Understanding Pre-trained and Fine-tuned model behaviour using Model Diffing

Anonymous Author(s)
Affiliation
Address
email

Abstract

Fine-tuning large language models (LLMs) for specialized domains can alter both their output distributions and internal mechanisms in ways that standard task metrics obscure. We study *model diffing* between a pretrained base (DeepSeek-R1-Distill-Qwen-1.5B) and a LoRA-adapted variant trained for medical reasoning on HuatuoGPT-o1 style data. Our protocol couples (i) *next-token KL divergence*, measured across general and medical corpora—to quantify output-level shift, with (ii) *activation patching* to localize where domain knowledge and reasoning procedures are encoded. We target the LoRA-modified projections and MLP pathways and analyze the behavioral impact of swapping per-layer activations across models. Complementary experiments explore Complex Chain-of-Thought fine-tuning and a *Kahneman-Tversky Optimization (KTO)* objective from the Human-Aware Loss family to encourage structured reasoning without preference labels. Empirically, we observe domain-selective distributional drift with minimal degradation on general text, and layerwise concentration of medical competence consistent with prior findings that factual/semantic knowledge often resides in mid-to-late MLP blocks. Our contributions are: (1) a unified, reproducible KL plus patching diffing protocol; (2) evidence on how LoRA placements mediate domain specialization; and (3) an analysis of reasoning-oriented post-training (CoT/KTO) and its interaction with representational localization. We release code and scripts to support systematic model diffing in domain-specific alignment.

1 Introduction

Large language models (LLMs) have achieved remarkable performance across open-domain and specialized reasoning tasks, yet understanding how fine-tuning reshapes their internal mechanisms remains an open challenge. In domains such as medical reasoning, where reliability and interpretability are essential, it is not sufficient to evaluate models solely on accuracy. Instead, one must also interrogate the representational shifts and causal mechanisms that emerge when a general-purpose model is adapted to a specialized domain.

Recent advances in mechanistic interpretability have begun to expose the circuits and abstractions underpinning transformer models. Foundational work established mathematical frameworks for analyzing attention and feed-forward pathways Elhage et al. [2021], Geva et al. [2020] and revealed that factual associations can be localized and edited via targeted interventions Meng et al. [2022]. Techniques such as activation patching Syed et al. [2023], Zhang and Nanda [2023] and attribution methods Vig et al. [2020], Conmy et al. [2023] allow researchers to causally probe how specific layers and heads contribute to predictions, while recent benchmarks Mueller et al. [2025] and circuit-analysis efforts Wu et al. [2024a] seek to standardize evaluation. These approaches have shown that domain

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

knowledge, heuristics, and even undesirable behaviors such as sycophancy Li et al. [2025] can often be traced to mid-to-late transformer blocks, underscoring the importance of layer-level analysis.

Parallel strands of work have investigated representation divergence as a quantitative measure of model adaptation. Studies of mode connectivity and representation finetuning Zhou et al. [2023,?], Wu et al. [2024b] suggest that fine-tuning induces structured but non-trivial shifts in embedding spaces. Kullback–Leibler (KL) divergence has been widely used to quantify such distributional changes, both in knowledge distillation Wu et al. [2024a] and in tracking training dynamics Kishino et al. [2025]. These methods highlight that fine-tuning not only alters outputs but also reorganizes internal representations in ways that can be systematically measured.

At the same time, alignment research has advanced techniques to encourage structured reasoning and human-centric behavior. Chain-of-Thought prompting Wei et al. [2022] demonstrated that eliciting intermediate steps enhances reliability. Preference-based fine-tuning Ziegler et al. [2019], **?** and Direct Preference Optimization Rafailov et al. [2023]introduced frameworks for aligning models with human judgments, though they require costly data collection. To reduce this dependency, cognitive-inspired approaches such as Kahneman-Tversky Optimization Ethayarajh et al. [2024] and human-aware loss functions Qi et al. [2025] leverage decision-theoretic principles Kahneman and Tversky [2013] to regularize models toward trustworthy reasoning without explicit preference labels. Emerging methods like chain-of-preference optimization Zhang et al. [2024] and steering vectors for reasoning Venhoff et al. [2025] further suggest that interpretability and alignment can be jointly optimized.

Against this backdrop, our work contributes a unified framework for model diffing that combines KL divergence and activation patching to study how fine-tuning for medical reasoning reshapes an LLM's internals. We focus on DeepSeek-R1-Distill-Qwen-1.5B one of the Deepseek series of models Guo et al. [2025]and its LoRA-adapted variant Hu et al. [2022] trained on domain-specific medical reasoning tasks. By quantifying output divergence and performing layerwise patching on LoRA-modified projections and MLP blocks, we identify where medical knowledge and reasoning capabilities are embedded. We further evaluate reasoning-oriented objectives such as Complex Chain-of-Thought fine-tuning and KTO-based HALOs, probing how they interact with representational dynamics.

Our contributions are threefold

1. We introduce a **KL + activation** patching diffing protocol for systematically comparing base and fine-tuned LLMs.

2. We provide **empirical evidence** of layerwise specialization and representational divergence induced by medical fine-tuning under LoRA adaptation.

3. We analyze how **reasoning-oriented objectives** shape both output distributions and internal mechanisms, offering insights for building interpretable, trustworthy medical AI systems.

2 Dataset

To evaluate how fine-tuning reshapes model internals for specialized reasoning, we trained on the HuatuoGPT-o1 medical reasoning dataset developed by Freedom Intelligence [1]. We used the instruction-tuned subset released as medical-o1-reasoning-SFT [2]

This dataset was chosen because it aligns with our research objective of studying how fine-tuning for domain-specific reasoning alters internal representations. Unlike general medical QA corpora, HuatuoGPT-o1 explicitly encodes reasoning traces, which makes it particularly well-suited for model diffing experiments that combine output divergence (via KL) and causal analysis (via activation patching).

[1] https://github.com/FreedomIntelligence/HuatuoGPT-o1

[2] https://huggingface.co/datasets/FreedomIntelligence/medical-o1-reasoning-SFT

3 Methodology

We fine-tuned the DeepSeek-R1-Distill-Qwen-1.5B model using Low-Rank Adaptation (LoRA) with memory-efficient optimizations. To enable large sequence processing, we set the maximum sequence length to 4096 tokens. The base model was loaded with 4-bit quantization (bnb-4bit), reducing GPU memory requirements while maintaining competitive performance. A custom chat template was applied to align the model with medical reasoning tasks, encouraging structured chain-of-thought style responses.

For parameter-efficient fine-tuning, LoRA adapters (rank = 16, scaling factor = 16, no bias) were injected into the attention and MLP projection layers (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, and down_proj). Gradient checkpointing was enabled to reduce memory overhead, and random initialization seeds ensured reproducibility.

We align the LoRA-augmented DeepSeek-R1-Distill-Qwen-1.5B with Kahneman–Tversky Optimization (KTO) Ethayarajh et al. [2024] via TRL's KTOTrainer [3]. KTO treats feedback as binary desirability signals per sample and maximizes a prospect-theoretic utility using asymmetric value/weighting functions, placing greater penalty on undesirable generations relative to comparable gains for desirable ones. Practically, this removes the need for pairwise comparisons while targeting the same alignment goals as preference-based methods. We train with an effective batch size of 8 (4 × grad-accum 2), AdamW-8bit, cosine LR schedule with 10% warmup, weight decay 0.01, and a base learning rate of 5e-7. Mixed precision (FP16/BF16) is used based on hardware support.

To further optimize memory usage during training, we implemented a custom memory management callback. This routine explicitly deletes input and output tensors at the end of each step, clears the CUDA cache, and logs memory statistics every 100 steps. GPU memory utilization was monitored to ensure stable training under constrained resources.

The model was fine-tuned on a subset of the medical reasoning dataset, using the tokenizer for data preprocessing. The overall setup allows efficient domain adaptation of a 1.5B parameter model on modest hardware, and training was conducted on an NVIDIA A100 GPU, enabling controlled experimentation with model diffing and mechanistic interpretability tools.

4 Results

4.1 KL Divergence Analysis

4.1.1 Token-level KL Divergence Analysis

To quantify how fine-tuning alters the model's behavior, we computed the KL divergence between the pretrained and fine-tuned models at each token position of a sample prompt. Both models produced logits of shape (1, 47, 151,936), corresponding to 47 tokens and a shared vocabulary. The summed KL divergence across tokens was 0.2449, with individual values ranging from approximately –0.0027 to 0.0183.

The divergence values indicate that most tokens exhibit small but non-negligible shifts, while a few tokens show localized spikes in divergence (e.g., >0.015). This suggests that fine-tuning does not uniformly shift the output distribution but instead induces targeted representational changes at specific positions. Such differences likely correspond to tokens where medical reasoning knowledge has been reinforced or reweighted.

The bar plot in Figure 1 illustrates these token-level divergences. While many tokens cluster near zero, the variance in divergence magnitudes highlights which lexical items were most impacted by domain adaptation. These findings support the hypothesis that fine-tuning selectively alters the model's probability landscape, rather than producing a homogeneous distributional drift.

4.1.2 Token-level KL Divergence Across Generated Responses

We evaluated token-level KL divergence between the pretrained and fine-tuned models across five generated medical reasoning responses. The average per-token divergence was very small (7×10 to

[3] https://huggingface.co/docs/trl/main/en/kto$_trainer$

8×10^3), confirming that fine-tuning preserves the overall distributional structure of the base model. However, divergence did not remain constant across the sequence. As shown in Figure 2, the per-token KL values fluctuate along the response, with localized spikes at medically salient positions. The moving average reveals that divergence gradually rises during intermediate reasoning steps before tapering off, suggesting that the model makes targeted adjustments when introducing or justifying domain-specific knowledge.

While the cumulative KL divergence varied more substantially (0.39–2.29 across responses), this reflects the accumulation of small but meaningful shifts over longer generations. Together, these patterns indicate that fine-tuning does not globally distort the pretrained model's behavior; instead, it introduces selective representational refinements concentrated at key reasoning junctures, while maintaining consistency elsewhere.

Table 1: Aggregate results over 10 generated responses: averaged token-level KL divergence and text quality scores.

N	Avg Mean KL	Avg Sum KL	FT ROUGE-L	FT BERT F1	Base ROUGE-L	Base BERT F1
10	0.00696	0.34786	0.2222	0.8503	0.2197	0.8504

4.2 Activation Patching Analysis

While KL divergence quantifies output-level shifts, it does not directly reveal where in the network such shifts are encoded. Activation patching provides a causal tool for localizing representational changes: by selectively substituting hidden activations from one model into another, we can test which layers or submodules are most responsible for domain-specific behaviors. This technique has been widely used in mechanistic interpretability to identify circuits, attribute predictions to specific components, and trace factual knowledge to mid-to-late transformer blocks. In the context of model diffing, activation patching is crucial because it moves beyond correlation to establish causal responsibility for divergence between pretrained and fine-tuned models.

4.2.1 Methods and Results

We applied activation patching to the LoRA-augmented DeepSeek-R1-Distill-Qwen-1.5B. Specifically, we targeted the LoRA-injected projection pathways (`q_proj`, `k_proj`, `v_proj`, `o_proj`) and the MLP subcomponents (`gate_proj`, `up_proj`, `down_proj`). For each layer, we replaced the fine-tuned model's activations with their pretrained counterparts during forward passes on medical reasoning prompts. By measuring the change in the fine-tuned model's output distributions, we isolated which layers are most critical for encoding the medical reasoning capability.

Figure 3 shows the mean KL divergence per layer under patching interventions. The attention projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`) exhibit minimal divergence (≈ 0.0068), suggesting that fine-tuning preserved the pretrained model's attention mechanisms. In contrast, the MLP projections show markedly higher divergence, peaking in `up_proj` (0.0106) and `down_proj` (0.0086). These results indicate that fine-tuning primarily reshaped the feed-forward pathways, consistent with prior findings that factual and semantic knowledge tends to be stored in MLP layers.

Notably, patching experiments revealed that medical reasoning ability is causally localized in the LoRA-modified MLP layers, whereas attention pathways remain broadly intact. Furthermore, by restricting fine-tuning to selected LoRA adapters, we not only maintained strong domain adaptation but also substantially reduced computational costs, lowering GPU memory usage and thereby reducing the carbon footprint of training. These findings underscore that activation patching, combined with divergence analysis, provides a layered view of how parameter-efficient fine-tuning embeds domain knowledge into compact feed-forward pathways while leaving general-purpose attentional scaffolding largely unchanged.

References

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2020.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

Ryo Kishino, Yusuke Takase, Momose Oyama, Hiroaki Yamagiwa, and Hidetoshi Shimodaira. Revealing language model trajectories via kullback-leibler divergence. *arXiv preprint arXiv:2505.15353*, 2025.

Jin Li, Keyu Wang, Shu Yang, Zhuoran Zhang, and Di Wang. When truth is overridden: Uncovering the internal origins of sycophancy in large language models. *arXiv preprint arXiv:2508.02087*, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, et al. Mib: A mechanistic interpretability benchmark. *arXiv preprint arXiv:2504.13151*, 2025.

Xuan Qi, Jiahao Qiu, Xinzhe Juan, Yue Wu, and Mengdi Wang. Shallow preference signals: Large language model aligns even better with truncated data? *arXiv preprint arXiv:2505.17122*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. *arXiv preprint arXiv:2310.10348*, 2023.

Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. Understanding reasoning in thinking language models via steering vectors. *arXiv preprint arXiv:2506.18167*, 2025.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. Rethinking kullback-leibler divergence in knowledge distillation for large language models. *arXiv preprint arXiv:2404.02657*, 2024a.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Reft: Representation finetuning for language models. *Advances in Neural Information Processing Systems*, 37:63908–63962, 2024b.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*, 2023.

Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *Advances in Neural Information Processing Systems*, 37:333–356, 2024.

Zhanpeng Zhou, Yongyi Yang, Xiaojiang Yang, Junchi Yan, and Wei Hu. Going beyond linear mode connectivity: The layerwise linear feature connectivity. *Advances in neural information processing systems*, 36:60853–60877, 2023.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Technical Appendices and Supplementary Material

A.1 Limitations

Our study is limited to a single base model and dataset, so the findings may not generalize across domains or architectures. The KL and activation patching analyses were performed on a relatively small set of prompts, making the results more illustrative than statistically exhaustive. In addition, patching identifies layer-level effects but does not capture finer-grained circuits within neurons or attention heads. Finally, while we highlight reduced computational costs and emissions, we do not provide a full quantitative carbon analysis.

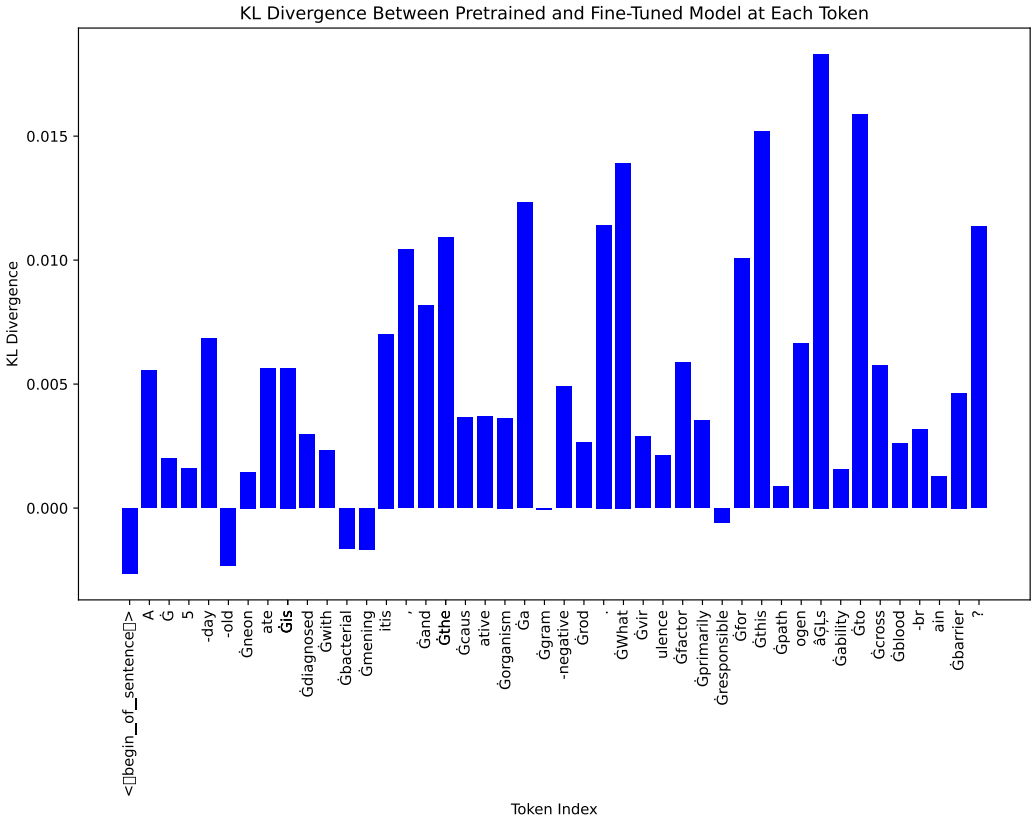

Figure 1: Token-level KL divergence between the pretrained and fine-tuned model. The plot shows localized spikes in divergence, indicating that fine-tuning selectively alters the probability distribution at specific tokens.

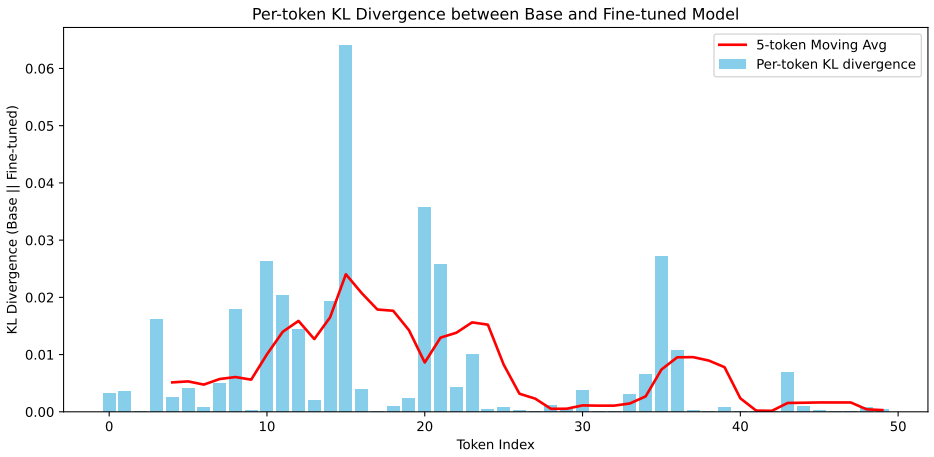

Figure 2: Token-level KL divergence between the pretrained and fine-tuned model. The divergence plot shows localized spikes, indicating that fine-tuning selectively alters the probability distribution at specific tokens.

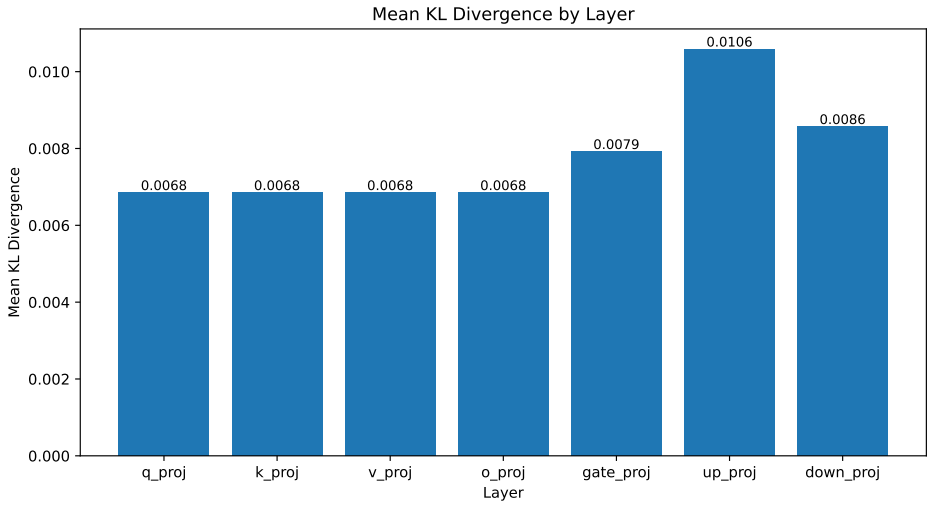

Figure 3: Mean KL divergence between the base and fine-tuned model across projection layers. Fine-tuning introduces the largest divergence in `up_proj` and `down_proj` layers, while attention projection layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`) show smaller shifts.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.

- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: As discussed in the abstract. We have

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification: We have added the limitations section in Appendix

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.

- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We are providing the full code used in this paper over the GitHub repository.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The model and dataset used in the paper is openly accessable from Huggingface models and datasets library which we already provided the code for the same.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In the methodology section we have completely provided our training details including the hyperparameters

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, we have compared using activation patching mechanism which is best suitable for comparing the effect of individual modules.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided full information about the code, data and fine-tuning in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We found while fine-tuning the model on selected LoRA adaptors we can reduce the computational costs which greately impacts on the carbon emission.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Models and data have taken from huggingface platform.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.