FROM POST-TRAINING PLASTICITY TO FUNCTION-TARGETED ADAPTATION IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training is central to developing and enhancing the capabilities of pretrained large language models (LLMs), yet its effects on their internal functional organization remain poorly understood. Inspired by experience-driven plasticity in biological intelligence, we investigate *post-training plasticity*: how the functional organization of attention heads evolves as LLMs acquire or enhance capabilities through post-training. We find that post-training primarily strengthens functional structures already present in pretrained models. This observation motivates *Function-Targeted Adaptation (FTA)*, a selective fine-tuning paradigm that translates findings on post-training plasticity into targeted parameter updates. FTA identifies attention heads associated with a target capability and restricts updates to their corresponding weights, allowing us to examine whether existing functional structures provide a shared basis for both capability enhancement and knowledge suppression. We evaluate this hypothesis through skill learning and knowledge unlearning, assessing target adaptation alongside retention of non-target capabilities. Across multiple models and adaptation objectives, FTA updates at most around 10% of attention heads yet matches or exceeds full fine-tuning in target adaptation, while substantially reducing interference with non-target capabilities and retaining performance close to that of the original model. These results connect a mechanistic understanding of post-training plasticity with selective model adaptation, demonstrating how existing functional structures can guide both learning and unlearning.

1 INTRODUCTION

Human intelligence is remarkably plastic: learning and experience can reshape existing neural systems to support new skills, as illustrated by hippocampal structural changes associated with spatial-navigation expertise in London taxi drivers (Maguire et al., 2000; 2006). Similarly, post-training (Ouyang et al., 2022; Rafailov et al., 2023; Guo et al., 2025) enhances the capabilities of pretrained large language models (LLMs), including reasoning, coding, and alignment with human preferences. Yet the understanding of this process remains largely behavioral, with limited insight into how capability gains relate to changes in internal functional organization.

Inspired by experience-driven plasticity in biological intelligence, we investigate this question through the lens of **post-training plasticity**: how internal functional organization evolves as LLMs acquire or enhance capabilities through post-training (as shown in Fig. 1). We focus on attention heads, which have been shown to exhibit functional specialization (Li et al., 2023; Wu et al.; Zhou et al., 2024; Ma et al.). Using probing, we identify heads associated with mathematical reasoning and harmlessness and trace their functional patterns before and after post-training. We find that the **top-k head sets remain largely consistent** for each target capability, although their relative rankings shift. Post-training generally increases head importance scores for the target capability, with larger gains concentrated among its top-k heads, while scores for non-target capabilities remain largely unchanged. Together, these findings reveal selective and non-uniform functional plasticity, suggesting that post-training primarily strengthens existing function-associated patterns related to the target capability.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

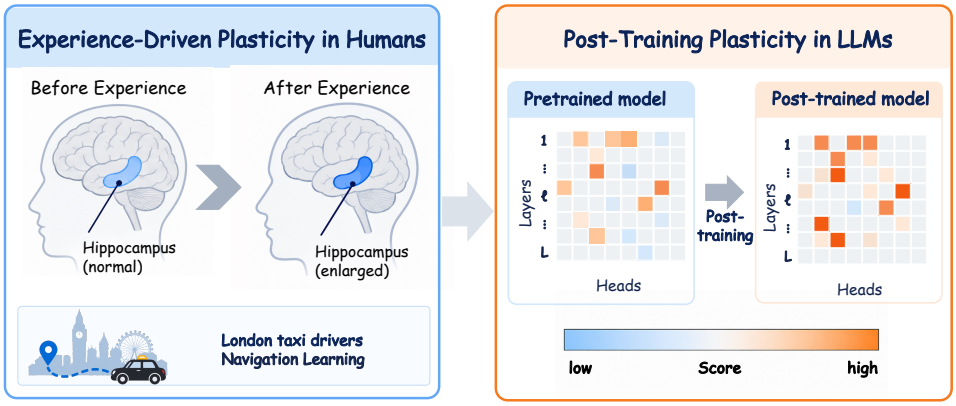

Figure 1: From experience-driven plasticity to post-training plasticity. Left: experience-driven plasticity in humans, where prolonged navigation experience is associated with structural changes in the hippocampus. Right: post-training plasticity in LLMs, where training reorganizes and strengthens pre-existing function-related head patterns.

These observations motivate a more selective approach to model adaptation. If capability enhancement is concentrated in functionally relevant heads, updating the entire model may be unnecessary. Conventional full fine-tuning modifies all parameters, potentially disrupting components that support non-target capabilities and interfering with previously acquired behaviors. Parameter-efficient fine-tuning (PEFT) methods reduce adaptation costs through low-rank updates (Hu et al., 2021) or lightweight adapters (Houlsby et al., 2019), but trade adaptation capacity for efficiency by constraining the space of trainable updates. These constraints typically reflect architectural or parameterization choices rather than the functional roles of the underlying components.

Building on our findings on post-training plasticity, we introduce *Function-Targeted Adaptation (FTA)*, a selective fine-tuning paradigm that uses the model's intrinsic functional organization to guide adaptation. Given a target capability, FTA identifies the associated attention heads and restricts parameter updates to them while keeping the remaining parameters frozen. By concentrating updates within functionally relevant components, FTA aims to achieve effective target adaptation while minimizing interference with non-target capabilities.

We evaluate FTA through two complementary settings: *skill learning* and *knowledge unlearning*. Skill learning examines whether function-targeted updates can enhance a target capability while preserving existing capabilities, whereas knowledge unlearning tests whether they can suppress designated knowledge while retaining broader model utility. Together, these two settings allow us to investigate whether function-associated components provide a general adaptation subspace for modifying model capabilities in either direction. Across multiple LLMs and adaptation objectives, FTA updates only a small subset of attention heads-typically no more than 10%—while matching or outperforming full fine-tuning in target adaptation, while retaining non-target performance close to that of the original model. Notably, in knowledge unlearning setting, FTA suppresses forget-set responses to zero ROUGE-L on TOFU (Maini et al., 2024), with substantially less degradation of non-target capabilities than full fine-tuning. Efficiency measurements further show better target and non-target performance than LoRA, alongside lower wall-clock time and peak GPU memory usage in the evaluated configuration. Together, these results bridge mechanistic understanding and practical model adaptation, establishing functional organization as a common basis for guiding both learning and unlearning while preserving broader model capabilities.

Our main contributions are summarized as follows:

- We investigate *post-training plasticity*, revealing that post-training primarily strengthens functional structures already present in pretrained LLMs.
- We propose Function-Targeted Adaptation, translating these mechanistic insights into selective updates of capability-relevant attention heads.
- Comprehensive evaluations across multiple models, tasks, and adaptation objectives, including SFT and DPO, show that FTA updates only 10% of attention heads yet achieves

superior target adaptation while better preserving non-target capabilities compared with full fine-tuning and other adaptation baselines.

2 Related Work

Post-Training and Mechanism Analysis Post-training plays a central role in transforming pretrained language models into systems with stronger task-specific and instruction-following capabilities. Common approaches include supervised fine-tuning (SFT) (Chung et al., 2024), reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022), and direct preference optimization (DPO) (Rafailov et al., 2023). Recent advances, including reinforcement learning with verifiable rewards, have further improved reasoning (Guo et al., 2025), coding (Cao et al., 2026), and instruction following and alignment with human preferences (Lambert et al., 2024). Despite differences in their optimization objectives, these approaches share the goal of adapting pretrained models toward desired downstream behaviors. Beyond behavioral improvements, recent work examines whether post-training creates new capabilities or primarily elicits and reorganizes existing ones, highlighting the distinct effects of SFT and RL and the role of explicit reasoning in capability expression (Rajani et al., 2025; Chen et al., 2026c; Ma et al., 2025). Complementary mechanistic analyses identify localized refusal representations (Arditi et al., 2024), selective activation and recombination of task-specific attention heads during SFT (Zhao et al., 2024), and newly recruited heads in post-trained reasoning circuits (Park et al., 2025). Recent work further reports changes in circuit activation intensity and diversity under RL and uneven representational changes across layers during SFT (Zhang et al., 2026; Zhao et al., 2026). These findings suggest that post-training reshapes internal mechanisms in regime- and task-dependent ways. Building on this, we trace functionally specialized heads across capabilities and use their evolution to guide selective adaptation.

Parameter-Efficient and Selective Model Adaptation Parameter-efficient fine-tuning (PEFT) reduces adaptation costs through low-rank updates (Hu et al., 2021), lightweight adapters (Houlsby et al., 2019), or continuous prompts and prefixes (Lester et al., 2021; Li & Liang, 2021). Selective methods further allocate updates according to parameter importance or task-dependent statistics (Sung et al., 2021; Zhang et al., 2023; Yang et al., 2025), while recent approaches exploit attention-head importance or specialization to guide data selection and adaptation (Chen et al., 2026b; Liu et al., 2025; Luo et al., 2025). Related work on model unlearning and task arithmetic explores targeted knowledge removal or capability suppression while preserving overall utility (Maini et al., 2024; Li et al., 2024a; Ilharco et al., 2022). Building on these, we introduce FTA to connect post-training changes in capability-associated components with selective adaptation, supporting capability enhancement and knowledge suppression while better preserving non-target capabilities.

3 Post-Training Plasticity

To investigate how the functional organization of an LLM evolves during post-training, we focus on harmlessness alignment, math reasoning and knowledge unlearning. Using probing, we identify attention heads associated with each capability and examine how their functional patterns change during post-training.

3.1 Localizing Capability-Relevant Heads

We consider Anthropic HH-RLHF (Bai et al., 2022), the mathematical subset of UltraInteract (Yuan et al., 2025) (UltraInteract-math), and TOFU (Maini et al., 2024) for harmlessness alignment, mathematical reasoning, and knowledge unlearning respectively. HH-RLHF provides chosen and rejected responses based on human preferences, whereas UltraInteract-math provides correct and incorrect responses for mathematical reasoning. For unlearning, which does not involve paired responses, TOFU provides a designated forget set, allowing us to identify relevant heads directly from the forget-set objective (see Appendix A.1 for details).

For harmlessness alignment and mathematical reasoning, we concatenate each question with its response and extract the activation of every attention head at the final token. For head h in layer ℓ,

this yields a probing dataset:

$$\mathcal{D}_{c,\ell,h}^{\text{probe}} = \left\{ \left(\mathbf{x}_i^{\ell,h}, y_i \right) \right\}_{i=1}^{N_c}, \tag{1}$$

where $\mathbf{x}_i^{\ell,h}$ denotes the head activation for the i-th question–response pair, N_c is the number of pairs, and $y_i = 1$ and $y_i = 0$ indicate chosen/correct and rejected/incorrect responses, respectively, depending on the dataset. Data splitting details are provided in Appendix A.2. Each head feature is first passed through a trainable linear projection for dimensionality reduction, after which a two-layer MLP probe is trained for classification (training details are provided in Appendix A.1). To quantify the contribution of individual heads to the target capability, we apply gradient-based attribution using the gradient×activation score. Specifically, for head h in layer ℓ, we compute $I_{c,\ell,h} = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_c^{\text{probe}}} \left[\frac{\partial \hat{y}_y}{\partial \mathbf{z}^{\ell,h}} \odot \mathbf{z}^{\ell,h} \right]$, where $\mathbf{z}^{\ell,h}$ denotes the projected feature of head h in layer ℓ, $\hat{y}_y$ denotes the logit corresponding to the ground-truth class y, and $\odot$ denotes element-wise multiplication. This yields an importance score for each attention head, with higher scores indicating a stronger association with the target capability. We then rank heads by their importance scores to identify the heads most strongly associated with each capability.

3.2 Evolution of Functional Organization

To characterize how functional organization evolves throughout post-training, we analyze intermediate checkpoints rather than comparing only the initial and final models. We save checkpoints at 0%, 10%, 33%, 66%, and 100% of optimizer steps and recompute head importance scores at each checkpoint. We examine three one-epoch adaptation settings in the Llama-3.1-8B family: enhancing mathematical reasoning by fine-tuning Llama-3.1-8B-Instruct on UltraInteract-Math, acquiring harmlessness by fine-tuning Llama-3.1-8B-Base on HH-RLHF chosen responses, and knowledge unlearning with GradDiff on TOFU `forget05`. The SFT settings use a learning rate of 10^{-5}, while unlearning starts from the TOFU Llama-3.1-8B-Instruct with a learning rate of 5×10^{-5}.

For each target capability c, we select the top-k capability-associated heads according to their importance scores. We determine k from the elbow of the initial score curve and keep it fixed throughout each trajectory, yielding $k = 49$, 116, 100 for mathematical reasoning, harmlessness, and unlearning, respectively (see details in Appendix A.1). For probing-based trajectories, we additionally re-score the same checkpoint using independent probe seeds to assess overlap variability due to probe training. We then track how these identified heads evolve over the course of post-training. Our analysis reveals three main findings as illustrated in Fig. 2.

Pre-existence. More than 60% of the top-k heads identified at each SFT checkpoint also belong to the initial model's top-k (i.e., the 0% checkpoint). Moreover, around 90% fall within the initial top-$2k$ for harmlessness and mathematical reasoning, and about 75% for knowledge unlearning, despite changes in their relative rankings. These overlaps remain high throughout training, suggesting that post-training primarily builds on existing capability-relevant heads.

Non-uniform functional plasticity. Changes in head importance are concentrated among heads that are already associated with the target capability before adaptation, with detailed heatmaps provided in Appendix A.1. As measured by the $|\Delta I|$ concentration, initial top-k heads exhibit systematically larger importance shifts than other heads across both SFT and unlearning, with AUROC consistently above the random baseline of 0.5 in Fig. 2(b). The direction of change further reflects the adaptation objective: SFT generally strengthens these capability-associated heads, whereas unlearning tends to reduce their importance. Together, these results suggest that post-training preferentially reshapes existing capability-associated patterns rather than modifying all heads uniformly.

Cross-Capability Stability. Post-training on one capability leaves the probing scores of heads associated with other capabilities largely unchanged, indicating limited interference with non-target functions. Consistently, Fig. 2(c) shows that cross-capability top-k overlap remains low and relatively stable throughout training. For the SFT settings, the overlap can even decrease from its initial level, suggesting increased specialization of capability-associated heads rather than growing entanglement across functions.

Together, these findings motivate us to explore function-targeted adaptation.

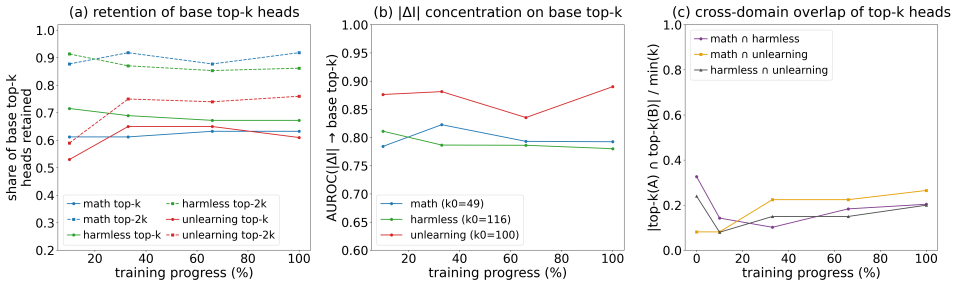

Figure 2: Functional evolution during post-training: (a) Retention: fraction of base top-k heads remaining in the top-k/2k. (b) $|\Delta I|$ concentration: AUROC measuring whether importance changes concentrate on base top-k heads (0.5 = random). (c) Cross-domain overlap: normalized top-k overlap between capabilities.

4 METHOD

Preliminaries. We consider a decoder-only Transformer with L layers, H attention heads per layer, and hidden dimension d. Given a sequence of n tokens, let $X^{\ell-1} \in \mathbb{R}^{n\times d}$ denote the residual stream entering layer ℓ. Each layer applies multi-head attention followed by a position-wise feed-forward network (MLP), both with pre-normalization and residual connections:

$$\widetilde{X}^{\ell} = X^{\ell-1} + \mathrm{MHA}^{\ell}\Big(\mathrm{Norm}^{\ell}_{\mathrm{attn}}(X^{\ell-1})\Big), \quad X^{\ell} = \widetilde{X}^{\ell} + \mathrm{MLP}^{\ell}\Big(\mathrm{Norm}^{\ell}_{\mathrm{mlp}}(\widetilde{X}^{\ell})\Big). \tag{2}$$

For brevity, let $Z^{\ell} = \mathrm{Norm}^{\ell}_{\mathrm{attn}}(X^{\ell-1})$. The output of attention head h in layer ℓ is

$$A^{\ell,h} = \mathrm{softmax}\left(\frac{(Z^{\ell}W_Q^{\ell,h})(Z^{\ell}W_K^{\ell,h})^{\top}}{\sqrt{d_{\mathrm{head}}}} + M\right) Z^{\ell}W_V^{\ell,h}, \tag{3}$$

where M is the causal mask and $W_Q^{\ell,h}, W_K^{\ell,h}, W_V^{\ell,h} \in \mathbb{R}^{d\times d_{\mathrm{head}}}$ are the query, key, and value projections. Multi-head attention aggregates the head outputs as $\mathrm{MHA}^{\ell}(Z^{\ell}) = \sum_{h=1}^{H} A^{\ell,h}W_O^{\ell,h}$, where $W_O^{\ell,h} \in \mathbb{R}^{d_{\mathrm{head}}\times d}$ is the corresponding block of the output projection. For architectures with grouped-query attention, query heads within a group share their key and value projections.

4.1 FUNCTION-TARGETED ADAPTATION

For each target capability c, we apply the probing method described in Section 3.1 only to the initial model (0% checkpoint) to identify capability-relevant attention heads, which remain fixed throughout subsequent fine-tuning. Let $\mathcal{S}_c$ denote the top-k heads ranked by their probing scores. FTA restricts adaptation to their associated projection parameters:

$$\Theta_c = \bigcup_{(\ell,h)\in\mathcal{S}_c} \left\{W_Q^{\ell,h}, W_K^{\ell,h}, W_V^{\ell,h}, W_O^{\ell,h}\right\}. \tag{4}$$

All parameters outside Θ_c remain frozen. Let M_c denote a binary parameter mask with entries equal to one for parameters in Θ_c and zero otherwise. Starting from pretrained parameters $\theta^{(0)}$, FTA optimizes

$$\Delta\theta_c^{*} = \arg\min_{\Delta\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}_c^{\mathrm{adapt}}} \left[\mathcal{L}_c\left(f_{\theta^{(0)}+M_c\odot\Delta\theta}(x), y\right)\right], \tag{5}$$

where $\mathcal{D}_c^{\mathrm{adapt}}$ is the adaptation dataset and $\mathcal{L}_c$ specifies the target objective. The adapted parameters are $\theta_c^{*} = \theta^{(0)} + M_c \odot \Delta\theta_c^{*}$. The head selection $\mathcal{S}_c$ remains fixed throughout adaptation, while the corresponding parameters are updated to support capability enhancement or suppression.

This formulation separates where adaptation occurs from what it optimizes: the probing scores determine the trainable components, while $\mathcal{L}_c$ specifies capability enhancement or suppression. For grouped-query attention, selecting a head also includes its shared key and value projections, so updates to these projections can affect other query heads in the same group.

5 EXPERIMENTS

We evaluate FTA on two LLM families at multiple scales: LLaMA (Touvron et al., 2023) (Llama3.1-8B and Llama3.2-3B) and OLMo (OLMo et al., 2024)(OLMo-2-13B and OLMo-2-7B). Our experiments cover learning tasks in value alignment (harmlessness and honesty) and mathematical reasoning, as well as knowledge unlearning task, assessing whether FTA can selectively enhance target capabilities or remove designated knowledge while preserving non-target capabilities.

Datasets. We use the HH-RLHF (Bai et al., 2022), TruthfulQA (Li et al., 2023), and UltraInteract-math) (Yuan et al., 2025) for harmlessness alignment, honesty, and mathematical reasoning, respectively. For knowledge unlearning, we use the TOFU `forget05` split (Maini et al., 2024). These datasets support both probing-based functional-head identification and subsequent post-training, using capability-specific labels for probing and objective-specific training examples for adaptation.

Baselines and training protocol. We compare FTA with full-parameter fine-tuning, LoRA (r=8) (Hu et al., 2021), and P2D (Chen et al., 2026a), which identifies task-sensitive heads from weight-distribution shifts after a short proxy fine-tuning run to guide sparse updates. We also include a random-head control matched to FTA in trainable-parameter count and update structure. For fairness, all methods are trained on the same data with the same optimizer and schedule; additional details are provided in Appendix A.4. We evaluate skill learning under SFT (Chung et al., 2024) and DPO (Rafailov et al., 2023), and knowledge unlearning under the gradient-difference (GradDiff) objective (Maini et al., 2024).

Evaluation. Across settings, we jointly evaluate target-task performance and non-target capability retention to assess adaptation effectiveness and functional selectivity. Mathematical reasoning is evaluated on GSM8K (Cobbe et al., 2021), SVAMP (Patel et al., 2021), ASDiv (Miao et al., 2020), and GSM+ (Li et al., 2024b) using exact-match accuracy under fixed few-shot prompts. Harmlessness is measured by the safe-response rate on JailbreakBench (Chao et al., 2024), assessed by Qwen3Guard-Gen-8B, and honesty by truthfulness on TruthfulQA testing split (Li et al., 2023), judged by GPT-5.4 (OpenAI, 2026)). For TOFU (Maini et al., 2024) unlearning, we assess forgetting using forget-set ROUGE-L (Lin, 2004). Non-target retention is evaluated on ten benchmarks spanning reasoning, code, knowledge, commonsense, and reading comprehension (Clark et al., 2018; Liu et al., 2023a;b; Gu et al., 2024; Hendrycks et al., 2020; Talmor et al., 2019; Mihaylov et al., 2018; Sap et al., 2019; Rajpurkar et al., 2016; Clark et al., 2019). We use option accuracy for multiple-choice tasks, exact match for open-ended tasks, and pass@1 with the reference test harness for HumanEval+, with fixed benchmark-specific few-shot prompts.

5.1 LEARNING TASK

Table 1 presents results on Llama-3.1-8B and OLMo-2-13B, with results for the other two models reported in Table 8 (Appendix A.5). FTA achieves the highest average scores on both target and non-target capabilities across the two main models and both training objectives. Under SFT, FTA improves OLMo-2-13B's average mathematical reasoning score from 61.3% to 74.6%, outperforming full fine-tuning and LoRA by 2.4 and 4.3 percentage points, respectively. Meanwhile, its non-target average remains close to the base model (61.0% vs. 61.4%), whereas full fine-tuning reduces it to 57.1%. On Llama-3.1-8B, FTA under SFT also slightly improves the non-target average, from 67.5% to 68.3%. These gains span logical reasoning on LogiQA2 (54.0% to 56.1%), code generation on HEval+ (62.8% to 63.4%), and general knowledge on MMLU (62.5% to 63.2%), alongside improvements on SQuAD and BoolQ. As expected, full fine-tuning yields the lowest non-target average across all four model–objective settings, consistent with the risk that unrestricted updates modify components supporting other capabilities and thereby degrade their performance. Notably, all methods are fine-tuned exclusively on UltraInteract-math, while target-capability performance is evaluated on other mathematical reasoning benchmarks. Fine-tuning does not consistently yield gains on these benchmarks: under SFT, full fine-tuning, LoRA, and P2D reduce Llama-3.1-8B's target average from 80.4% to 71.6%, 78.4%, and 66.7%, respectively. This degradation suggests that parameter-efficient updates, such as LoRA, do not necessarily prevent impaired generalization across related benchmarks. For P2D, heads selected based on task sensitivity may capture dataset-specific patterns rather than broader capability-relevant representations.

Table 1: Target and non-target capability performance (%) after post-training on mathematical reasoning. Base models serve as references. Better value in **bold**.

Model	Train	Method	Target					Non-Target									
			GSM8K	SVAMP	ASDiv	GSM+	**Avg**	ARC-C	LogiQA2	HEval+	CRUX	MMLU	CSQA	OBQA	SIQA	SQuAD	BoolQ	**Avg**
Llama-3.1-8B	–	base	81.9	88.2	84.0	67.5	80.4	82.7	54.0	62.8	30.9	62.5	76.2	84.8	73.7	65.6	81.3	67.5
	SFT	full	68.4	83.7	82.6	51.5	71.6	53.8	36.0	59.8	10.6	40.1	70.9	74.4	70.9	69.1	81.5	56.7
		lora	81.3	84.9	81.5	65.8	78.4	78.6	52.0	61.6	**30.8**	58.1	74.8	81.8	**74.2**	**74.1**	**85.2**	67.1
		P2D	70.6	75.6	66.6	54.1	66.7	77.6	37.5	**64.0**	17.0	57.5	76.2	80.2	67.0	35.3	82.3	59.5
		Ours	**82.8**	**89.4**	**83.4**	**68.4**	**81.0**	**81.9**	**56.1**	63.4	30.0	**63.2**	**76.5**	**83.2**	73.5	71.1	84.0	**68.3**
	DPO	full	76.3	80.2	83.0	61.3	75.2	60.7	48.2	34.1	12.9	53.8	63.6	68.8	67.3	55.3	**82.8**	54.8
		lora	81.6	**84.4**	83.2	64.2	78.4	78.9	48.7	**64.6**	26.1	58.0	76.4	83.8	71.0	63.3	80.6	65.1
		P2D	80.1	83.4	83.3	63.4	77.6	81.0	51.2	61.0	**27.8**	60.0	**76.8**	83.8	**71.8**	**67.1**	82.4	66.3
		Ours	**81.8**	82.3	**84.4**	**70.0**	**79.6**	**81.7**	**52.3**	62.8	23.9	**63.2**	76.5	**84.2**	71.0	66.5	82.4	**66.5**
OLMo-2-13B	–	base	52.3	78.7	77.6	36.8	61.3	81.0	46.6	29.9	15.6	58.2	75.8	79.8	70.2	71.2	85.2	61.4
	SFT	full	72.3	80.1	79.4	57.1	72.2	78.1	45.6	12.2	14.4	**59.9**	72.6	78.4	61.4	62.8	86.0	57.1
		lora	66.3	81.7	79.1	54.2	70.3	74.9	45.6	25.6	14.4	59.4	76.8	81.4	69.3	73.2	85.8	60.6
		P2D	67.7	83.2	**81.2**	51.5	70.9	**80.5**	**51.1**	**34.8**	13.6	59.3	**77.2**	**82.6**	**70.2**	**75.8**	**86.4**	**63.1**
		Ours	**74.4**	**84.1**	80.4	**59.6**	**74.6**	78.2	45.0	26.8	**14.6**	58.3	76.6	80.6	69.4	75.6	85.3	61.0
	DPO	full	50.5	77.4	77.0	34.1	59.8	76.0	42.6	14.6	14.4	53.0	68.4	71.4	70.1	56.4	80.9	54.8
		lora	53.6	78.1	77.0	37.5	61.5	75.0	44.3	3.0	13.1	54.0	69.4	78.6	**70.8**	53.8	81.7	54.4
		P2D	**53.7**	78.3	77.4	**38.3**	61.9	79.1	44.7	**24.2**	13.6	55.6	73.7	78.0	68.0	**67.0**	83.8	58.8
		Ours	53.4	**79.2**	**78.1**	37.5	**62.1**	**81.1**	**45.7**	22.0	**15.9**	**56.5**	**75.8**	**78.4**	70.6	66.0	**85.4**	**59.7**

Beyond math reasoning, we examine whether functional selection can improve harmlessness and honesty while preserving broader model utility. Table 2 shows that FTA achieves the best or joint-best alignment performance across the evaluated settings, while LoRA generally yields smaller gains. Results on additional models are reported in Table 9 (Appendix A.5). Across this broad evaluation of non-target retention on ten benchmarks, FTA maintains average performance close to that of the original model, suggesting that its alignment gains do not come at the expense of widespread capability degradation. Although LoRA occasionally achieves slightly higher non-target averages, its alignment improvements are generally smaller, whereas full fine-tuning often substantially degrades non-target performance. Together with the math results, these findings support capability-associated selection as a basis for balancing target adaptation and capability retention across diverse domains.

5.2 Unlearning Task

We next examine whether FTA can selectively suppress designated knowledge while preserving broader model utility. Table 3 reports results on TOFU forget05 using the GradDiff objective Maini et al. (2024) on Llama-3.1-8B, with results on Llama-3.2-3B provided in Table 10 in the Appendix. FTA reduces forget-set ROUGE-L (Lin, 2004) from 98.6% to 0.0%, compared with 0.3% for full fine-tuning and 5.4% for LoRA, indicating the strongest suppression of reference-answer reproduction under this metric. Meanwhile, FTA achieves the highest non-target average of 61.0%, exceeding full fine-tuning by 14.6%. The retention benefits are particularly evident in code generation and execution reasoning (HEval+, CRUX), logical reasoning (LogiQA2), and reading comprehension (SQuAD). Overall, FTA combines stronger measured forgetting with slightly better average retention than LoRA and substantially less degradation than full fine-tuning. These results extend capability-associated selection from enhancement to targeted knowledge suppression, although low forget-set ROUGE-L alone does not establish complete knowledge removal.

5.3 Ablation Study

To assess the contribution of capability-related head selection, we compare FTA with random-head adaptation on Llama-3.1-8B across learning and unlearning tasks in Table 4. Detailed comparisons for each capability are provided in Appendix A.5. Functional selection improves target outcomes in six of seven settings and matches random selection in the remaining setting, with statistical significance reported in Table 16 in Appendix. The largest gains occur in mathematical reasoning under

Table 2: Target and non-target capability performance after post-training for harmlessness and honesty, evaluated on JailbreakBench (Safe%) and TruthfulQA (Honest%), respectively. Base models serve as references. Better value in **bold**.

Model	Value	Train	Method	Target	Non-Target										
				Safe/Hon	ARC-C	LogiQA2	HEval+	CRUX	MMLU	CSQA	OBQA	SIQA	SQuAD	BoolQ	**Avg**
Llama-3.1-8B	harmless	–	base	62.0	79.6	51.9	31.1	35.4	66.6	73.7	78.4	71.0	77.1	86.5	65.1
		SFT	full	94.0	60.0	35.2	22.6	32.4	52.8	62.2	62.2	60.3	51.0	74.5	51.3
			lora	92.0	76.6	50.5	28.7	34.2	66.1	**74.3**	**79.4**	70.2	73.7	81.3	63.5
			P2D	91.0	**79.1**	**51.7**	**30.5**	34.1	**66.6**	73.9	79.2	**70.5**	**76.6**	84.0	**64.6**
			Ours	**99.0**	78.6	49.1	28.7	**34.9**	66.3	73.6	78.6	69.5	74.5	**85.1**	63.9
		DPO	full	**100.0**	60.0	34.1	17.7	24.2	45.7	30.9	61.0	46.6	63.3	74.1	45.8
			lora	**100.0**	**79.0**	41.0	21.1	**37.9**	**65.5**	**73.5**	76.6	68.9	77.1	80.5	62.1
			P2D	**100.0**	76.4	**47.6**	**29.9**	34.5	63.9	70.4	72.0	**69.2**	76.6	74.0	61.4
			Ours	**100.0**	77.7	47.5	24.4	36.2	64.8	72.8	**77.4**	66.3	**77.8**	**83.0**	**62.8**
	honesty	–	base	28.7	79.6	51.8	31.7	35.5	66.6	73.7	78.4	71.0	77.2	86.5	65.2
		SFT	full	26.8	38.4	34.9	20.7	24.8	40.6	31.3	26.0	48.3	8.2	62.0	33.5
			lora	28.0	79.4	51.9	32.3	35.5	66.6	73.9	79.4	70.5	**77.0**	83.8	65.0
			P2D	37.2	**79.6**	**52.0**	31.7	**36.5**	65.6	73.7	78.8	**71.9**	76.5	**86.3**	**65.3**
			Ours	**47.0**	79.5	51.5	**32.9**	35.9	**66.8**	**74.2**	**80.0**	70.9	70.6	84.7	64.7
		DPO	full	22.0	0.0	0.0	19.5	30.8	1.9	0.0	0.0	0.0	1.1	31.5	8.5
			lora	34.1	**79.9**	**52.0**	32.9	36.0	**66.8**	**74.0**	78.8	70.4	77.5	80.9	**64.9**
			P2D	42.7	76.7	48.7	33.4	34.1	63.7	70.6	77.6	68.9	75.6	**84.2**	63.3
			Ours	**51.2**	78.6	50.6	**35.4**	**35.5**	66.4	73.5	76.0	**70.6**	74.1	80.0	64.1
OLMo-2-13B	harmless	–	base	23.0	82.8	54.7	15.9	26.4	67.0	78.5	85.2	75.4	81.8	85.0	65.3
		SFT	full	67.0	82.4	**55.7**	13.4	26.8	66.0	78.5	83.6	74.5	64.6	**86.8**	63.2
			lora	53.0	**82.8**	55.2	15.9	27.0	66.8	78.5	**85.0**	**75.5**	81.5	85.5	65.4
			P2D	65.0	82.5	54.4	**17.7**	**27.4**	**67.1**	78.5	84.8	75.4	**82.0**	85.9	**65.6**
			Ours	**70.0**	82.6	55.0	17.1	26.5	66.8	**78.8**	**85.0**	75.2	81.7	85.7	65.4
		DPO	full	**100.0**	79.9	49.1	14.6	**27.9**	64.1	74.4	82.0	71.5	74.4	78.7	61.7
			lora	**100.0**	**82.8**	53.9	14.0	27.1	66.7	78.3	85.6	**75.2**	79.2	84.0	**64.7**
			P2D	**100.0**	82.6	**54.5**	14.6	26.5	66.6	76.7	84.0	74.8	**80.5**	84.1	64.5
			Ours	**100.0**	82.3	51.3	**15.2**	27.5	**67.1**	**78.3**	**85.6**	74.7	79.5	**84.4**	64.6
	honesty	–	base	46.3	82.8	54.7	15.9	26.2	67.0	78.5	85.2	75.5	81.8	85.0	65.3
		SFT	full	50.6	81.7	53.9	**19.5**	**29.2**	66.6	74.1	84.2	74.1	63.3	80.5	62.7
			lora	47.6	**82.8**	55.0	15.9	26.8	67.0	**78.7**	85.2	75.3	81.3	85.3	65.3
			P2D	54.9	82.7	**55.2**	17.1	26.6	66.9	78.5	85.0	**75.4**	81.3	**85.9**	**65.5**
			Ours	**56.7**	82.7	55.0	16.5	26.9	**67.1**	78.5	**85.8**	75.2	**81.8**	85.0	**65.5**
		DPO	full	62.8	82.3	52.2	11.0	24.4	66.5	76.1	82.8	73.2	73.2	77.5	61.9
			lora	53.7	**83.0**	**55.3**	17.1	27.0	66.5	78.6	85.0	**75.5**	**81.2**	81.8	65.1
			P2D	56.1	82.8	55.0	16.5	27.1	**67.1**	**78.9**	**85.6**	75.1	81.1	**85.7**	**65.5**
			Ours	**65.9**	82.7	53.4	**18.3**	**27.8**	66.9	78.7	85.0	74.6	80.5	83.8	65.2

SFT (+17.3 points) and honesty under both SFT (+19.0 points) and DPO (+16.4 points), highlighting the importance of head selection for effective adaptation. Functional selection also improves the non-target average in five of seven settings. The relatively strong harmlessness performance of random selection may reflect smaller differences in functional importance scores across heads, making the choice of heads less consequential for this capability. For honesty, FTA substantially outperforms random selection on the target metric under both SFT and DPO, with gains of 19.0 and 16.4 points, while incurring only small reductions of 0.9 and 0.7 points in non-target performance, respectively. Notably, random-head updates also suppress forget-set responses but cause greater non-target degradation. This may be because MLPs and attention jointly support factual recall (Geva et al., 2023), random-head updates may suppress target responses by disrupting knowledge retrieval rather than erasing stored knowledge, while also impairing computations shared with non-target capabilities. Overall, these results support functional relevance as an effective criterion for determining where to adapt, beyond simply restricting updates to a small subset of heads. We further evaluate FTA under different top-k selection ratios in Appendix A.5, supporting the effectiveness of our strategy.

Table 3: Target knowledge suppression and non-target capability retention after unlearning on TOFU `forget05` using GradDiff. Lower forget-set ROUGE-L indicates stronger suppression. Base models serve as references.

Model	Method	Target	Non-Target										
		Forget↓	ARC-C	LogiQA2	HEval+	CRUX	MMLU	CSQA	OBQA	SIQA	SQuAD	BoolQ	**Avg**
Llama-3.1-8B	base	98.6	81.3	54.1	46.3	32.8	67.3	76.2	85.0	72.0	56.2	88.0	65.9
	full	0.3	59.4	45.2	15.2	18.9	57.0	65.7	57.4	66.8	9.9	68.6	46.4
	lora	5.4	**78.5**	48.9	42.7	26.8	**64.2**	**72.8**	**80.0**	**70.6**	34.6	**86.3**	60.5
	P2D	1.5	26.0	32.6	28.0	25.9	16.6	19.9	25.8	34.8	17.4	52.8	28.0
	Ours	**0.0**	75.3	**52.2**	**46.3**	**31.6**	59.7	69.7	77.0	66.9	**48.6**	83.0	**61.0**

Training Efficiency. Table 5 compares the computational and storage costs of mathematical SFT on Llama-3.1-8B-Instruct. FTA updates 87M parameters (around 1% of the model) and runs on a single GPU with 33.4 GB peak memory, whereas full fine-tuning uses two GPUs with 85.8 GB combined peak memory. FTA also reduces end-to-end wall-clock time from 19.6 to 13.7 minutes and GPU usage from 39.3 to 13.7 GPU-minutes, while achieving higher target and non-target scores.

Table 4: Functional vs. random head selection on Llama-3.1-8B. Target reports task-specific performance; Non-Target Avg averages scores across 10 benchmarks. Δ = Ours − Random.

Setting	Target			Non-Target Avg		
	Ours	Random	Δ	Ours	Random	Δ
Math (SFT)	**81.0**	63.7	+17.3	**68.3**	62.6	+5.7
Math (DPO)	**79.6**	77.6	+2.0	**66.5**	63.9	+2.6
Harmless (SFT)	**99.0**	93.0	+6.0	**63.9**	62.8	+1.1
Harmless (DPO)	100.0	100.0	+0.0	**62.8**	61.8	+1.0
Honesty (SFT)	**47.0**	28.0	+19.0	64.7	**65.6**	−0.9
Honesty (DPO)	**51.2**	34.8	+16.4	64.1	**64.8**	−0.7
Unlearn. (GradDiff)↓	**0.0**	0.3	−0.3	**61.0**	57.2	+3.8

Compared with LoRA ($r = 8$), FTA updates 12.8× more parameters yet uses 19.1% less peak GPU memory and runs 15.4% faster. Increasing LoRA's rank improves target performance, but none of the evaluated adaptation baselines surpasses full fine-tuning on the target average, whereas FTA reaches 81.0; results for additional LoRA ranks are reported in Appendix A.5. Against parameter-matched LoRA ($r = 102$), FTA also reduces peak memory by 9.2% and runtime by 20.3%, while achieving higher target (81.0 vs. 69.2) and non-target (68.3 vs. 65.4) scores. Its 1.3 GB head-slice delta is also substantially smaller than a full checkpoint, though larger than LoRA adapters. Thus, FTA provides a favorable balance of adaptation quality, memory use, and runtime, while LoRA remains more storage-efficient.

Table 5: Training efficiency comparision on Llama-3.1-8B-Instruct with UltraInteract-math, measured on A100 80GB GPUs.

Method	Trainable params	#GPUs	Peak mem. (GB)	Wall-clock (min)	GPU·min	Storage	Target Avg	Non-Target Avg
full	8.03 B (100%)	2	85.8	19.6	39.3	15 GB	71.6	56.7
lora (r=8)	**6.8 M (0.08%)**	**1**	41.3	16.2	16.2	**43 MB**	44.2	67.1
lora (r=102)	86.9 M (1.08%)	**1**	36.8	17.2	17.2	348 MB	69.2	65.4
P2D	124.3 M (1.55%)	**1**	34.1	14.3	14.3	1.2 GB	66.7	59.5
Ours	87 M (1.08%)	**1**	**33.4**	**13.7**	**13.7**	1.3 GB	**81.0**	**68.3**
Ours vs. full	92× ↓	2× ↓	2.6× ↓	1.4× ↓	2.9× ↓	11.5× ↓	+9.4	+11.6

6 Conclusion

We investigated post-training plasticity by examining how the functional roles of attention heads evolve during post-training. Our findings show that post-training primarily strengthens and reorganizes functional structures already present in pretrained models, motivating adaptation guided by this intrinsic organization. Building on these insights, we introduced Function-Targeted Adaptation, which restricts parameter updates to components associated with a target capability. Across multiple LLMs and capabilities, FTA supports both capability enhancement and suppression while substantially reducing interference with non-target capabilities. Together, these results connect mechanistic understanding with practical model adaptation, showing that functional localization can inform not only how we interpret post-training, but also where parameter updates should occur.

AI USE STATEMENT

We used generative AI tools to improve the manuscript's clarity and readability, assist with finding related work, and support code development and debugging. The authors reviewed and revised the AI-assisted text, checked the cited references for accuracy and relevance, and reviewed and validated the AI-assisted code. We take full responsibility for the final content of this work, including all text, code, citations, and claims.

REPRODUCIBILITY STATEMENT

The main text describes our method and experimental setup, with additional implementation details, hyperparameter settings, and evaluation procedures provided in the appendix. We identify the models and datasets used in our experiments to facilitate replication. Upon acceptance, we will release the source code.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Ruisheng Cao, Mouxiang Chen, Jiawei Chen, Zeyu Cui, Yunlong Feng, Binyuan Hui, Yuheng Jing, Kaixin Li, Mingze Li, Junyang Lin, et al. Qwen3-coder-next technical report. *arXiv preprint arXiv:2603.00729*, 2026.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37:55005–55029, 2024.

Hao Chen, Qi Zhang, Liyao Li, Zhanming Shen, Wentao Ye, Lirong Gao, Ningtao Wang, Xing Fu, Xiaoyu Shen, and Junbo Zhao. From parameters to data: A task-parameter-guided fine-tuning pipeline for efficient llm alignment. *arXiv preprint arXiv:2605.21558*, 2026a.

Hao Chen, Qi Zhang, Liyao Li, Zhanming Shen, Wentao Ye, Lirong Gao, Ningtao Wang, Xing Fu, Xiaoyu Shen, and Junbo Zhao. From parameters to data: A task-parameter-guided fine-tuning pipeline for efficient llm alignment. *arXiv preprint arXiv:2605.21558*, 2026b.

Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *Advances in Neural Information Processing Systems*, 38:57654–57689, 2026c.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of machine learning research*, 25(70):1–53, 2024.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

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*, 2021.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, 2023.

Alex Gu, Baptiste Rozière, Hugh Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida I Wang. Cruxeval: A benchmark for code reasoning, understanding and execution. *arXiv preprint arXiv:2401.03065*, 2024.

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

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*, 2020.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 3045–3059, 2021.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024a.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. Gsm-plus: A comprehensive benchmark for evaluating the robustness of llms as mathematical problem solvers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2961–2984, 2024b.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pp. 4582–4597, 2021.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2947–2962, 2023a.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Advances in neural information processing systems*, 36:21558–21572, 2023b.

Jinrui Liu, Jeff Wu, Xuanguang Pan, Gavin Cheung, Shuai Ma, and Chongyang Tao. Air: Post-training data selection for reasoning via attention head influence. *arXiv preprint arXiv:2512.13279*, 2025.

Tiange Luo, Lajanugen Logeswaran, Jaekyeom Kim, Justin Johnson, and Honglak Lee. Selective lora for visual tokens and attention heads. *arXiv preprint arXiv:2512.19219*, 2025.

Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. Reasoning models can be effective without thinking. *arXiv preprint arXiv:2504.09858*, 2025.

Xueqi Ma, Jun Wang, Yanbei Jiang, Sarah Monazam Erfani, Tongliang Liu, and James Bailey. Cognitive mirrors: Exploring the diverse functional roles of attention heads in llm reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Eleanor A Maguire, David G Gadian, Ingrid S Johnsrude, Catriona D Good, John Ashburner, Richard SJ Frackowiak, and Christopher D Frith. Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences*, 97(8):4398–4403, 2000.

Eleanor A Maguire, Katherine Woollett, and Hugo J Spiers. London taxi drivers and bus drivers: a structural mri and neuropsychological analysis. *Hippocampus*, 16(12):1091–1101, 2006.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*, 2024.

Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. A diverse corpus for evaluating and developing english math word problem solvers. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pp. 975–984, 2020.

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 *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2381–2391, 2018.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

OpenAI. Introducing GPT-5.4. `https://openai.com/index/introducing-gpt-5-4/`, March 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Yein Park, Minbyul Jeong, and Jaewoo Kang. Thinking sparks!: Emergent attention heads in reasoning models during post training. *arXiv preprint arXiv:2509.25758*, 2025.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are nlp models really able to solve simple math word problems? In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 2080–2094, 2021.

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.

Neel Rajani, Aryo Pradipta Gema, Seraphina Goldfarb-Tarrant, and Ivan Titov. Scalpel vs. hammer: Grpo amplifies existing capabilities, sft replaces them. *arXiv preprint arXiv:2507.10616*, 2025.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 2383–2392, 2016.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 4463–4473, 2019.

Yi-Lin Sung, Varun Nair, and Colin A Raffel. Training neural networks with fixed sparse masks. *Advances in neural information processing systems*, 34:24193–24205, 2021.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality, 2024. *URL https://arxiv. org/abs/2404*, 15574.

Yibo Yang, Sihao Liu, Chuan Rao, Bang An, Tiancheng Shen, Philip HS Torr, Ming-Hsuan Yang, and Bernard Ghanem. Dynamic context-oriented decomposition for task-aware low-rank adaptation with less forgetting and faster convergence. *arXiv preprint arXiv:2506.13187*, 2025.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Boji Shan, Zeyuan Liu, Jia Deng, Huimin Chen, Ruobing Xie, et al. Advancing llm reasoning generalists with preference trees. In *International Conference on Learning Representations*, volume 2025, pp. 24897–24919, 2025.

Honglin Zhang, Qianyue Hao, Fengli Xu, and Yong Li. Reinforcement learning fine-tuning enhances activation intensity and diversity in the internal circuitry of llms. In *International Conference on Learning Representations*, volume 2026, pp. 63715–63733, 2026.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023.

Qinghua Zhao, Xueling Gong, Xinyu Chen, Zhongfeng Kang, and Xinlu Li. A layer-wise analysis of supervised fine-tuning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9981–9992, 2026.

Yang Zhao, Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. Supervised fine-tuning achieve rapid task adaption via alternating attention head activation patterns. *arXiv preprint arXiv:2409.15820*, 2024.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. On the role of attention heads in large language model safety. *arXiv preprint arXiv:2410.13708*, 2024.

A Appendix

A.1 More details of localizing capability-relevant heads

Identifying Related Heads for Unlearning on TOFU. Unlike harmlessness alignment and mathematical reasoning, TOFU does not provide paired positive and negative responses. We therefore identify unlearning-related heads directly from the forget-set objective. Specifically, for each of the 200 question–answer pairs in the TOFU `forget05` split (Maini et al., 2024), we compute the

cross-entropy loss over the target answer tokens on the TOFU-finetuned model. We score each attention head using the magnitude of its activation–gradient product, $|\mathrm{Act} \times \mathrm{Grad}|$, measured at the input to the output projection. The scores are summed over token positions and head dimensions and then averaged across all 200 examples. We rank heads by these importance scores and apply the same elbow criterion as in the paired-response setting, selecting 100 heads (9.8% of all heads). GradDiff updates are then restricted to these selected heads, which are most strongly associated with producing the designated forget-set answers.

MLP We train a two-layer multi-class MLP for probing. The first layer applies a shared linear projection to each multi-head representation vector, reducing each to a 64-dimensional embedding. These embeddings are then flattened and concatenated into a single vector of size $64 \times numberofheads$. This vector is fed into a hidden layer with 512 units, followed by a ReLU activation and a dropout with a rate of 0.3. The final output layer maps the 512-dimensional hidden representation to the set of cognitive function labels. The model is trained using the Adam optimizer with a learning rate of 10^{-4} and a cross-entropy loss. Training proceeds for 100 epochs.

Change Patterns. Heatmaps of post-training plasticity in Figure 3, showing changes in attention-head importance scores relative to the base model. Importance scores increase for most heads during post-training, with the largest gains concentrated in a small subset of top-k heads.

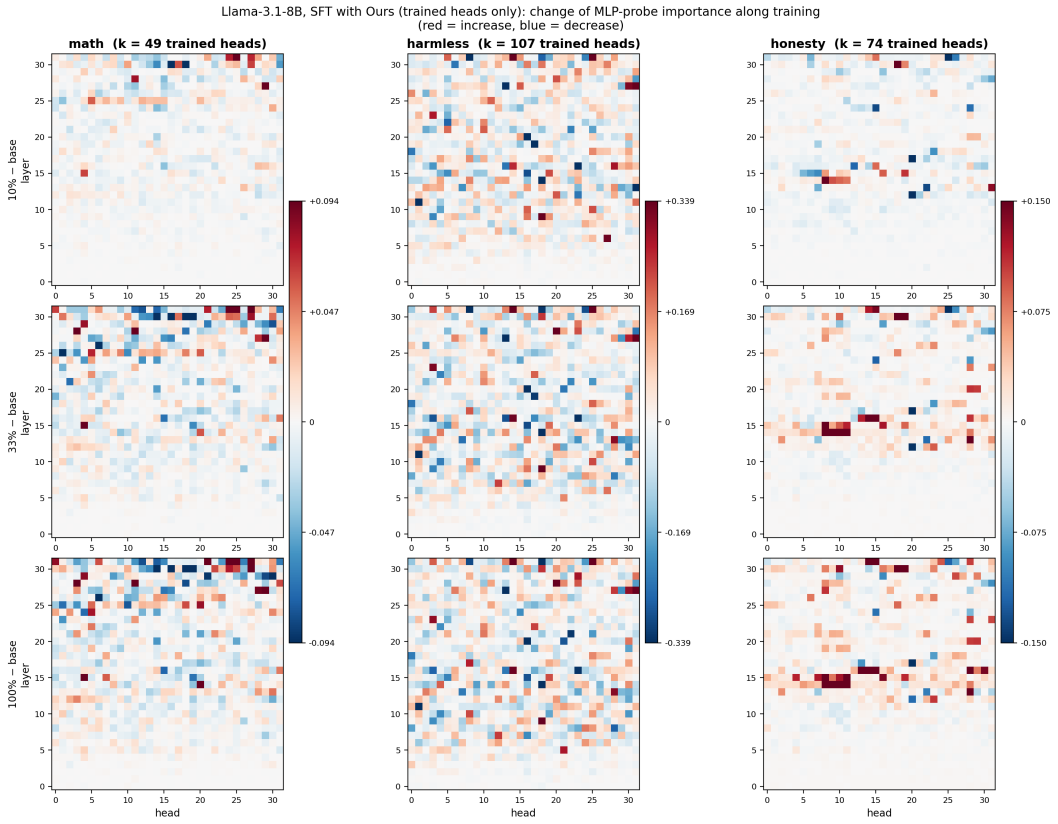

Figure 3: Heatmap of post-training plasticity. Changes in attention-head importance scores are measured relative to the base model.

We ranked the importance scores in descending order and identified the elbow point in the resulting curve, as illustrated in Figure 4.

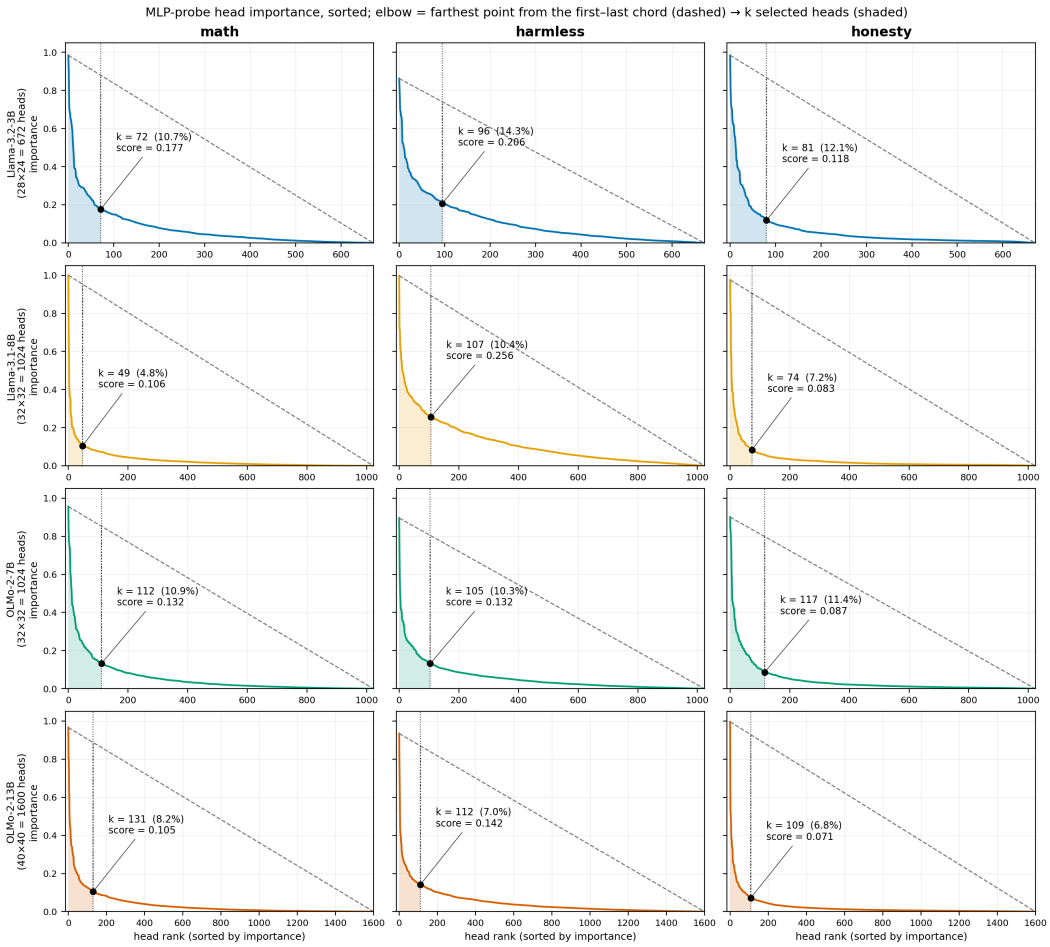

Figure 4: Ranked attention-head importance scores for mathematics, harmlessness, and honesty across Llama and OLMo models.

A.2 Training and Test Splits for Probing.

Capability-associated heads are identified on the same model that is subsequently adapted: Llama-3.1-8B-Instruct for mathematical reasoning, and Llama-3.1-8B-Base for harmlessness and honesty. Each probing example is constructed from a paired response, yielding two sequences—context + chosen/correct (label 1) and context + rejected/incorrect (label 0)—using the same prompt format as the corresponding adaptation data.

Math. Pairs are drawn from the UltraInteract-math training split. After a fixed-seed shuffle, the first 3,000 pairs are used for probe training and the next 600 for validation, corresponding to 6,000 and 1,200 sequences, respectively. Dialogues are formatted with the model's chat template; the instruction "Put your final answer in `\boxed{}`" is appended to the final user turn, and both responses are normalized to place the final answer in `\boxed{}`. Sequences are truncated to 1,024 tokens.

Harmlessness. Pairs are drawn from Anthropic HH-RLHF and restricted to single-turn dialogues containing exactly one human turn in both the chosen and rejected responses. The HH-RLHF training split provides 49,436 probe-training pairs, while the test split provides 2,612 validation pairs, corresponding to 98,872 and 5,224 sequences. We retain the native `Human:`/`Assistant:` format.

Honesty. We split the 817 TruthfulQA generation questions into 653 training and 164 held-out test questions using a fixed seed (80/20). Probing uses only the 653 training questions: the first 555 are

used for probe training and the remaining 98 for validation. For each question, the positive sequence combines the `Human:/Assistant:` scaffold with the reference best answer, while the negative sequence uses one incorrect answer sampled from the provided incorrect-answer list. Sequences are truncated to 1,024 tokens. The 164 held-out test questions are never used during probing.

A.3 Evaluation Metrics

Let H denote the total number of attention heads ($H = 1024$ for Llama-3.1-8B), and let $I_h^{(t)}$ denote the importance score of head h at training progress $t \in \{0, 10, 33, 66, 100\}\%$, where $t = 0$ corresponds to the initial model. We define the initial capability-associated head set as

$$\mathcal{S}_0 = \mathrm{top}_k\Big(I^{(0)}\Big), \tag{6}$$

where k is determined by the elbow point of the initial ranked importance curve ($k = 49$, 116, and 100 for mathematical reasoning, harmlessness, and unlearning, respectively).

Head Retention. To measure how stable the initial capability-associated heads remain during post-training, we compute the fraction of $\mathcal{S}_0$ that remains within the top-k or top-$2k$ heads at checkpoint t:

$$\mathrm{Ret}_k(t) = \frac{\big|\mathcal{S}_0 \cap \mathrm{top}_k(I^{(t)})\big|}{k}, \qquad \mathrm{Ret}_{2k}(t) = \frac{\big|\mathcal{S}_0 \cap \mathrm{top}_{2k}(I^{(t)})\big|}{k}. \tag{7}$$

$|\Delta I|$ Concentration. To assess whether post-training-induced importance changes are preferentially concentrated on the initial capability-associated heads, we first normalize the importance scores at each checkpoint and compute their change relative to the initial model:

$$\tilde{I}_h^{(t)} = \frac{I_h^{(t)}}{\max_{h'} |I_{h'}^{(t)}|}, \qquad \Delta I_h^{(t)} = \tilde{I}_h^{(t)} - \tilde{I}_h^{(0)}. \tag{8}$$

We then treat $|\Delta I_h^{(t)}|$ as a score for predicting whether head h belongs to $\mathcal{S}_0$ and compute the corresponding rank-based AUROC:

$$\begin{aligned}\mathrm{AUROC}(t) = \frac{1}{k(H-k)} \sum_{i \in \mathcal{S}_0} \sum_{j \notin \mathcal{S}_0} \Big(& \mathbb{1}[|\Delta I_i^{(t)}| > |\Delta I_j^{(t)}|] \\ & + \tfrac{1}{2}\mathbb{1}[|\Delta I_i^{(t)}| = |\Delta I_j^{(t)}|] \Big).\end{aligned} \tag{9}$$

Equivalently, using the Mann–Whitney rank formulation,

$$\mathrm{AUROC}(t) = \frac{\sum_{i \in \mathcal{S}_0} r_i - \frac{k(k+1)}{2}}{k(H-k)}, \tag{10}$$

where r_i is the ascending rank of $|\Delta I_i^{(t)}|$ among all H heads. An AUROC of 0.5 indicates no preferential concentration of importance changes on $\mathcal{S}_0$, whereas values approaching 1 indicate increasingly stronger concentration on the initial capability-associated heads.

For complementary analysis, we also compute the ratio $\overline{|\Delta I|}_{\mathcal{S}_0} / \overline{|\Delta I|}_{\bar{\mathcal{S}}_0}$ and the signed mean change $\frac{1}{k}\sum_{i \in \mathcal{S}_0} \Delta I_i^{(t)}$, which quantify the relative magnitude and direction of importance shifts, respectively.

A.4 Experiment Details

Optimization and Hyperparameters. All methods are trained for one epoch with AdamW using the same data and training schedule. SFT uses an effective batch size of 16 sequences ($4\times$ gradient accumulation; maximum length 1024), while DPO uses 16 preference pairs per update (2×8; maximum length 1024), $\beta = 0.1$, and the frozen base model as reference. The learning rate is 10^{-5} for both SFT and DPO across methods, except SFT LoRA, for which 5×10^{-5} performs better than 10^{-5} and is therefore reported. For all competing methods, we use their best-performing configurations after method-specific tuning to ensure a fair comparison.

More details for non-target evaluation. Non-target retention is evaluated on ten benchmarks shared across all domains and methods, covering scientific and logical reasoning (ARC-C (Clark et al., 2018), LogiQA2 (Liu et al., 2023a)), code (HumanEval+ (Liu et al., 2023b), CRUXEval (Gu et al., 2024)), knowledge (MMLU (Hendrycks et al., 2020)), commonsense (CSQA (Talmor et al., 2019), OBQA (Mihaylov et al., 2018), SIQA (Sap et al., 2019)), and reading comprehension (SQuAD (Rajpurkar et al., 2016), BoolQ (Clark et al., 2019)).

A.5 More Results

Detailed comparisons between FTA and random heads for each capability. Tables 11, 12, and 13 compare FTA with random-head adaptation on mathematical reasoning, value alignment, and knowledge unlearning, respectively. In each setting, the random baseline adapts the same number of heads under an identical training protocol. The two methods therefore differ only in which heads are updated.

Mathematical reasoning. Under SFT, the gap is large. FTA outperforms random heads by 17.3 points on the target average (81.0 vs. 63.7) and by 5.7 points on the non-target average (68.3 vs. 62.6). Random-head SFT not only fails to improve math but degrades it well below the base model. It also causes severe collateral damage on LogiQA2 (-15.8) and SQuAD (-20.6) relative to FTA. Under DPO, the target gap narrows to 2.0 points (79.6 vs. 77.6), though FTA remains clearly ahead on the hardest benchmark, GSM-Plus (70.0 vs. 63.4). FTA also better preserves non-target capabilities (66.5 vs. 63.9), with the largest margin again on SQuAD (66.5 vs. 58.4).

Value alignment. For harmlessness, FTA reaches a higher safe rate under SFT (99.0% vs. 93.0%). Under DPO, both methods saturate at 100%. In both cases FTA retains slightly more general capability (+1.1 and +1.0 non-target Avg), so even when the target is saturated, selecting capability-relevant heads reduces interference with unrelated functions. For honesty, the difference is most pronounced on the target: FTA improves honesty by 19.0 points under SFT (47.0% vs. 28.0%) and 16.4 points under DPO (51.2% vs. 34.8%). Random heads retain marginally higher non-target averages here ($<$1 point, mainly from SQuAD). This likely reflects that random-head updates barely modify honesty-related behavior, and thus also leave the rest of the model largely untouched.

Knowledge unlearning. Random-head updates also achieve effective unlearning, with forget-set ROUGE-L near zero for both methods (0.3 vs. 0.0). However, they cause substantially greater degradation of non-target capabilities (57.2 vs. 61.0 Avg), with the largest drops on LogiQA2 (-9.5), HumanEval+ (-8.5), SQuAD (-8.4), and CRUXEval (-6.1). This may be because factual recall is distributed across components: MLPs enrich subject representations while attention heads extract the relevant attributes. Perturbing arbitrary heads can therefore disrupt recall, but it also interferes with heads serving unrelated functions such as reasoning and code generation.

Effect of the Updated-Head Ratio. We vary the proportion of attention heads updated by FTA to examine sensitivity to the adaptation scope. As shown in Table 6, updating more heads does not consistently improve performance. For math, the elbow-selected 4.8% achieves the highest target score while also giving the best non-target average, whereas increasing the ratio to 10% or 15% substantially reduces target performance. For harmlessness, the elbow selection achieves the highest target score, while smaller fixed ratios slightly improve non-target retention. For honesty, increasing the ratio to 15% further improves the target metric but incurs a larger non-target drop, whereas the elbow selection provides stronger retention. Overall, these results show that FTA's gains are not simply due to updating more parameters and support the elbow-based selection as an effective way to balance target adaptation and non-target retention.

Robustness to Generation Sampling. To assess robustness to generation sampling, we evaluate five sampled generations per method (Table 14). FTA achieves the best mean target performance among the post-trained methods across all settings and both models, tying with full fine-tuning and LoRA on Harmless DPO, where all reach 100% safe responses. For unlearning, FTA reaches zero forget-set ROUGE-L. Non-target results are more mixed: FTA often matches or exceeds the baselines, although it does not lead in every setting. Despite some variation across samples, particularly on honesty and harmlessness, the target-performance advantage remains consistent. These results demonstrate that FTA consistently improves target performance across sampled generations while generally preserving non-target capabilities.

Table 6: Head-ratio ablation on Llama-3.1-8B under SFT. We compare the elbow-selected head set with fixed top-5%, 10%, and 15% subsets under the same training and evaluation setup. Target metrics are math4 Avg, JailbreakBench Safe%, and TruthfulQA Honest%; non-target performance is averaged over ten benchmarks. Best results within each domain are in **bold**.

			Target	Non-target										
Domain	Heads	k		ARC-C	LogiQA2	HEval+	CRUX	MMLU	CSQA	OBQA	SIQA	SQuAD	BoolQ	**Avg**
math	elbow (4.8%)	49	**81.0**	81.9	**56.1**	63.4	30.0	**63.2**	**76.5**	83.2	73.5	71.1	84.0	**68.3**
	5%	52	**81.0**	**82.0**	55.3	**64.6**	29.6	62.6	**76.5**	**83.6**	72.8	70.5	84.0	68.2
	10%	103	74.1	80.8	55.2	62.2	30.5	59.9	75.8	83.4	73.6	72.2	84.2	67.8
	15%	154	63.1	80.3	53.8	64.0	**30.6**	58.4	75.6	82.8	**74.0**	**73.0**	**84.7**	67.7
harmless	elbow (10.4%)	107	**99.0**	78.6	49.1	28.7	34.9	66.3	73.6	78.6	69.5	74.5	85.1	63.9
	5%	52	94.0	79.0	**51.0**	28.7	**35.8**	**67.3**	**73.9**	78.4	**70.0**	**74.9**	86.7	**64.6**
	10%	103	98.0	**79.4**	49.2	28.7	35.2	66.8	**73.9**	78.2	69.0	74.2	**87.0**	64.2
	15%	154	98.0	78.0	49.9	**29.3**	34.5	65.9	73.3	**78.8**	69.7	74.3	85.8	63.9
honesty	elbow (7.2%)	74	47.0	79.5	51.5	**32.9**	35.9	**66.8**	**74.2**	**80.0**	70.9	70.6	**84.7**	**64.7**
	5%	52	41.5	79.9	51.3	**32.9**	35.9	66.5	73.8	78.6	71.1	70.2	83.4	64.4
	10%	103	44.5	79.1	**51.8**	30.5	**36.2**	66.5	74.0	79.6	**71.3**	**71.7**	83.9	64.5
	15%	154	**54.9**	**80.0**	50.3	31.1	35.2	66.7	73.0	79.6	70.9	65.9	82.9	63.6

Table 7: Probe-seed robustness of functional-head selection (Llama-3.1-8B). For each seed pair we report the Spearman ρ between the two 1024-head importance vectors ($|$normalised Grad$\times$Act$|$, the selection score), the overlap of the two top-k sets at the paper's elbow k ($|\mathrm{top}_k(A) \cap \mathrm{top}_k(B)|/k$; random $= k/1024$), and the Jaccard index of the two sets each cut at its own seed's elbow.

Domain (k)	Seed pair	Spearman ρ	Overlap@k	Jaccard (own elbow)
Math (49)	0 vs 1	0.982	0.980	0.948
	0 vs 2	0.987	0.959	0.932
	1 vs 2	0.986	0.980	0.947
	mean	**0.985**	**0.973**	**0.943**
Harmless (107)	0 vs 1	0.805	0.757	0.527
	0 vs 2	0.790	0.785	0.638
	1 vs 2	0.799	0.776	0.579
	mean	**0.798**	**0.773**	**0.582**
Honesty (74)	0 vs 1	0.986	0.973	0.919
	0 vs 2	0.989	0.986	0.918
	1 vs 2	0.994	0.973	0.947
	mean	**0.990**	**0.977**	**0.928**

Comparison across LoRA ranks. Table 17 tests whether FTA's advantage over LoRA comes from its larger trainable-parameter budget. We increase the LoRA rank from r=8 (6.8 M) and r=16 (13.6 M) to r=102, which matches FTA's budget (86.9 M vs. 87.0 M).

Higher rank helps in-domain, but does not close the gap. Raising the rank steadily improves in-domain math accuracy, from 44.2 (r=8) to 49.6 (r=16) and 69.2 (r=102). Even so, every LoRA configuration stays well below the base model (80.4). At an equal parameter budget, LoRA (r=102) still trails FTA by 11.8 points ID Avg (69.2 vs. 81.0) and is behind on all four ID benchmarks. FTA is the only method that improves over the base model on the ID average (+0.6).

Higher rank hurts out-of-domain generalization. OOD accuracy declines monotonically as the rank increases: 67.1 (r=8), 66.3 (r=16), and 65.4 (r=102), all below the base model (67.5). The largest drop is on CRUXEval, which falls from 30.9 to 11.2 at r=102. This suggests that adding unconstrained low-rank capacity to all q/k/v/o projections increasingly overwrites general capabilities. By contrast, FTA raises OOD Avg to 68.3 (+0.8 over base) and keeps the best OOD average among all trained methods.

Takeaway. LoRA shows a clear trade-off: more capacity buys in-domain gains at the cost of out-of-domain generalization. FTA improves both at the same budget. Its benefit therefore comes from *where* the parameters are placed, in the selected functional head slices, rather than from *how many* are trained.

Table 8: Target and non-target capability performance (%) of Llama-3.2-3B and OLMo-2-7B after post-training for mathematical reasoning.

Model	Train	Method	Target					Non-Target									
			GSM8K	SVAMP	ASDiv	GSM+	**Avg**	ARC-C	LogiQA2	HEval+	CRUX	MMLU	CSQA	OBQA	SIQA	SQuAD	BoolQ	**Avg**
Llama-3.2-3B	–	base	53.9	64.4	65.0	51.0	58.6	77.6	48.0	51.8	18.9	58.5	73.7	77.0	66.2	64.8	71.4	60.8
	SFT	full	65.0	83.0	80.1	47.5	68.9	63.2	36.3	**0.0**	7.8	47.9	66.8	65.0	66.4	**69.6**	**76.8**	50.0
		lora	81.3	84.9	81.5	65.8	78.4	74.3	45.5	**0.0**	4.1	56.2	**73.7**	76.8	50.9	62.4	71.0	51.5
		Ours	**69.1**	**84.2**	**80.3**	**55.3**	**72.2**	**77.5**	**46.8**	**0.0**	**15.2**	**59.3**	73.3	**77.2**	**66.9**	65.6	71.2	**55.3**
	DPO	full	73.7	81.1	81.8	58.7	73.8	**70.4**	43.8	46.7	16.5	**54.0**	**70.0**	68.4	64.7	59.0	**72.2**	56.6
		lora	76.5	81.5	82.9	**62.7**	75.9	61.4	43.4	**59.8**	15.9	50.5	68.4	63.0	63.6	**69.4**	71.5	56.7
		Ours	**77.1**	**85.4**	**84.6**	60.9	**77.0**	67.9	**45.5**	54.3	**18.9**	52.6	69.7	**72.4**	**66.9**	51.7	70.8	**57.1**
OLMo-2-7B	–	base	38.2	71.6	71.0	25.8	51.7	67.5	43.3	32.9	11.5	49.5	73.6	73.0	72.4	55.9	82.4	56.2
	SFT	full	62.7	79.3	72.7	**54.2**	67.2	75.5	44.9	**32.9**	12.2	**55.5**	75.5	76.8	72.0	**69.8**	**84.7**	**60.0**
		lora	61.8	74.2	75.7	45.7	64.4	**77.8**	**46.6**	25.4	**13.0**	54.1	**76.7**	**77.0**	72.2	67.8	84.5	59.5
		Ours	**71.3**	**79.6**	**77.2**	53.6	**70.4**	75.9	**46.6**	25.0	10.5	54.5	**76.7**	75.0	**72.5**	69.3	84.5	59.1
	DPO	full	70.8	**79.1**	75.9	58.7	71.1	67.1	45.1	31.6	8.3	51.3	73.6	**74.6**	70.3	39.5	80.2	54.2
		lora	72.8	76.3	75.1	60.4	71.2	**75.7**	48.3	14.6	7.4	**56.2**	**74.9**	73.0	70.7	30.8	81.3	53.3
		Ours	**73.2**	**79.1**	**77.2**	**60.9**	**72.6**	72.4	**49.6**	**33.5**	**10.3**	54.3	73.1	70.8	**71.8**	**49.0**	**81.9**	**56.7**

Table 9: Target and non-target performance (%) of Llama-3.2-3B and OLMo-2-7B under harmlessness and honesty alignment via SFT and DPO. Target metrics are Safe% on JailbreakBench and Honest% on TruthfulQA. Best results within each model's SFT or DPO group are **bolded**.

Model	Value	Train	Method	Target	Non$_t arget$										
				Safe/Hon%	ARC-C	LogiQA2	HEval+	CRUX	MMLU	CSQA	OBQA	SIQA	SQuAD	BoolQ	**Avg**
Llama-3.2-3B	harmless	–	base	34.0	68.5	44.2	25.6	28.5	57.1	65.1	67.0	64.9	71.9	80.0	57.3
		SFT	full	**95.0**	59.9	34.9	22.6	28.0	49.8	61.8	61.4	57.4	63.0	79.0	51.8
			lora	63.0	**68.1**	**44.1**	**26.2**	**28.5**	**56.9**	65.2	**67.4**	64.6	**71.2**	80.2	**57.2**
			Ours	**95.0**	67.7	43.8	24.4	27.9	56.8	**65.7**	**67.4**	**65.5**	69.2	**80.5**	56.9
		DPO	full	**100.0**	55.5	31.6	17.1	26.1	45.6	51.7	60.6	53.9	62.3	71.0	47.5
			lora	**100.0**	66.9	**43.4**	20.8	29.4	51.5	63.3	63.6	**63.7**	70.3	72.4	54.5
			Ours	**100.0**	**67.2**	42.7	**22.6**	**30.1**	**56.1**	**64.5**	**65.4**	61.5	**70.5**	**77.1**	**55.8**
	honesty	–	base	24.4	68.5	44.3	25.6	28.2	57.1	65.0	67.0	64.8	71.9	80.0	57.2
		SFT	full	26.2	58.7	35.7	22.6	**28.9**	51.4	55.0	58.0	55.7	0.2	71.8	43.8
			lora	24.4	67.7	**44.3**	25.0	28.2	**57.1**	64.9	67.4	**64.9**	**71.5**	**79.8**	**57.1**
			Ours	**32.9**	**68.5**	42.9	**25.6**	28.4	56.6	**65.6**	**67.8**	64.8	68.4	77.4	56.6
		DPO	full	22.3	62.3	41.2	**28.7**	25.1	54.2	62.2	62.8	57.6	40.9	39.8	47.5
			lora	26.8	**68.1**	44.1	25.6	28.4	57.0	**64.9**	67.0	64.7	71.7	**79.6**	**57.1**
			Ours	**29.3**	67.7	**44.5**	24.4	**28.5**	**57.2**	64.7	**67.2**	**64.9**	**72.0**	79.5	57.1
OLMo-2-7B	harmless	–	base	14.0	79.9	49.6	9.1	24.4	62.7	74.0	79.6	71.6	81.8	80.3	61.3
		SFT	full	68.0	70.5	44.2	**11.6**	23.1	60.4	72.6	74.2	65.8	69.4	**85.3**	57.7
			lora	35.0	80.1	**49.4**	10.4	**24.8**	62.8	**73.9**	**79.4**	**71.3**	**81.8**	82.0	**61.6**
			Ours	**89.0**	**80.2**	48.9	11.0	24.2	**63.0**	73.7	79.2	71.1	**81.8**	82.3	61.5
		DPO	full	97.0	66.6	38.4	7.3	22.4	56.0	64.6	64.2	52.8	59.0	79.0	51.0
			lora	96.0	**80.2**	**49.6**	**9.1**	**24.6**	62.7	**74.0**	79.4	71.2	80.1	80.0	61.1
			Ours	**100.0**	80.1	**49.6**	7.3	**24.6**	**63.1**	**74.0**	**79.6**	**71.5**	**80.6**	**80.7**	**61.1**
	honesty	–	base	42.7	79.9	49.6	9.1	24.5	62.7	73.9	79.6	71.6	81.8	80.3	61.3
		SFT	full	39.6	64.2	44.7	**12.8**	**26.5**	58.1	60.6	64.2	50.3	50.4	77.1	50.9
			lora	41.5	80.2	**49.4**	9.8	24.6	62.7	73.8	**79.4**	**71.8**	**81.7**	**80.5**	**61.4**
			Ours	**46.3**	**80.6**	**49.4**	9.1	24.9	**62.9**	**73.9**	79.0	71.6	81.2	77.8	61.0
		DPO	full	44.5	77.6	48.4	3.0	20.5	62.5	72.2	75.8	62.0	71.2	71.3	56.5
			lora	52.4	80.1	49.4	**9.1**	**24.6**	62.8	73.8	**79.0**	**71.8**	71.5	**75.0**	59.7
			Ours	**59.8**	**80.5**	**50.1**	6.7	24.5	**62.9**	**74.3**	78.6	**71.8**	**77.0**	73.6	**60.0**

Table 10: Target forgetting and non-target capability retention (%) of Llama-3.2-3B and Llama-3.1-8B on TOFU forget05 using GradDiff. Lower Forget scores indicate stronger forgetting; higher non-target scores indicate better retention. Best post-training results within each model are **bolded**.

Model	Method	Target	Non-Target										
		Forget↓	ARC-C	LogiQA2	HEval+	CRUX	MMLU	CSQA	OBQA	SIQA	SQuAD	BoolQ	**Avg**
Llama-3.2-3B	base	92.3	74.1	46.2	37.8	26.5	60.4	69.5	76.4	68.8	65.5	80.6	60.6
	full	1.0	72.0	41.1	11.6	13.8	55.3	64.2	73.2	62.2	46.8	71.3	51.2
	lora	47.7	70.9	41.2	31.0	22.1	**60.1**	66.5	74.8	**68.4**	**66.6**	79.4	58.1
	Ours	**0.0**	**72.2**	**42.7**	**32.3**	**22.9**	60.0	**67.1**	**75.0**	67.4	65.1	**79.7**	**58.4**
Llama-3.1-8B	base	98.6	81.3	54.1	46.3	32.8	67.3	76.2	85.0	72.0	56.2	88.0	65.9
	full	0.3	59.4	45.2	15.2	18.9	57.0	65.7	57.4	66.8	9.9	68.6	46.4
	lora	5.4	**78.5**	48.9	42.7	26.8	**64.2**	**72.8**	**80.0**	**70.6**	34.6	**86.3**	60.5
	Ours	**0.0**	75.3	**52.2**	**46.3**	**31.6**	59.7	69.7	77.0	66.9	**48.6**	83.0	**61.0**

Table 11: Ablation comparing capability-guided and random head selection for mathematical reasoning on Llama-3.1-8B-Instruct.

Train	Method	Target					Non-Target									
		GSM8K	SVAMP	ASDiv	GSM+	**Avg**	ARC-C	LogiQA2	HEval+	CRUX	MMLU	CSQA	OBQA	SIQA	SQuAD	BoolQ	**Avg**
SFT	Ours	**82.8**	**89.4**	**83.4**	**68.4**	**81.0**	**81.9**	**56.1**	63.4	**30.0**	**63.2**	**76.5**	**83.2**	**73.5**	**71.1**	**84.0**	**68.3**
	Random	66.7	71.8	64.8	51.6	63.7	78.2	40.3	**64.0**	28.0	56.8	75.0	80.4	69.3	50.5	83.1	62.6
DPO	Ours	**81.8**	82.3	**84.4**	**70.0**	**79.6**	**81.7**	**52.3**	**62.8**	23.9	**63.2**	**76.5**	**84.2**	**71.0**	**66.5**	**82.4**	**66.5**
	Random	80.5	**83.6**	82.7	63.4	77.6	77.8	48.9	62.2	**26.1**	59.0	75.2	81.6	70.0	58.4	80.2	63.9

Table 12: Ablation comparing capability-guided and random head selection for harmlessness and honesty alignment on Llama-3.1-8B. Target metrics are Safe (%) on JailbreakBench and Honest (%) on TruthfulQA.

Value	Train	Method	Target	Non-Target									
			Safe/Hon%	ARC-C	LogiQA2	HEval+	CRUX	MMLU	CSQA	OBQA	SIQA	SQuAD	BoolQ	**Avg**
harmless	SFT	Ours	**99.0**	**78.6**	**49.1**	28.7	**34.9**	**66.3**	73.6	**78.6**	**69.5**	**74.5**	**85.1**	**63.9**
		Random	93.0	75.3	48.1	**30.5**	34.5	64.2	**73.8**	76.6	68.4	72.4	83.8	62.8
	DPO	Ours	**100.0**	**77.7**	47.5	24.4	**36.2**	**64.8**	**72.8**	**77.4**	**66.3**	**77.8**	**83.0**	**62.8**
		Random	**100.0**	76.7	**48.3**	**27.4**	34.5	64.2	70.3	75.0	66.0	76.7	79.3	61.8
honesty	SFT	Ours	**47.0**	79.5	51.5	**32.9**	**35.9**	66.8	**74.2**	**80.0**	**70.9**	70.6	84.7	64.7
		Random	28.0	**79.8**	**51.8**	**32.9**	35.8	**67.0**	74.0	79.4	70.7	**77.5**	**86.7**	**65.6**
	DPO	Ours	**51.2**	78.6	50.6	**35.4**	35.5	**66.4**	73.5	76.0	**70.6**	74.1	80.0	64.1
		Random	34.8	**79.7**	**50.8**	33.5	**35.8**	66.3	**74.4**	**80.0**	70.2	**77.2**	**80.2**	**64.8**

Table 13: Ablation comparing capability-guided and random head selection for knowledge unlearning on Llama-3.1-8B using TOFU forget05 and GradDiff. Forgetting is measured by ROUGE-L↓.

Method	Target	Non-Target									
	Forget↓	ARC-C	LogiQA2	HEval+	CRUX	MMLU	CSQA	OBQA	SIQA	SQuAD	BoolQ	**Avg**
Ours	**0.0**	**75.3**	**52.2**	**46.3**	**31.6**	**59.7**	**69.7**	**77.0**	**66.9**	**48.6**	**83.0**	**61.0**
Random	0.3	73.1	42.7	37.8	25.5	59.4	68.2	75.0	**66.9**	40.2	**83.0**	57.2

Table 14: Robustness of main-table comparisons to generation sampling (mean±std). Target metrics are the four-benchmark math average, JailbreakBench Safe%, TruthfulQA Honest%, and TOFU forget ROUGE-L↓; non-target performance is averaged over ten benchmarks.

		Target				Non-Target Avg			
Model	Setting	base	full	LoRA	Ours	base	full	LoRA	Ours
Llama-3.1-8B	Math SFT	80.3±0.3	70.5±2.0	74.5±4.6	**80.6±0.4**	67.6±0.3	58.6±3.6	67.0±0.6	**68.6±0.3**
	Math DPO	80.7±0.3	75.6±0.6	78.5±0.4	**79.5±0.2**	67.4±0.3	54.5±0.4	65.4±0.5	**66.6±0.3**
	Harmless SFT	61.2±3.1	91.6±3.9	90.0±3.2	**96.5±3.8**	64.1±2.2	46.7±5.1	64.1±3.7	**66.2±3.0**
	Harmless DPO	60.0±3.1	**100.0±0.5**	**100.0±0.0**	**100.0±0.0**	66.7±2.2	48.4±3.7	**62.8±1.4**	62.5±2.0
	Honesty SFT	29.3±1.1	25.2±3.0	29.7±2.5	**47.5±2.9**	66.1±2.2	34.6±2.1	63.5±1.9	**64.1±1.6**
	Honesty DPO	28.2±1.1	21.2±1.6	33.5±2.1	**52.6±2.7**	66.8±2.2	8.2±0.5	63.4±2.2	**64.2±0.5**
	Unlearning (GradDiff)↓	98.6±0.4	0.3±0.0	5.4±0.2	**0.0±0.0**	65.8±0.5	46.6±0.6	60.7±0.2	**60.8±0.3**
OLMo-2-13B	Math SFT	60.3±1.5	72.0±3.5	66.6±4.1	**75.1±1.2**	61.8±0.4	56.6±1.0	**61.2±0.6**	60.6±0.4
	Math DPO	60.2±1.5	58.8±1.1	60.4±1.1	**61.4±1.7**	61.8±0.4	55.2±0.7	53.7±0.7	**59.0±0.7**
	Harmless SFT	23.7±4.6	62.1±6.3	54.2±2.7	**71.2±6.9**	67.2±2.1	63.5±3.5	**66.2±2.0**	64.7±2.2
	Harmless DPO	19.9±4.6	**100.0±0.0**	**100.0±0.0**	**100.0±0.0**	63.4±2.1	61.4±3.4	**65.7±1.7**	64.6±1.4
	Honesty SFT	49.0±2.8	51.9±3.8	45.9±1.9	**55.5±3.8**	65.5±2.1	62.0±2.0	63.4±1.9	**66.3±1.8**
	Honesty DPO	45.8±2.8	63.3±1.8	54.6±3.7	**66.0±1.5**	66.4±2.1	62.2±0.6	**67.0±2.2**	65.5±1.2

Table 15: Approximate two-sided Welch t-test p-values reconstructed from the reported means and standard deviations, assuming five independent runs per method. Full and LoRA are each compared with Ours. Values below 0.05 are **bolded**.

		ID		OOD Avg	
Model	Setting	full vs. Ours	LoRA vs. Ours	full vs. Ours	LoRA vs. Ours
Llama-3.1-8B	Math SFT	<**0.001**	<**0.001**	**0.003**	**0.002**
	Math DPO	<**0.001**	**0.003**	<**0.001**	**0.003**
	Harmless SFT	0.079	**0.020**	<**0.001**	0.354
	Harmless DPO	−−	−−	<**0.001**	0.791
	Honesty SFT	<**0.001**	<**0.001**	<**0.001**	0.604
	Honesty DPO	<**0.001**	<**0.001**	<**0.001**	0.468
	Unlearning (GradDiff)↓	<**0.001**	<**0.001**	<**0.001**	0.555
OLMo-2-13B	Math SFT	0.121	**0.008**	<**0.001**	0.105
	Math DPO	**0.025**	0.307	<**0.001**	<**0.001**
	Harmless SFT	0.061	**0.003**	0.538	0.292
	Harmless DPO	−−	−−	0.106	0.298
	Honesty SFT	0.173	**0.002**	**0.007**	**0.038**
	Honesty DPO	**0.034**	**0.0011**	**0.0016**	0.228

Table 16: Capability-guided vs. random head selection on Llama-3.1-8B. The t-tests compare Random with FTA. Better results and p-values below 0.05 are **bolded**.

	ID			OOD Avg		
Setting	Ours	Random	p	Ours	Random	p
Math (SFT)	**81.0**	62.1±8.8	**0.009**	**68.3**	64.6±2.3	**0.023**
Harmless (SFT)	**99.0**	84.0±8.5	**0.017**	**63.9**	63.1±1.3	0.260
Honesty (SFT)	**47.0**	28.6±1.0	<**0.001**	64.7	**65.4±0.1**	<**0.001**
Unlearning (GradDiff)↓	**0.0**	0.4±0.1	<**0.001**	**61.0**	58.6±3.4	0.190
Math (DPO)	**79.6**	76.8±0.8	**0.001**	**66.5**	64.4±0.5	<**0.001**
Harmless (DPO)	100.0	100.0±0.0	–	**62.8**	61.6±0.6	**0.012**
Honesty (DPO)	**51.2**	40.3±6.2	**0.017**	64.1	**64.6±0.4**	0.063

Table 17: LoRA rank ablation on Llama-3.1-8B-Instruct under the math SFT. LoRA ($r = 102$) matches FTA's parameter budget. Best results among trained methods are **bolded**.

Method	Trainable	ID					OOD										
	params	GSM8K	SVAMP	ASDiv	GSM+	**Avg**	ARC-C	LogiQA2	HEval+	CRUX	MMLU	CSQA	OBQA	SIQA	SQuAD	BoolQ	**Avg**
base	–	81.9	88.2	84.0	67.5	80.4	82.7	54.0	62.8	30.9	62.5	76.2	84.8	73.7	65.6	81.3	67.5
Ours (49 heads)	87.0 M	**82.8**	**89.4**	**83.4**	**68.4**	**81.0**	**81.9**	**56.1**	63.4	30.0	**63.2**	**76.5**	**83.2**	73.5	71.1	84.0	**68.3**
lora (r=8)	6.8 M	49.2	43.0	42.5	42.0	44.2	78.6	52.0	61.6	**30.8**	58.1	74.8	81.8	**74.2**	74.1	85.2	67.1
lora (r=16)	13.6 M	59.5	44.3	47.4	47.2	49.6	77.7	52.2	62.2	30.0	57.1	72.2	79.2	73.6	73.4	85.5	66.3
lora (r=102)	86.9 M	71.3	75.8	75.1	54.6	69.2	76.3	53.1	**67.1**	11.2	55.7	73.9	80.4	74.1	**76.0**	**85.9**	65.4