

Demystifying the Roles of LLM Layers in Retrieval, Knowledge, and Reasoning

Xinyuan Song¹, Keyu Wang², Pengxiang Li³, Lu Yin⁴, Shiwei Liu^{5,6}

¹Emory University, USA

²University of Tuebingen, Netherland

³Hong Kong Polytechnic University, Hong Kong

⁴University of Surrey, UK

⁵Max Planck Institute for Intelligent Systems,

⁶ELLIS Institute Tübingen,

Abstract

Recent studies suggest that the deeper layers of Large Language Models (LLMs) contribute little to representation learning and can often be removed without significant performance loss. However, such claims are typically drawn from narrow evaluations and may overlook important aspects of model behavior. In this work, we present a systematic study of depth utilization across diverse dimensions, including evaluation protocols, task categories, and model architectures. Our analysis confirms that very deep layers are generally less effective than earlier ones, but their contributions vary substantially with the evaluation setting. Under likelihood-based metrics without generation, pruning most layers preserves performance, with only the initial few being critical. By contrast, generation-based evaluation uncovers indispensable roles for middle and deeper layers in enabling reasoning and maintaining long-range coherence. We further find that knowledge and retrieval are concentrated in shallow components, whereas reasoning accuracy relies heavily on deeper layers—yet can be reshaped through distillation. These results highlight that depth usage in LLMs is highly heterogeneous and context-dependent, underscoring the need for task-, metric-, and model-aware perspectives in both interpreting and compressing large models.

1 Introduction

Deep neural networks have long exhibited depth-related challenges (Hanin, 2023), including vanishing or exploding gradients (Hochreiter, 1991), rank collapse (Bhojanapalli et al., 2020), and representational redundancy (Raghu et al., 2017). These effects reduce the marginal contribution of later layers during training and inference, even when optimization succeeds at lowering the loss (Dong et al., 2021). Large language models (LLMs) amplify these issues by pushing depth and width to the extreme: they contain billions of parameters and tens to hundreds of Transformer blocks (Brown et al., 2020). Such scale increases sensitivity to variance growth (Takase et al., 2025) and weak inter-layer signal propagation (Dong et al., 2021), making the effectiveness of very deep layers a central concern. A recent study, **Curse of Depth** (Sun et al., 2025), reports that deeper layers in modern LLMs can become ineffective due to variance explosion. Other analyses indicate that deeper layers may suffer from rank collapse or excessive feature overlap, limiting their ability to produce new, useful representations (Gromov et al., 2024; Li et al., 2025; Men et al., 2024).

This phenomenon presents both opportunities and challenges. On the opportunity side, deeper layers can be compressed more aggressively (Dumitru et al., 2024; Lu et al., 2024), and in some cases entire deep layers can be pruned without compromising performance (Muralidharan et al., 2024; Siddiqui et al., 2024). However, such claims are often based on narrow evaluations that may overlook important aspects of model behavior. Conventional benchmarks typically emphasize a limited set of tasks or metrics, which fail to capture the full range of capabilities that deeper layers might provide (Skean et al., 2025; Srivastava et al., 2022). In

more complex scenarios—such as knowledge-intensive reasoning (Petroni et al., 2021), mathematical problem solving (Cobbe et al., 2021b; Srivastava et al., 2022), or retrieval-augmented tasks (Lewis et al., 2020)—these layers may play a much more significant role. In light of these concerns, a key challenge remains unresolved: *How can we rigorously evaluate the contribution of each layer to overall LLM performance under different evaluation protocols?*

Building on these concerns, we present a systematic study of depth utilization across evaluation protocols, task categories, and model architectures. We first examine how different **evaluation protocols** affect conclusions about layer-wise importance, considering three protocols: *log-likelihood default* (Hendrycks et al., 2021), *log-likelihood continuation* (Paperno et al., 2016), and *generation until* (Gao et al., 2024; Liang et al., 2022). For each, we prune layers and measure performance degradation to quantify their contribution. Our analysis yields three **key findings**: (1) the perceived importance of layers is highly heterogeneous across evaluation metrics, underscoring the need for more principled and task-aware evaluation methodologies. (2) likelihood-based metrics emphasize shallow representations, with degradation concentrated in early layers; (3) generation exposes vulnerabilities in middle and deep layers, highlighting their role in reasoning and coherence; and

We next examine **knowledge and retrieval tasks** (Amini et al., 2019; Zellers et al., 2019) to probe how different functional demands shape layer usage. These tasks range from commonsense reasoning, which emphasizes shallow contextual plausibility, to retrieval-oriented settings (Gu and Dao, 2023; Mihaylov et al., 2018) that require accessing stored or external information. Through layer pruning, we identify three **key findings**: (1) shallow layers are critical, with their removal causing sharp degradation while deeper layers contribute less directly; (2) retrieval augmented with external evidence extends robustness into middle and deeper layers; and (3) knowledge and retrieval abilities are not uniformly distributed but localized to specific layers and even individual heads, highlighting both opportunities for targeted compression and challenges in preserving task fidelity.

Finally, we study **reasoning tasks and distilled models** (Mathematical Association of America, 1983; OpenAI, 2023; Rein et al., 2023). On benchmarks such as GSM8k (Cobbe et al., 2021a), reasoning accuracy depends strongly on middle and deep layers, with specific reasoning heads identified as critical. Distillation not only improves baseline accuracy but also redistributes reasoning capacity more evenly across depth, thereby enhancing robustness under ablation. These findings highlight how reasoning differs from knowledge and retrieval tasks and how distillation reshapes depth dependence in LLMs.

Our extensive controlled experiments across diverse tasks, evaluation metrics, and model families consistently show that the contribution of layers is highly non-uniform. Different tasks activate different depths, different metrics emphasize different subsets of layers, and different model designs further modulate these effects. These findings demonstrate that understanding depth usage in LLMs requires a **task-, metric-, and model-aware perspective**, which is essential to avoid experimental bias and to ensure that conclusions about layer importance are both reliable and broadly applicable.

2 Evaluation Protocol Matters

In this section, we examine how various evaluation protocols influence the assessment of layer pruning. Our goal is to understand whether likelihood-based metrics and generation-based metrics reveal consistent or divergent patterns of layer importance.

We evaluate the impact of different evaluation metrics under layer pruning using two representative LLMs, **LLaMA-3.1-8B** (Touvron et al., 2023) and **Qwen3-8B** (Bai et al., 2023). All experiments are conducted on the **MMLU benchmark** (Hendrycks et al., 2021), a widely adopted testbed for general knowledge and reasoning. We consider **log-likelihood default**, which scores multiple-choice answers based on option likelihood; **log-likelihood continuation**, which evaluates token-level continuation probability; and **generation until**, which measures open-ended autoregressive generation until the final answer. For each setting, we systematically ablate layers and measure performance degradation as a function of layer index, reporting both accuracy (μ) and relative change ($\Delta\mu$, defined as the difference between the full model and the layer-pruned model).

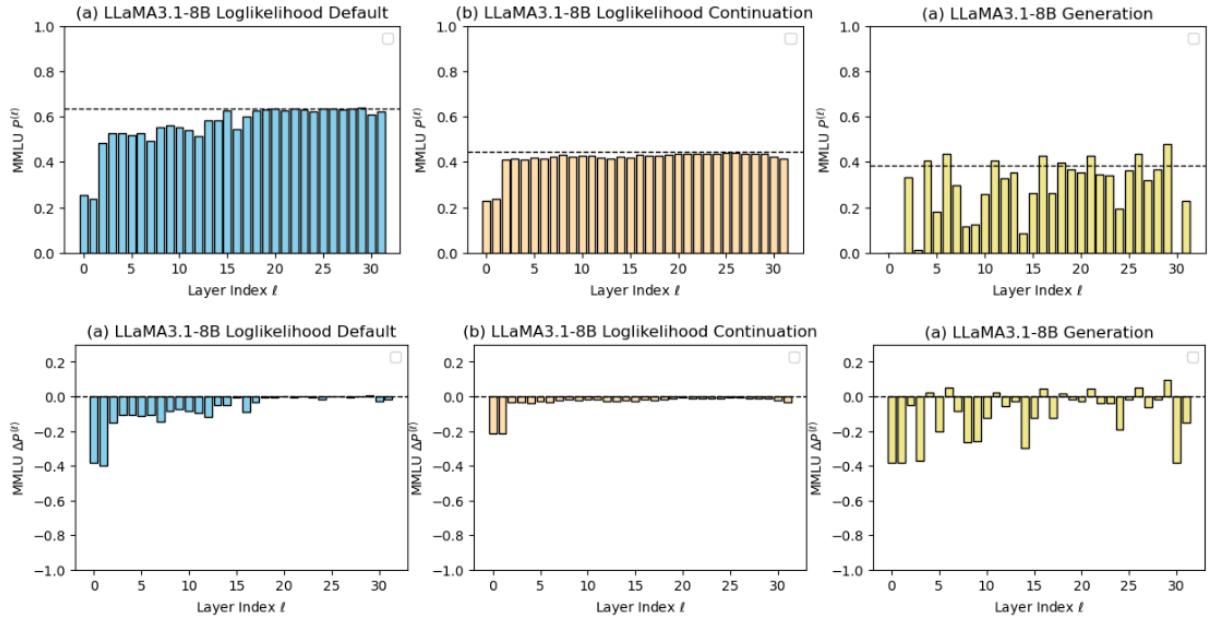


Figure 1: Layer pruning results of **LLaMA-3.1-8B** on the **MMLU** benchmark under three evaluation protocols: log-likelihood default (left), log-likelihood continuation (middle), and generation until (right). We report accuracy (μ) and relative change ($\Delta\mu$) across layers.

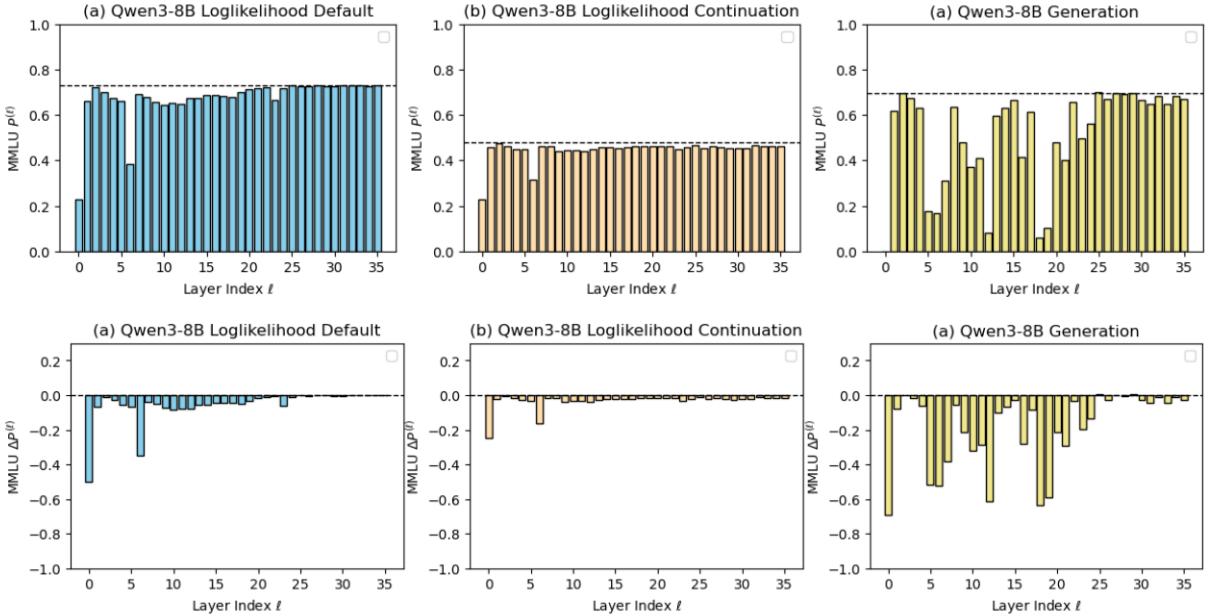


Figure 2: Layer pruning results of **Qwen3-8B** on the **MMLU** benchmark under the same three evaluation protocols. Accuracy (μ) and relative change ($\Delta\mu$) are shown across layer indices.

Results in Figures 1 and 2 show that in log-likelihood-based evaluations (multiple-choice or continuation), degradation is concentrated in the earliest layers, suggesting reliance on shallow representations to maintain

token-level coherence. In contrast, the generation-based evaluation reveals substantial drops across middle and deeper layers, indicating that multi-step reasoning and long-range consistency require contributions from the entire depth of the network. Together, these results demonstrate that likelihood-based evaluations substantially underestimate the fragility of compressed models, whereas generation more faithfully captures the dependence of LLMs on hierarchical depth.

3 Layer Importance in Knowledge Tasks

In this section, we analyze how layer pruning affects performance on knowledge-intensive tasks. We focus on commonsense reasoning benchmarks to identify whether shallow or deep layers contribute more critically to accuracy.

3.1 commonsense dataset

We evaluate **LLaMA-3.1-8B** on the **HellaSwag** (Zellers et al., 2019) commonsense dataset under a **layer pruning** setting. Three evaluation metrics are considered: the standard multiple-choice accuracy (**acc**), a cross-entropy-based accuracy (**acc_ce**), based on the **log-likelihood default** evaluation protocol as described in Section 2. For each metric, we report both the absolute performance and the relative difference compared to the unablated model as functions of layer index. The results are summarized in Figure 3.

As shown in the figure, ablating early layers leads to substantial performance degradation, with accuracy drops up to -0.3 in **acc** and -0.5 in **acc_ce**. In contrast, middle and later layers exhibit negligible changes, and in some cases even slight improvements. These findings suggest that commonsense continuation tasks such as HellaSwag rely heavily on shallow representations, while deeper layers play a less critical role in maintaining accuracy.

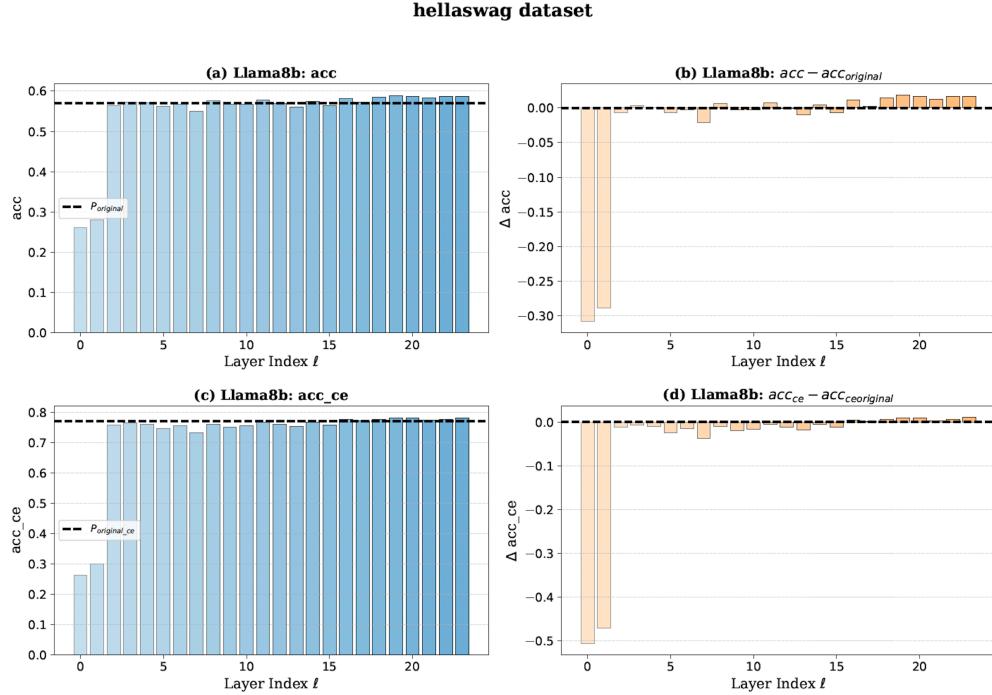


Figure 3: layer pruning results of **LLaMA-3.1-8B** on the **HellaSwag** dataset. We report both standard accuracy (**acc**) and cross-entropy-based accuracy (**acc_ce**), along with their relative differences compared to the unablated model across layers.

3.2 Math Problem Solving

We also evaluate **LLaMA-3.1-8B** on the **MathQA** dataset (Amini et al., 2019), a large-scale benchmark for mathematical problem solving. Our goal is to examine whether reasoning in math benchmarks exhibits broader or different sensitivity to layer pruning compared to commonsense tasks.

As in the HellaSwag evaluation, we report both the standard multiple-choice accuracy (**acc**) and the cross-entropy-based accuracy (**acc_ce**), analyzing their absolute values and relative deviations with respect to the unablated model. The results are shown in Figure 4.

In contrast to **HellaSwag**, **MathQA** exhibits broader sensitivity to layer pruning across the network. Removing early layers leads to moderate drops, with reductions up to -0.06 in **acc** and -0.08 in **acc_ce**, while degradations also persist into the middle layers. This pattern indicates that mathematical reasoning tasks require both shallow and intermediate representations, reflecting a cumulative reliance on symbolic manipulation and semantic integration distributed across model depth.

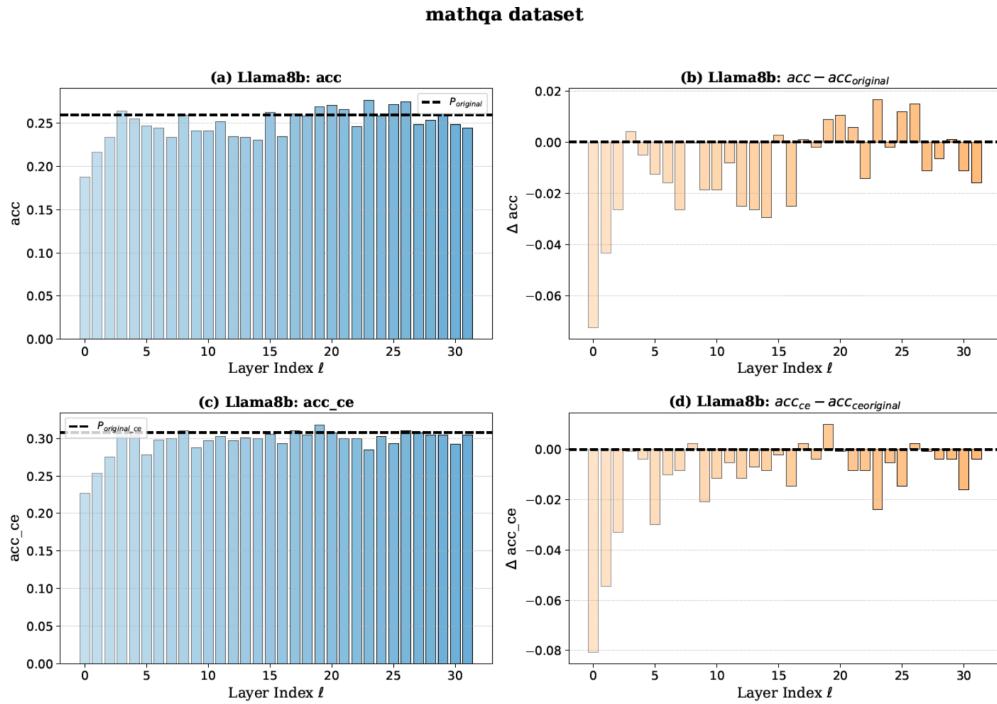


Figure 4: layer pruning results of **LLaMA-3.1-8B** on the **MathQA** dataset.

4 Layer Importance in Retrieval Tasks

A key question in understanding large language models is how retrieval ability is distributed across depth. To investigate this, we evaluate the effect of layer pruning on retrieval-oriented tasks.

4.1 KV retrieval

We evaluate **LLaMA-3.1-8B** on the **KV Retrieval task** (Bick et al., 2025) under layer pruning. The task requires retrieving the correct key-value pair from stored memory. We report accuracy (μ) and relative change ($\Delta\mu$), based on the **log-likelihood default** evaluation protocol across layers.

See Figure 5. Results show that shallow layers are critical for retrieval, as ablations in the first few layers cause sharp accuracy drops (up to -0.8). Beyond the lower layers, the model maintains near-perfect retrieval

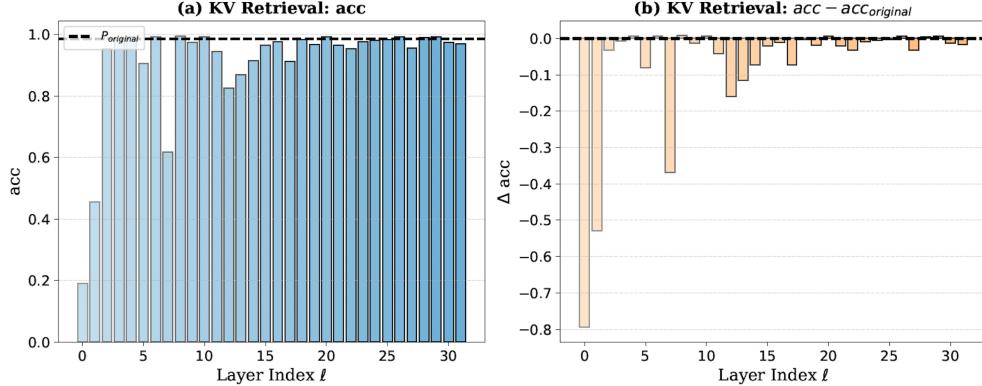


Figure 5: layer pruning results of **LLaMA-3.1-8B** on the **KV Retrieval task**. (a) accuracy μ (blue), (b) $\Delta\mu$ (yellow).

performance, and $\Delta\mu$ remains close to zero. This indicates that retrieval ability is encoded primarily in early representations, while middle and deeper layers contribute less to this task.

4.2 Retrieval Augmentation

Building on the analysis of **KV Retrieval**, which highlights the role of shallow layers in memorization-based retrieval, we next examine how retrieval compares to non-retrieval settings in question answering tasks. Specifically, we investigate the effect of retrieval augmentation on **LLaMA-3.1-8B** by performing layer pruning experiments on the **OpenBookQA** and **CloseBookQA** benchmarks (Mihaylov et al., 2018) (Figure 6). In this setting, **OpenBookQA** serves as the retrieval-augmented mode, where the model incorporates external evidence before answering, while **CloseBookQA** serves as the non-retrieval baseline, where the model answers questions directly without access to retrieval. We report both standard accuracy (**acc**) and cross-entropy-based accuracy (**acc_ce**) as evaluation metrics across layers.

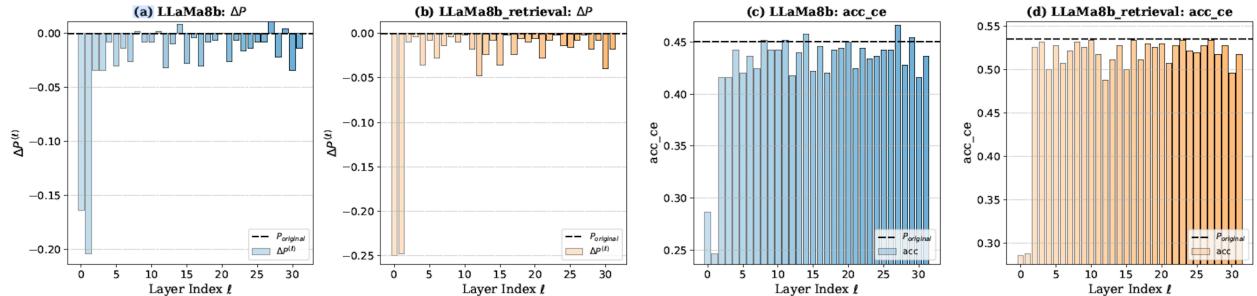


Figure 6: Layer pruning results of **LLaMA-3.1-8B** on the **OpenBookQA** benchmark under non-retrieval and retrieval-augmented settings. Blue curves denote standard accuracy (**acc** μ and $\Delta\mu$), while yellow curves denote cross-entropy-based accuracy (**acc_ce**).

Results show that retrieval consistently improves robustness across almost all layers, with the largest benefits appearing in middle and deeper layers. While ablation degrades both settings, the retrieval-augmented model maintains higher accuracy and exhibits smaller performance drops. This indicates that retrieval not only boosts baseline accuracy but also enhances stability against layer pruning.

4.3 LLaMA-1 Results

Since our experiments on **LLaMA-3.1-8B** did not reveal a clear dependence of retrieval tasks on intermediate layers, we hypothesize that the strength of the larger model may obscure such effects. To further probe this question, we turn to a simpler model, **LLaMA-1 7B** (Touvron et al., 2023), and perform layer pruning on the KV Retrieval task using the same evaluation metrics—accuracy (μ) and relative change ($\Delta\mu$) based on the **log-likelihood default** evaluation protocol.

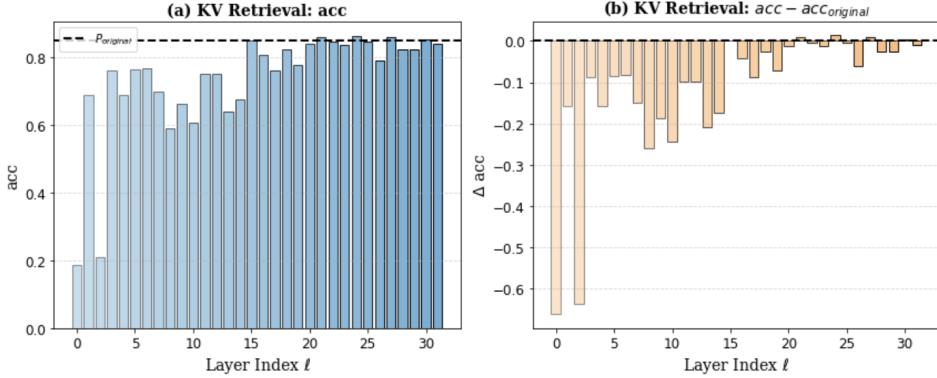


Figure 7: **LLaMA-1 7B** on the KV Retrieval task under layer pruning. (a) accuracy μ (blue), (b) $\Delta\mu$ (yellow).

For **KV Retrieval** (Figure 7), performance remains near-perfect once the lower layers are preserved, but pruning within the first two layers leads to sharp accuracy drops (up to -0.6). Unlike the results observed on **LLaMA-3.1-8B**, where accuracy stabilizes quickly after shallow layers, the **LLaMA-1 7B** model also shows noticeable degradations in some middle layers. This contrast suggests that the impact of layer pruning on retrieval is model-dependent, and smaller models may distribute retrieval capacity less compactly across depth compared to larger models.

5 Retrieval Head

To further probe the phenomenon observed in Section 4.3 and Figure 7, where **LLaMA-1 7B** shows pronounced accuracy drops at layers 8 and 10 on the **KV Retrieval task**, we conduct head pruning on these two layers. The goal is to localize the source of degradation and identify specific attention heads that dominate retrieval performance.

Head pruning experiments reveal that retrieval ability is concentrated in specific attention heads rather than being uniformly distributed. At layer 8, several heads prove to be critical, with individual ablations causing notable performance drops ($\Delta\mu$ up to -0.05), indicating that retrieval is strongly tied to particular heads in this shallow layer (Figure 8). At layer 10, a similar but weaker pattern emerges, as some heads disproportionately affect performance, though the magnitude of degradation is smaller compared to layer 8 (Figure 9). These results highlight that retrieval depends on a sparse set of specialized heads in shallow and mid-level layers, making their identification crucial for both understanding retrieval mechanisms and maintaining performance under pruning or compression.

6 Layer Importance in Reasoning Tasks

Previous sections examined the role of depth in knowledge and retrieval tasks, here we investigate whether a distinct reasoning layer emerges and how its position within the network affects performance.

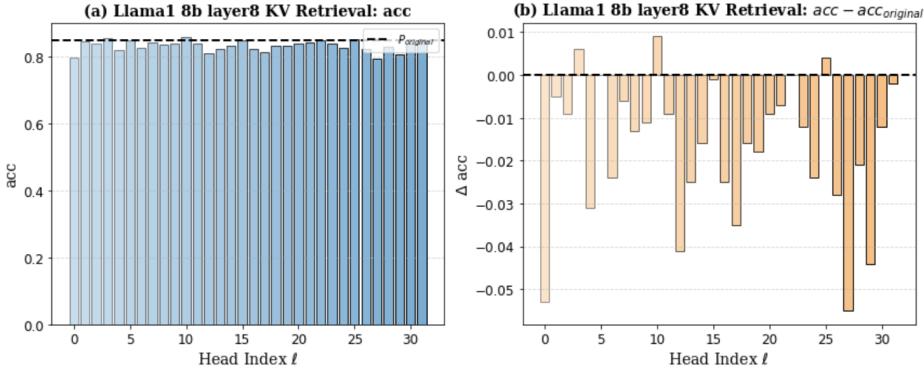


Figure 8: LLaMA-1 7B head pruning at layer 8 on KV Retrieval: (a) accuracy per head, (b) $\Delta\mu$ per head.

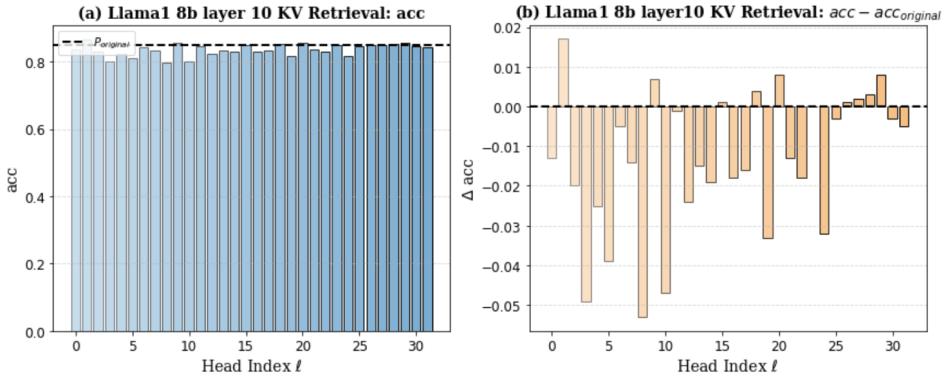


Figure 9: LLaMA-1 7B head pruning at layer 10 on KV Retrieval: (a) accuracy per head, (b) $\Delta\mu$ per head.

6.1 QWEN Result

We conduct **layer pruning** experiments on the **GSM8K 8-shot** (Cobbe et al., 2021a) benchmark using four models: **Qwen3-8B** (Bai et al., 2023), **Qwen3-8B without thinking mode**, and **LLaMA-3.1-8B** (Touvron et al., 2023). We selected this range of models to capture varying levels of chain-of-thought (CoT) (Wei et al., 2022) capability. Qwen3-8B generally exhibit stronger CoT performance, whereas LLaMA-3.1-8B and the Qwen3-8B variant with disabled thinking mode display comparatively weaker CoT. For each model, we measure performance degradation across layers, reporting accuracy (μ) and relative accuracy change ($\Delta\mu$) based on the **generate-until** evaluation protocol from Section 2.

Results show that reasoning performance is highly sensitive to middle and deep layers, with ablations in these regions causing sharp drops in GSM8K accuracy. The effect is consistent across different model families (Qwen vs. LLaMA), though models with explicit CoT training (e.g., Qwen3-8B) exhibit higher baseline robustness compared to those without CoT enhancement. This confirms that multi-step mathematical reasoning relies more heavily on deeper hierarchical layers than shallow continuation tasks.

6.2 Qwen Multi-shot Result

To further examine the presence of a **reasoning layer**, we study **Qwen3-8B** under the **generate_until** setup. We design two tasks: a **1-shot** setting with a single in-context example and a **4-shot** setting with four exemplars. For each case, we conduct layer pruning experiments on GSM8K and report absolute accuracy

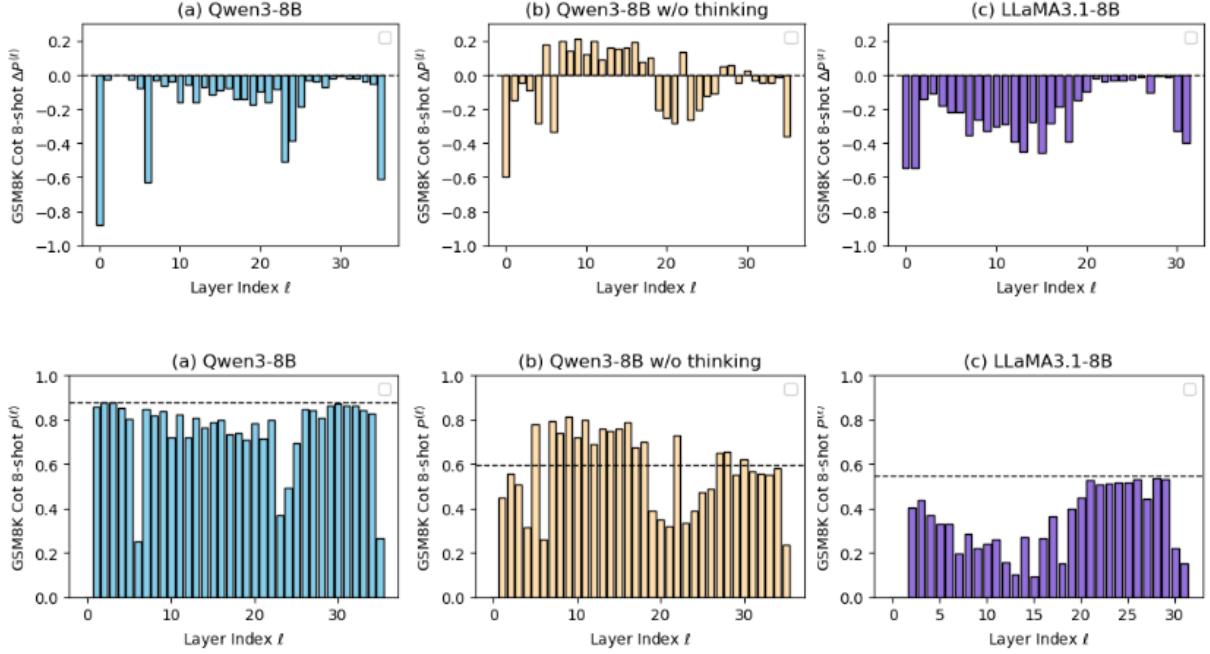


Figure 10: Layer pruning results on the **GSM8K 8-shot** benchmark under **Qwen3-8B**, **Qwen3-8B without thinking mode**, and **LLaMA-3.1-8B**.

(μ) and accuracy difference ($\Delta\mu$) based on the **generate-until** evaluation protocol.

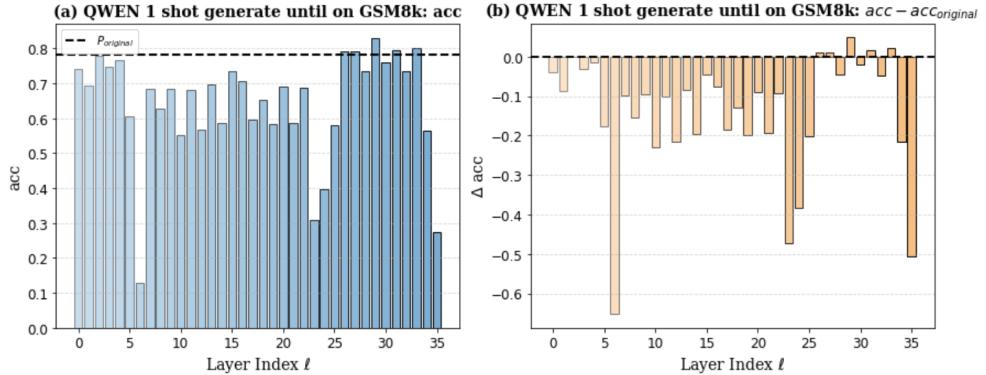


Figure 11: Qwen3-8B on GSM8K with **1-shot generate_until**: (a) accuracy μ (blue), (b) $\Delta\mu$ (yellow).

Results in Figure 11 and Figure 12 reveal similar patterns across the two prompting strategies. In the **4-shot** setup, baseline accuracy is consistently higher than in the **1-shot** case, reflecting the benefit of additional exemplars. However, both **1-shot** and **4-shot** exhibit sharp degradations when certain layers are pruned. In particular, several intermediate layers—such as layers 6, 23, and 35—show substantial drops, with $\Delta\mu$ reaching below -0.5 in the most affected regions. These findings suggest that while more exemplars improve overall performance, CoT-style reasoning remains highly dependent on specific mid-to-deep layer representations.

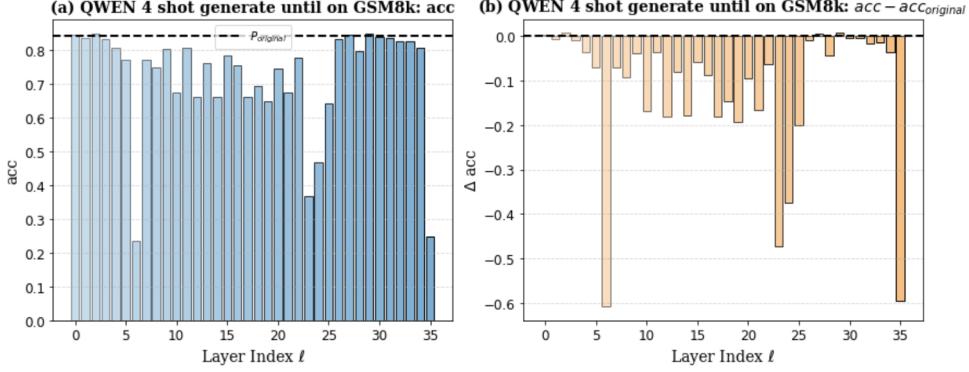


Figure 12: Qwen3-8B on GSM8K with 4-shot generate_until: (a) accuracy μ (blue), (b) $\Delta\mu$ (yellow).

6.3 Qwen Reasoning Head

Building on the layer pruning analysis, which showed that CoT-style reasoning in Qwen3-8B depends on specific intermediate and deep layers, we further investigate whether reasoning ability can be localized to individual attention heads. To this end, we conduct head pruning experiments on layer 35, one of the most critical layers identified in the GSM8K 1-shot CoT setting in Figure 11. For each head, we measure absolute accuracy (μ) and accuracy difference ($\Delta\mu$) based on the **generate-until** evaluation protocol.

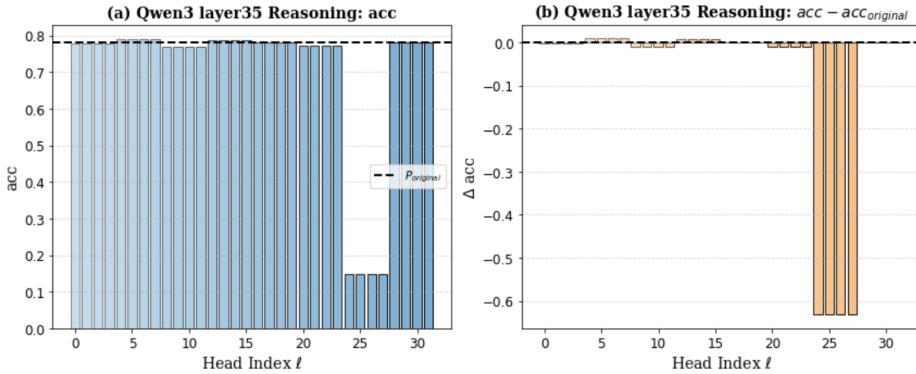


Figure 13: Qwen3-8B on GSM8K with Layer 35 head pruning: (a) accuracy μ (blue), (b) $\Delta\mu$ (yellow). The x-axis denotes the head index (1–32) within layer 35.

As shown in Figure 13, reasoning ability is highly concentrated in a sparse subset of heads. Pruning certain heads at layer 35 leads to severe degradation (with $\Delta acc \approx -0.6$), while most other heads cause little to no loss when ablated. This indicates that a small number of reasoning heads dominate the completion of multi-step reasoning in the final layer. Moreover, since **Qwen3-8B** employs Grouped Query Attention (Ainslie et al., 2023), the heads within the same group exhibit similar degradation patterns, further confirming that reasoning capacity is bottlenecked by specialized groups of attention heads in deep layers.

7 Distilled Reasoning Model

In this section, we examine whether distillation alters the depth distribution of reasoning ability. Our goal is to determine if distilled models concentrate reasoning in fewer layers or heads compared to their base counterparts.

In addition to **Qwen**, we also evaluate a distilled **LLaMA-3.1-8B** on GSM8K to probe the location of reasoning layers.

7.1 LLaMA-3.1-8B (distilled) on GSM8K under CoT

Distilled models retain strong reasoning ability. To locate where this ability resides in depth, we analyze **LLaMA-3.1-8B** and its distilled variant Deepseek model (DeepSeek-AI et al., 2025) on the **GSM8K** benchmark using chain-of-thought (CoT) prompting. We perform layer pruning and measure raw accuracy (**acc**) and accuracy difference relative to the model ($\Delta\mu$) as functions of the layer index.

Figure 14 shows the base model under CoT, and Figure 15 shows the distilled model under the same protocol. Across both models, reasoning performance is highly sensitive to shallow and middle layers, with $\Delta\mu$ drops reaching -0.6 or lower in the most affected regions. The distilled variant yields a higher baseline accuracy and slightly improved robustness in deeper layers, but it still exhibits notable vulnerability when early and mid-depth layers are ablated. These results indicate that, even after distillation, CoT-style reasoning depends on representations that form primarily in the shallow-to-mid depth range, with deeper layers contributing to stabilization rather than fully offsetting losses from earlier layers.

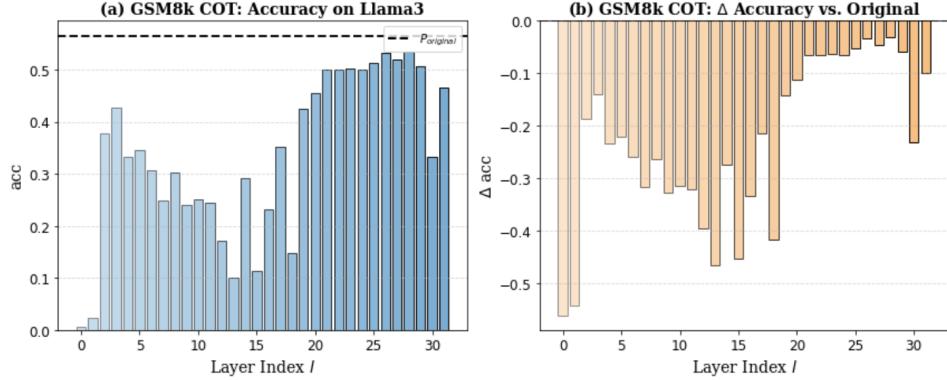


Figure 14: **LLaMA-3.1-8B** on GSM8K with **CoT prompting**: (a) accuracy, (b) Δacc , as functions of layer index.

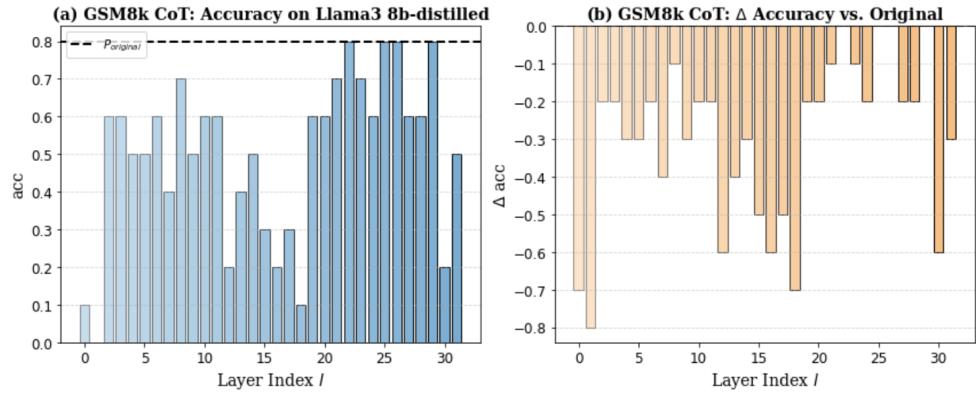


Figure 15: **LLaMA-3.1-8B (distilled)** on GSM8K with **CoT prompting**: (a) accuracy, (b) Δacc , as functions of layer index.

7.2 LLaMA-distilled Reasoning Head

Building on the layer pruning analysis in Figure 15, which revealed that reasoning in LLaMA-3.1-8B-distilled emerges in specific middle and deep layers, we further investigate whether this functionality can be localized to individual attention heads. To this end, we conduct head pruning experiments at three representative layers: layer 12 (early–middle depth), layer 18 (middle depth), and layer 30 (final layer). For each case, we measure absolute accuracy (μ) and accuracy difference relative to the unablated model ($\Delta\mu$).

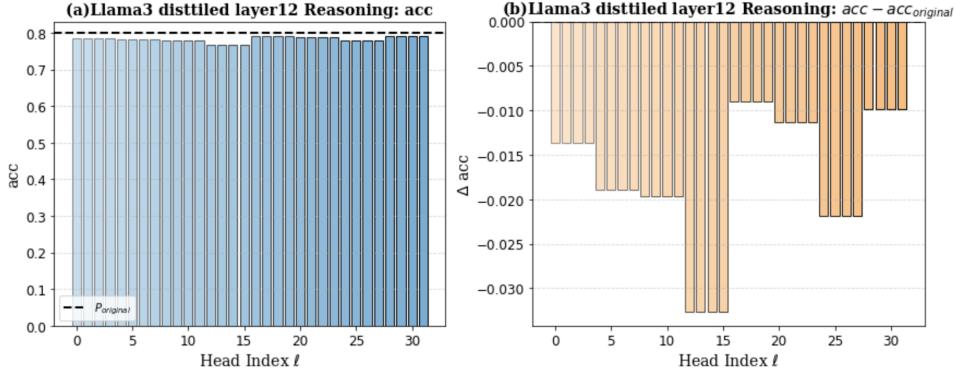


Figure 16: LLaMA-3.1-8B-distilled on GSM8K with **Layer 12 head pruning**: (a) accuracy μ per head, (b) $\Delta\mu$ per head.

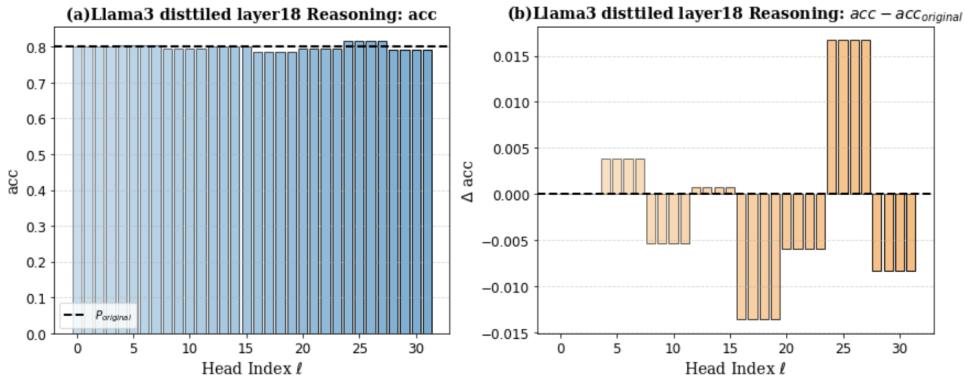


Figure 17: LLaMA-3.1-8B-distilled on GSM8K with **Layer 18 head pruning**: (a) accuracy μ per head, (b) $\Delta\mu$ per head.

The results show that reasoning ability in the distilled model is indeed localized to a sparse subset of heads across different depths. At layer 12, pruning certain heads leads to moderate degradation ($\Delta\mu \approx -0.03$), revealing early reasoning-sensitive heads. At layer 18, we observe mixed effects: some heads contribute positively, slightly boosting performance when pruned, while others cause drops ($\Delta\mu \approx -0.015$). At layer 30, pruning specific deep heads reduces performance ($\Delta\mu \approx -0.008$), suggesting that reasoning is consolidated by a few specialized heads in the final layer. Overall, these findings indicate that distilled reasoning models distribute reasoning functionality across multiple layers but still rely heavily on a small number of specialized attention heads.

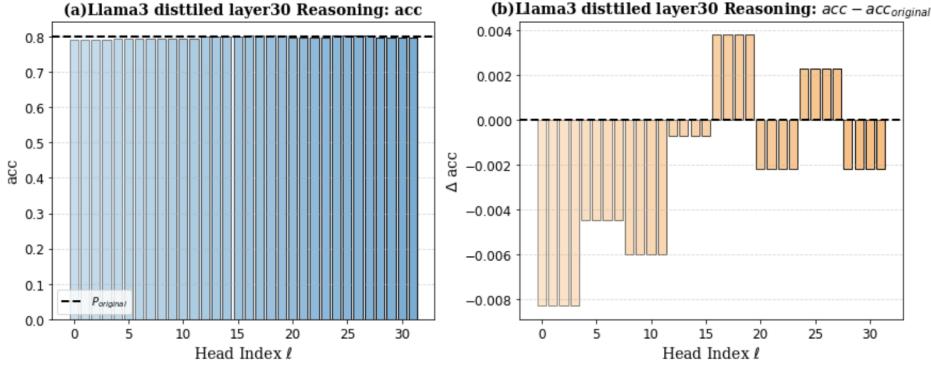


Figure 18: LLaMA-3.1-8B-distilled on GSM8K with **Layer 30 head pruning**: (a) accuracy per head, (b) $\Delta\mu$ per head.

8 Delta Model

In this section, we investigate how distillation effects can be isolated and analyzed through delta model replacement. Our goal is to determine whether improvements from distillation are concentrated in specific layers, and how replacing projections between models influences reasoning robustness.

8.1 Delta Model: Distillation-Base

We evaluate the effect of pruning (ΔW) combined with output projection W on the **GSM8K Chain-of-Thought (CoT)** benchmark, comparing the **DeepSeek-LLaMA3-distilled** model against the original **LLaMA-3.1** under layer pruning. The goal is to examine whether certain layers are crucial for distillation robustness.

Specifically, we adopt a delta model setting, where the output projection matrix of a given layer in one model is replaced by that of another model. Formally, let $W_{\text{src}}^{(l)}$ denote the output projection at layer l of the source model (e.g., DeepSeek-LLaMA3-distilled), and $W_{\text{tgt}}^{(l)}$ the corresponding matrix of the target model (e.g., LLaMA-3.1). The delta replacement is defined as:

$$W_{\text{tgt}}^{(l)} \leftarrow W_{\text{tgt}}^{(l)} + \Delta W^{(l)}, \quad \text{where} \quad \Delta W^{(l)} = W_{\text{src}}^{(l)} - W_{\text{tgt}}^{(l)}. \quad (1)$$

This operation effectively injects the representation learned by one model into the corresponding layer of the other, allowing us to isolate whether improvements from distillation are concentrated in specific layers.

For both models, we report (i) absolute accuracy and (ii) accuracy difference relative to the unablated model across all layers (Figure 19).

Results show that the **DeepSeek-distilled model** maintains high accuracy across almost all layers, with only mild fluctuations in $\Delta\mu$ and even small improvements in certain middle layers (Figure 19). When layers of **LLaMA-3.1** are replaced with those from the distilled model, we observe slight gains in the middle layers, suggesting that distilled representations at intermediate depths help enhance the reasoning ability of the base LLaMA model.

8.2 Delta Model: Reverse Replacement

We further evaluate **pruning (ΔW) combined with output projection W_o** under a **reverse replacement** setting, in which we progressively substitute each layer of the **LLaMA-3.1 distilled** model with the corresponding layer from the original **LLaMA-3.1-8B** (i.e., for layer l , $W_{o,\text{distilled}}^{(l)} \leftarrow W_{o,\text{base}}^{(l)}$). We run this procedure on GSM8K CoT and report absolute accuracy (μ) and accuracy difference ($\Delta\mu$) relative to the distilled baseline. The results are presented in Figure 20.

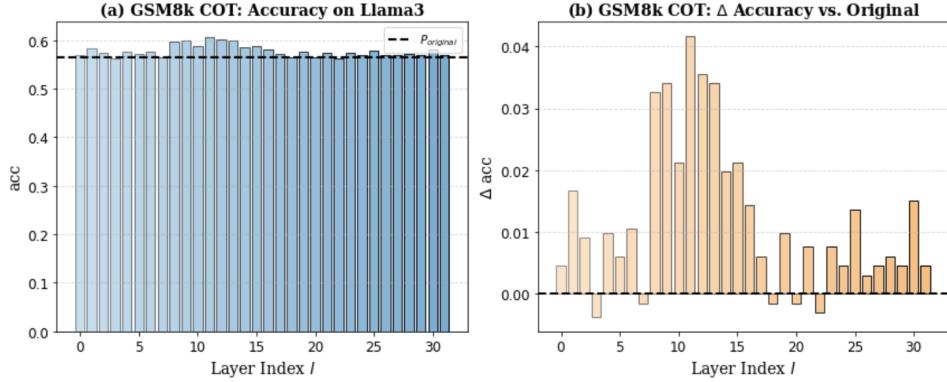


Figure 19: layer pruning results of **DeepSeek-LLaMA3-distilled** on GSM8k CoT. Left: absolute accuracy (μ). Right: relative accuracy difference ($\Delta\mu$) compared to the unablated model.

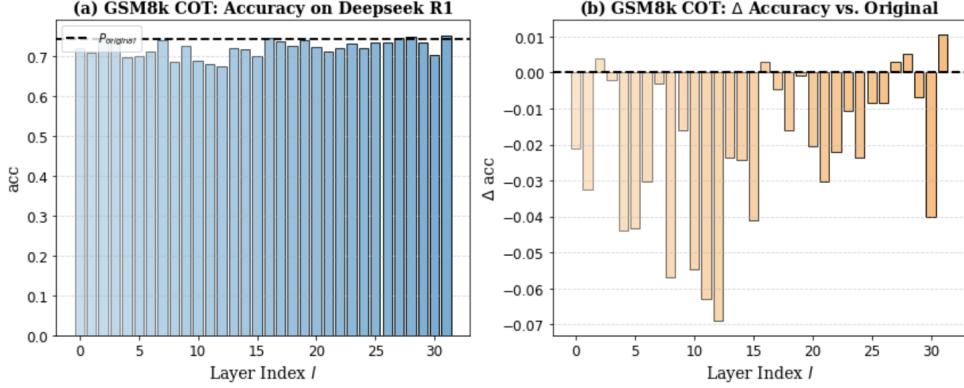


Figure 20: Reverse replacement on **LLaMA-3.1 distilled** for GSM8K CoT: for each layer l , the distilled layer’s output projection $W_o^{(l)}$ is replaced by the corresponding **LLaMA-3.1-8B** layer. Left: absolute accuracy (μ). Right: accuracy difference ($\Delta\mu$) relative to the distilled baseline. Lower $\Delta\mu$ indicates performance loss after substituting a distilled layer with a base layer.

Results show that **LLaMA-3.1 distilled** suffer performance degradation once its layers are replaced by those of the original **LLaMA-3.1**. In Figure 20, early and middle layers exhibit negative $\Delta\mu$ (up to -0.06), while later layers remain comparatively stable, indicating that replacing distilled representations with base LLaMA layers weakens reasoning robustness. These results demonstrate that distilled models encode reasoning capacity in a more robust layer distribution, and that reverse replacement with base model layers diminishes this advantage, confirming that distillation plays a key role in strengthening reasoning resilience against pruning.

8.3 Accumulative Layer Replacement Analysis

We further investigate the effect of **accumulative layer replacement** under pruning (ΔW) combined with output projection W_o on the **GSM8K Chain-of-Thought** (CoT) benchmark. In the first setting, we progressively replace layers of **DeepSeek R1** with those of **LLaMA-3.1-8B** (Figure 21), while in the second setting we reverse the process, gradually substituting layers of **LLaMA-3.1-8B** with those of **DeepSeek R1** (Figure 22).

Results show that in the **DeepSeek → LLaMA replacement** setting (Figure 21), accuracy initially

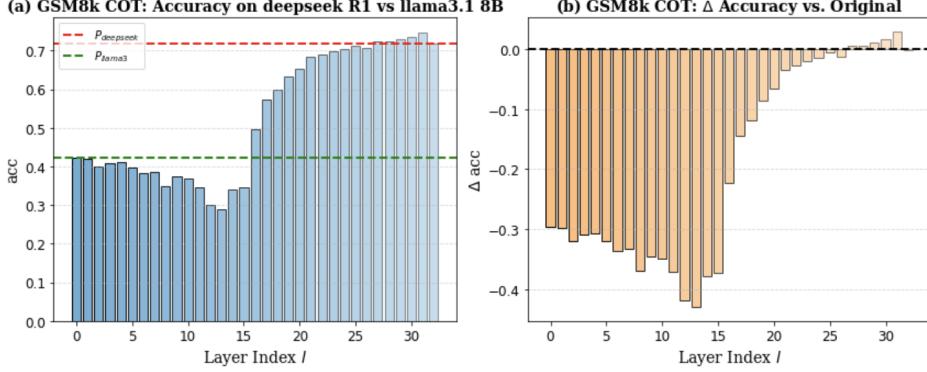


Figure 21: Accumulative layer replacement: **DeepSeek R1 replaced by LLaMA-3.1-8B** on GSM8k CoT. Left: absolute accuracy (μ). Right: relative accuracy difference ($\Delta \mu$) compared to the original DeepSeek R1 baseline.

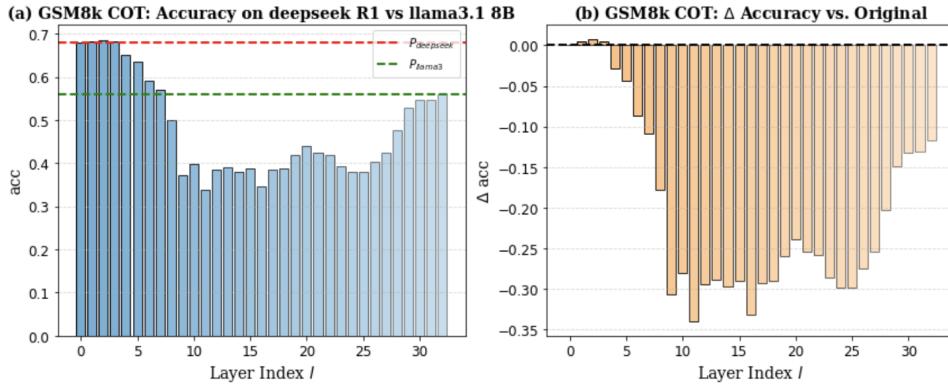


Figure 22: Accumulative layer replacement: **LLaMA-3.1-8B replaced by DeepSeek R1** on GSM8k CoT. Left: absolute accuracy (μ). Right: relative accuracy difference ($\Delta \mu$) compared to the original LLaMA-3.1-8B baseline.

drops sharply when the first few layers are replaced (up to -0.4), suggesting that shallow layers are critical for preserving DeepSeek’s distilled reasoning ability. However, as more layers are replaced, performance gradually converges to the LLaMA baseline. Conversely, in the **LLaMA → DeepSeek replacement** setting (Figure 22), we observe steady accuracy gains as more DeepSeek layers accumulate, indicating that distilled DeepSeek layers introduce robustness and reasoning improvements even when partially integrated.

Overall, these experiments highlight that early and middle layers are crucial for transferring distilled knowledge, while later layers can be replaced with relatively smaller impact. This supports the view that distillation redistributes reasoning capacity across the depth of the network but still depends critically on shallow-to-mid representations.

9 Conclusion

We systematically analyzed depth usage in large language models via layer pruning across tasks, metrics, and model families. Results show that layer contributions are highly uneven: shallow layers dominate likelihood and retrieval, while mid-to-deep layers are essential for reasoning and generation. Taken together, depth usage is inherently task-dependent, highly metric-sensitive, and strongly model-specific, highlighting the critical

need for task-aware evaluation and offering practical guidance for compression and future model design.

References

- Joshua Ainslie, Santiago Ontanon, James Lee-Thorp, Michiel de Jong, Jianmo Ni, Honglei Zhuang, and Ed Chi. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023. URL <https://arxiv.org/abs/2305.13245>.
- Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *NAACL-HLT*, pages 2357–2367, 2019. doi: 10.18653/v1/N19-1245.
- Yutao Bai, Yuxiao Bai, Yichang Cui, Zeyu Deng, Zihan Dong, Yifan Fu, Fan Guo, Junjie Han, Le Hou, Zhen Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Tamas Li, Raghu Meka, Sashank Reddi, Sujay Sanghavi, and Suvrit Sra. Low-rank bottleneck in multi-head attention models. In *International Conference on Machine Learning*, pages 864–873. PMLR, 2020.
- Aviv Bick, Eric Xing, and Albert Gu. Understanding the skill gap in recurrent language models: The role of the gather-and-aggregate mechanism, 2025. URL <https://arxiv.org/abs/2504.18574>.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, John Schulman, James Betker, Sam Grey, Michael Dong, Bruno Sauce, Steven Chan, Clemens Winter, Daniel Tetelbaum, Natasha McAleese, Timothy Lillicrap, Ilya Sutskever, Wojciech Zaremba, and Amanda Power. Training verifiers to solve math word problems. In *Advances in Neural Information Processing Systems*, 2021a.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021b.
- DeepSeek-AI, Z. Z. Ren, Zhihong Shao, Junxiao Song, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL <https://arxiv.org/abs/2501.12948>.
- Xiaohan Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pages 2793–2803. PMLR, 2021.
- Razvan-Gabriel Dumitru, Paul-Ioan Clotan, Vikas Yadav, Darius Peteleaza, and Mihai Surdeanu. Change is the only constant: Dynamic llm slicing based on layer redundancy, 2024. URL <https://arxiv.org/abs/2411.03513>.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL <https://zenodo.org/records/12608602>.
- Andrey Gromov, Kushal Tirumala, et al. The unreasonable ineffectiveness of the deeper layers. *arXiv preprint arXiv:2403.17887*, 2024.

- Albert Gu and Tri Dao. On the skill gap in recurrent sequence models: Kv-retrieval as a benchmark for long-range dependencies. In *Advances in Neural Information Processing Systems*, 2023.
- Boris Hanin. Random fully connected neural networks as perturbatively solvable hierarchies, 2023. URL <https://arxiv.org/abs/2204.01058>.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2021.
- Sepp Hochreiter. Untersuchungen zu dynamischen neuronalen netzen. *Diploma, Technische Universität München*, 1991.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Luke Zettlemoyer. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.
- Pengxiang Li, Lu Yin, and Shiwei Liu. Mix-In: Unleashing the power of deeper layers by combining pre-In and post-In, 2025. URL <https://arxiv.org/abs/2412.13795>.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Diba Mirza Soylu, Michihiro Yasunaga, Yian Zhang, Hanlin Zha, Aditi Raghunathan Kumar, Tatsunori Hashimoto, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.
- Haiquan Lu, Yefan Zhou, Shiwei Liu, Zhangyang “Atlas” Wang, Michael W. Mahoney, and Yaoqing Yang. Alphapruning: Using heavy-tailed self regularization theory for improved layer-wise pruning of large language models. In *Proceedings of the Thirty-Eighth Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024.
- Mathematical Association of America. American invitational mathematics examination (aime). <https://www.maa.org/math-competitions>, 1983.
- Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. Shortgpt: Layers in large language models are more redundant than you expect. In *arXiv preprint arXiv:2403.03853*, 2024.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, pages 2381–2391, 2018.
- Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. Compact language models via pruning and knowledge distillation, 2024. URL <https://arxiv.org/abs/2407.14679>.
- OpenAI. Math500 benchmark. https://github.com/openai/math_eval, 2023.
- Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Nghia The Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context. In *ACL*, pages 1525–1534, 2016.
- Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. Kilt: a benchmark for knowledge intensive language tasks. In *Proceedings of NAACL*, pages 2523–2544, 2021. doi: 10.18653/v1/2021.nacl-main.200.
- Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

David Rein, Niklas Muennighoff, Ross Taylor, Colin Anderson, Keiran Purohit, Chirag Barot, Kyunghyun Cho Kim, Douwe Kiela, Y-Lan Boureau, Jason Weston, et al. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Shoaib Ahmed Siddiqui, Xin Dong, Greg Heinrich, Thomas Breuel, Jan Kautz, David Krueger, and Pavlo Molchanov. A deeper look at depth pruning of llms, 2024. URL <https://arxiv.org/abs/2407.16286>.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models, 2025. URL <https://arxiv.org/abs/2502.02013>.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022. URL <https://arxiv.org/abs/2206.04615>.

Wenfang Sun, Xinyuan Song, Pengxiang Li, Lu Yin, Yefeng Zheng, and Shiwei Liu. The curse of depth in large language models, 2025. URL <https://arxiv.org/abs/2502.05795>.

Sho Takase, Shun Kiyono, Sosuke Kobayashi, and Jun Suzuki. Spike no more: Stabilizing the pre-training of large language models, 2025. URL <https://arxiv.org/abs/2312.16903>.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL <https://arxiv.org/abs/2201.11903>.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *ACL*, pages 4791–4800, 2019.