answer from the provided documents. If the answer cannot be logically inferred from the documents, respond with NO-RES ... Document(Title: Terry Norris (actor)) ... As an actor, he has starred in TV Shows such as "Bellbird" & "Cop Shop", and in films like "Romulus, My Father" and "Paper Planes" ... Document(Title: Romulus, My Father (film)) ... Romulus, My Father is a 2007 Australian drama film directed by Richard Roxburgh ... Based on the memoir by Raimond Gaita, the film tells the story of Romulus (Eric Bana) and his wife Christine (Franka Potente), and their struggle in the face of great adversity to raise their son, Raimond (Kodi Smit-McPhee) ... Question: Terence Richard "Terry" Norris starred in a 2007 Australian drama film that was directed by Richard Roxburgh, and was based on a memoir by who? Answer:			
Ground Truth	Raimond Gaita			
Answer	ERIC BANA	ERIC BANA	NO-RES	**Based on the documents provided, Terry Norris starred in the film "Romulus, My Father" directed by Richard Roxburgh, based on the memoir by Raimond Gaita.**

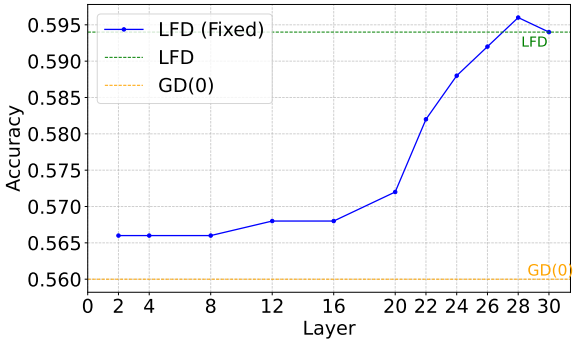
Table 11: Case study on GD(0), CS, DoLA, and LFD on the HotpotQA dataset using LLaMA2-7B.



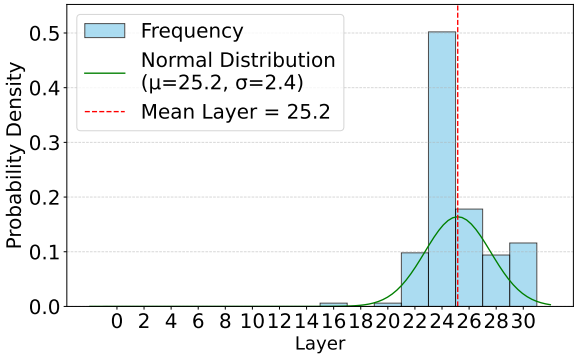
(a) LFD vs. LFD (Fixed) on NQ dataset.



(b) Histgram of layer selection in LFD on NQ dataset.



(c) LFD vs. LFD (Fixed) on HotpotQA dataset.

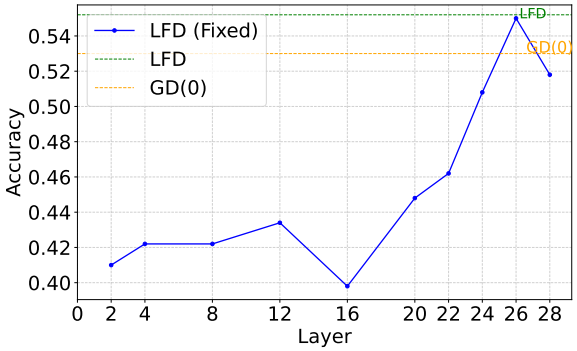


(d) Histgram of layer selection in LFD on HotpotQA dataset.

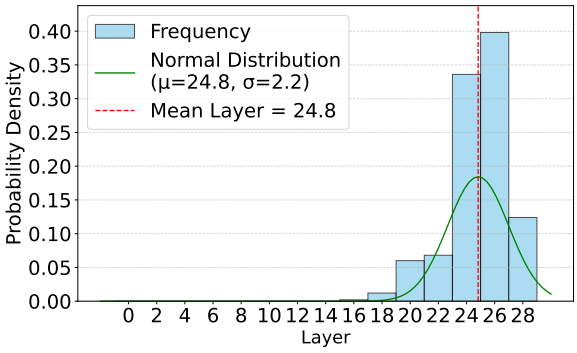
Figure 11: Comparison between LFD and LFD (Fixed) using the Mistral-7B model on the NQ and HotpotQA datasets.

Method	GD(0)	CS	DoLA	LFD
Prompt	You are given a question and you MUST respond by EXTRACTING or DERIVING the answer from the provided documents. If the answer cannot be logically inferred from the documents, respond with NO-RES ... Document(Title: A Tale of Winter) A Tale of Winter is a 1992 French drama film directed by Éric Rohmer, and starring Charlotte Véry, Frédéric van den Driessche and Michael Voletti ... Document(Title: Éric Rohmer) ... He edited the influential film journal, "Cahiers du cinéma", from 1957 to 1963 ... Question: Where does the director of film A Tale Of Winter work at? Answer:			
Ground Truth	Cahiers du cinéma			
Answer	Éric Rohmer works at a film production company.	Éric Rohmer	NO-RES: The answer cannot be logically inferred from the provided documents.	**Éric Rohmer works at a film journal called "Cahiers du cinéma".**

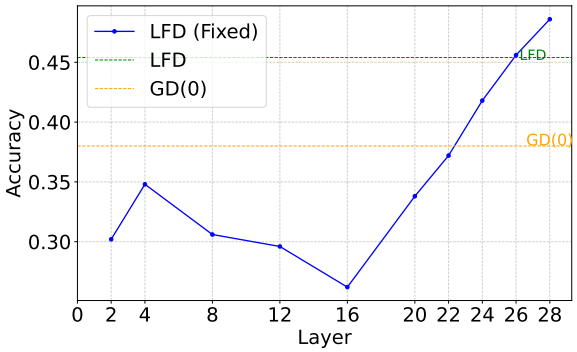
Table 12: Case study on GD(0), CS, DoLA, and LFD on the 2WikiMultihopQA dataset using LLaMA2-7B.



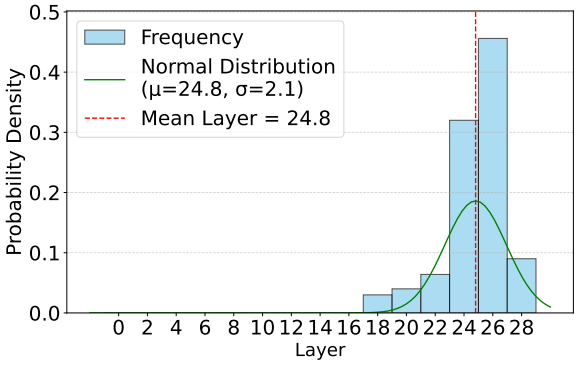
(a) LFD vs. LFD (Fixed) on NQ dataset.



(b) Histgram of layer selection in LFD on NQ dataset.

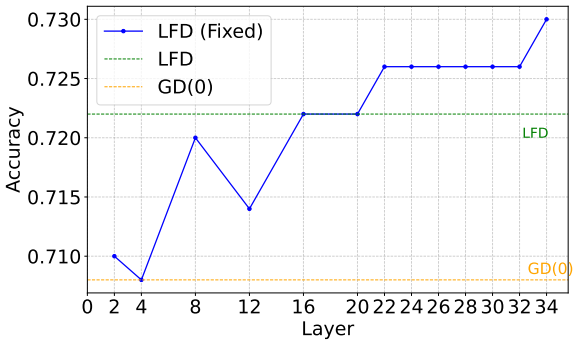


(c) LFD vs. LFD (Fixed) on HotpotQA dataset.

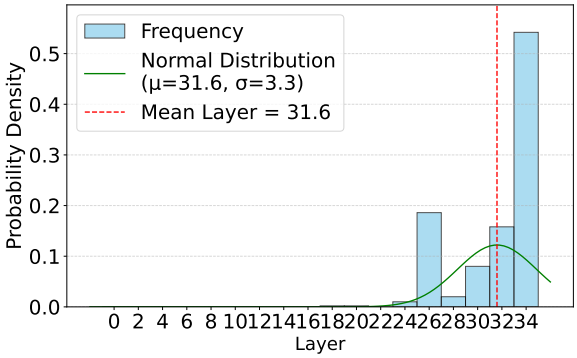


(d) Histgram of layer selection in LFD on HotpotQA dataset.

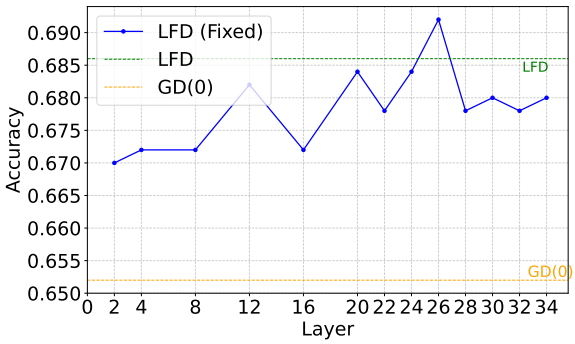
Figure 12: Comparison between LFD and LFD (Fixed) using the DeepSeek-7B model on the NQ and HotpotQA datasets.



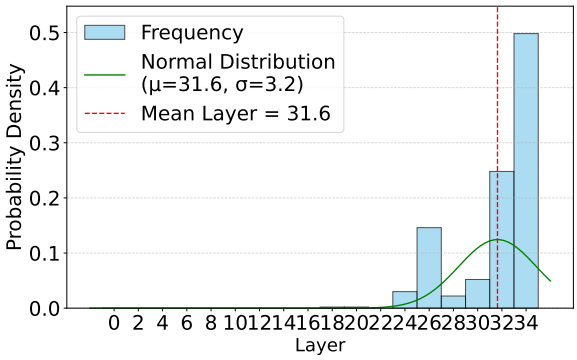
(a) LFD vs. LFD (Fixed) on NQ dataset.



(b) Histgram of layer selection in LFD on NQ dataset.



(c) LFD vs. LFD (Fixed) on HotpotQA dataset.



(d) Histgram of layer selection in LFD on HotpotQA dataset.

Figure 13: Comparison between LFD and LFD (Fixed) using the Qwen3-8B model on the NQ and HotpotQA datasets.