# How Relevance Emerges: A Mechanistic Analysis of LoRA Fine-Tuning in Reranking LLMs

ATHARVA NIJASURE, TANYA CHOWDHURY, and JAMES ALLAN,

Center for Intelligent Information Retrieval, Manning College of Information and Computer Sciences, University of Massachusetts Amherst, USA

We conduct a behavioral exploration of LoRA fine-tuned LLMs for Passage Reranking to understand how relevance signals are learned and deployed by Large Language Models. By fine-tuning Mistral-7B, LLaMA3.1-8B, and Pythia-6.9B on MS MARCO under diverse LoRA configurations, we investigate how relevance modeling evolves across checkpoints, the impact of LoRA rank (1, 2, 8, 32), and the relative importance of updated MHA vs. MLP components. Our ablations reveal which layers and projections within LoRA transformations are most critical for reranking accuracy. These findings offer fresh explanations into LoRA's adaptation mechanisms, setting the stage for deeper mechanistic studies in Information Retrieval. All models used in this study can be found here[1].

## 1 Background

**Motivation:** Fine-tuning Large Language Models (LLMs) using Low-Rank Adaptation (LoRA) has become a popular approach for adapting pre-trained transformers to Information Retrieval tasks. However, the inner workings of these fine-tuned models remain largely opaque, limiting our ability to understand how and where relevance signals are encoded and used. In this study, we attempt to demistify the workings of the LoRA updates that grant ranking ability to LLMs. This work contributes toward our long-term goal of uncovering novel latent features encoded in the MLP layers of LLMs and integrating them into conventional statistical ranking models.

**Reranking Task:** We use the Tevatron repository[2] to fine-tune LLMs for passage reranking on MS MARCO [6], similar to the existing work [1, 5]. Given a query-document pair, the model learns to predict a relevance score via a cross-entropy loss, using hard negatives for contrastive learning. We evaluate performance using nDCG@10 on the TREC DL19 and DL20 benchmarks [2].

**LoRA Fine-Tuning:** Instead of updating all model parameters, LoRA [3] injects a lightweight, low-rank approximation into existing layers, drastically reducing computational overhead and storage needs. This low-rank module is merged with the base model at inference time, often achieving parity with full fine-tuning in tasks such as passage reranking [4, 7].

## 2 Experiments

We examine how relevance modeling evolves in LoRA fine-tuning, the influence of LoRA rank, the distinct roles of MHA and MLP updates, and layer-wise contributions of LoRA to final predictions. All of these results are from the test set (DL19).

**Emerging relevance during LoRA fine-tuning:** We track how reranking performance evolves across fine-tuning checkpoints, as shown in Table 1. All models improve steadily over time, with LLaMA3 and Pythia surpassing Mistral by Step 50 in terms of MRR and NDCG@10. By Step 300, LLaMA3 slightly edges out Pythia, indicating stronger late-stage

[1]https://huggingface.co/AtharvaNijasureUMass/
[2]https://github.com/texttron/tevatron

Authors' Contact Information: Atharva Nijasure; Tanya Chowdhury; James Allan, {anijasure,tchowdhury,allan}@cs.umass.edu,
Center for Intelligent Information Retrieval, Manning College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, MA, USA.

performance, although all models converge by then. Please refer to the *figure 1* for how NDCG@10 evolves across fine-tuning checkpoints for each base model fine-tuned with LoRA rank 8.

**Effect of varying LoRA rank on model adaptation:** We experiment with ranks 1,2,8,32 to fine-tune the LLMs for the passage reranking task. The results are presented in Table 2. We observe that the models achieve comparable ranking performance across different LoRA ranks, suggesting that they adapt similarly irrespective of rank. Notably, a LoRA rank of 1 is already sufficient to encode effective ranking behavior, at least on the MS MARCO dataset. Please refer to *Figure 2* and *Figure 3* to see how NDCG@10 scores evolve across different LoRA ranks.

**MHA vs MLP LoRa updates:** We perform experiments zeroing out MHA updates and MLP updates of the LoRA fine-tuned model individually, as shown in Table 3. MLP-only and MHA-only updates both significantly improve ranking quality, with MHA giving a stronger boost than MLP. However, combining both leads to the highest MRR and NDCG@10 for all three models, indicating that each component contributes complementary benefits. As we are primarily interested in extracting novel features within LLMs, we here-on only fine-tune on the MLP LoRA components.

**Layer and Projection-Specific Insights in LoRA:** We fine-tune only the MLP layers of LLMs and perform ablation studies by selectively zeroing out specific layers and projections – i.e., replacing their fine-tuned weights with those from the base model. Our results (Table 4 and *figure 4*) show that layers 5–15 contribute most to relevance prediction because when just those updates are removed, the performance plummets. We also see that the *up* and *gate* projections are substantially more impactful than the *down* projection in the MLP blocks. Additionally, we find that LoRA updates applied only to the MLP layers can recover up to 98% of the relevance modeling performance achieved by updating both MHA and MLP components. This potentially indicates that ranking specific information is localized to certain layers and projections within the LLM.

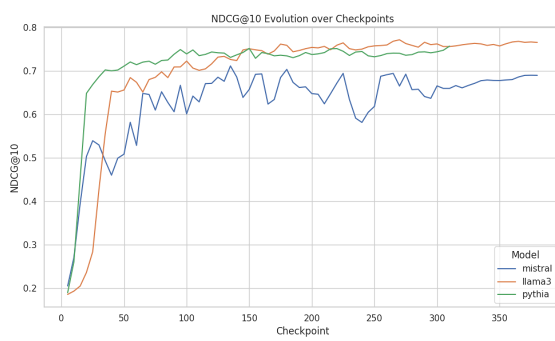


Fig. 1. NDCG@10 evolution across LoRA fine-tuning checkpoints for LLaMa3, Mistral, and Pythia models (rank = 8). Relevance improves rapidly within the first 50 steps and stabilizes after 300 steps.

## 3 Conclusion

Our experiments provided a detailed glance into where and how relevance signals emerge in LoRA-based fine-tuning for passage reranking. We found that even a LoRA rank of 1 can capture effective ranking behavior, that MHA and MLP

Table 1. Tracking Relevance score across LoRA fine tuning checkpoints (steps) for Llama3, Mistral and Pythia.

| Step | LLaMA3 | | | Mistral | | | Pythia | | |
|---|---|---|---|---|---|---|---|---|---|
| | MAP | MRR | NDCG@10 | MAP | MRR | NDCG@10 | MAP | MRR | NDCG@10 |
| 5 | 0.1887 | 0.4220 | 0.1855 | 0.1969 | 0.4147 | 0.2050 | 0.1955 | 0.4339 | 0.1895 |
| 10 | 0.1932 | 0.4485 | 0.1931 | 0.2350 | 0.5224 | 0.2708 | 0.2296 | 0.4947 | 0.2606 |
| 100 | 0.4861 | 0.9380 | 0.7225 | 0.4094 | 0.7657 | 0.6010 | 0.4957 | 0.9884 | 0.7390 |
| 300 | 0.5141 | 0.9709 | 0.7620 | 0.4652 | 0.8948 | 0.6654 | 0.5104 | 0.9814 | 0.7439 |

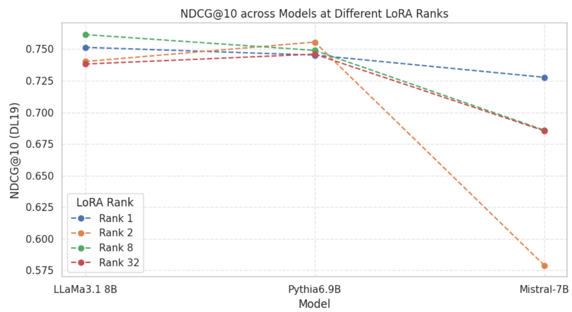


Fig. 2. NDCG@10 scores for rerankers fine-tuned with different LoRA ranks (1, 2, 8, 32). Scores are closely clustered, indicating limited sensitivity to rank choice for LLaMA and Pythia models.

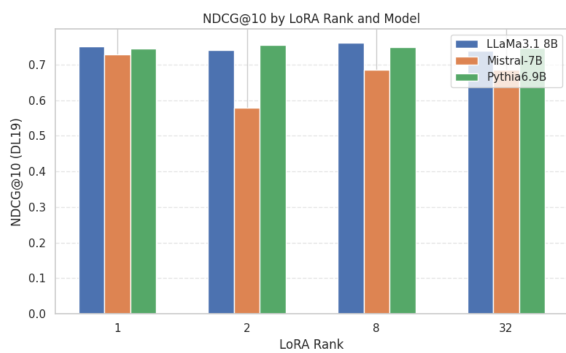


Fig. 3. NDCG@10 scores for rerankers fine-tuned with different LoRA ranks (1, 2, 8, 32). Scores are closely clustered, indicating limited sensitivity to rank choice.

Table 2. NDCG@10 on DL-19 for LLMs fine tuned with different LoRA Ranks, along with Checkpoint at which they converge.

| **Rank** | **Base Model** | | |
|---|---|---|---|
| | **LLaMa3.1 8B** | **Mistral-7B** | **Pythia6.9b** |
| 1 | 0.7514 (150) | 0.7277 (200) | 0.7453 (75) |
| 2 | 0.7402 (100) | 0.5790 (400) | 0.7556 (100) |
| 8 | 0.7615 (175) | 0.6859 (140) | 0.7490 (95) |
| 32 | 0.7382 (175) | 0.6854 (125) | 0.7460 (75) |

updates contribute complementary gains (with MHA offering a stronger boost), that mid-range MLP layers (especially up and gate projections) are most critical for encoding relevance and that MLP only LoRA updates can recover upto

Table 3. Evaluation after zeroing out MLP or MHA components across all layers in a reranker model finetuned with LoRA Rank 8.

| Setting | LLaMa3 | | Mistral | | Pythia | |
|---|---|---|---|---|---|---|
| | MRR | NDCG@10 | MRR | NDCG@10 | MRR | NDCG@10 |
| No LoRA updates | 0.422 | 0.1855 | 0.2978 | 0.1450 | 0.4339 | 0.1895 |
| MLP updated only | 0.6971 | 0.4341 | 0.5899 | 0.4390 | 0.9048 | 0.6262 |
| MHA updated only | 0.8508 | 0.5987 | 0.7101 | 0.5164 | 0.9593 | 0.7009 |
| Both MLP & MHA updated | 0.9543 | 0.7655 | 0.9050 | 0.6891 | 0.9709 | 0.7569 |

Table 4. Evaluation with various layer/projection ablations with MLP-only LoRA models on DL19 (NDCG@10).

| LoRA Updates | LLaMA | Mistral | Pythia | LoRA Updates | LLaMA | Mistral | Pythia |
|---|---|---|---|---|---|---|---|
| All Layers | 0.7497 | 0.7318 | 0.7570 | Layers 0 to 4 and 16 to 31 | 0.2900 | 0.2494 | 0.4016 |
| Layers 5 to 15 | 0.6678 | 0.5559 | 0.5895 | All Layers Up & gate_proj | 0.7373 | 0.6947 | 0.7262 |
| Layers 5 to 31 | 0.7184 | 0.6831 | 0.6192 | Layers 0 to 15 | 0.7065 | 0.5640 | 0.7463 |

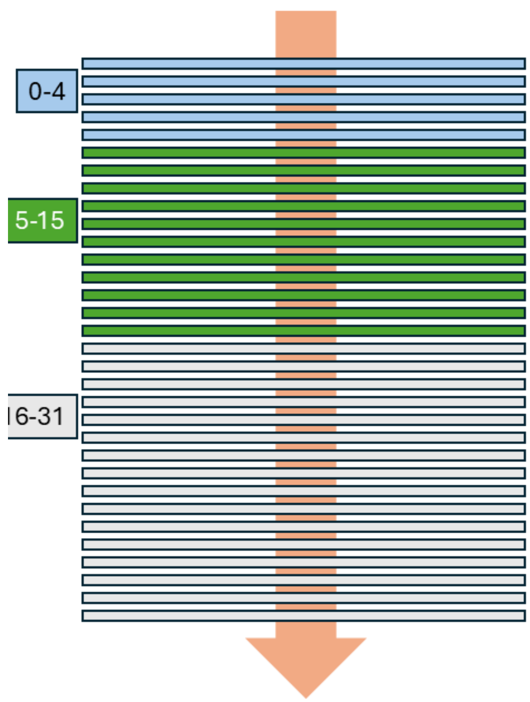


Fig. 4. Effect of LoRA updates applied to different layer ranges and components (MLP-only) in LLaMA3, Mistral, and Pythia models. Updating only layers 5–15 retains 88–92% of full performance, while updating only the Up+Gate projections recovers 96%.

98% of relevance information. These insights explain the role of of low-rank adaptations in IR, and they pave the way for further mechanistic investigations into how LLMs learn and store retrieval-specific knowledge.

## References


[1] Tanya Chowdhury and James Allan. 2024. Probing Ranking LLMs: Mechanistic Interpretability in Information Retrieval. *arXiv preprint arXiv:2410.18527* (2024).

[2] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the TREC 2019 deep learning track. *arXiv preprint arXiv:2003.07820* (2020).

[3] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.

[4] Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023. Fine-Tuning LLaMA for Multi-Stage Text Retrieval. CoRR abs/2310.08319 (2023).

[5] Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2421–2425.

[6] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset. (2016).

[7] Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. 2024. Lora vs full fine-tuning: An illusion of equivalence. *arXiv preprint arXiv:2410.21228* (2024).