Flexi-LoRA: Efficient LoRA Finetuning with Input-Adaptive Dynamic Ranks

Zongqian Li [1]   Yixuan Su [1]   Han Zhou [1]   Zihao Fu [1]   Nigel Collier [1]

Abstract

Parameter-efficient fine-tuning methods like Low-Rank Adaptation (LoRA) have become essential for deploying large language models, yet their static parameter allocation remains suboptimal for inputs of varying complexity. We present Flexi-LoRA, a novel framework that dynamically adjusts LoRA ranks based on input complexity during both training and inference. Through empirical analysis across question answering and mathematical reasoning tasks, we demonstrate that maintaining consistency between training and inference dynamics is important for effective adaptation, particularly for sequential reasoning tasks. Our findings reveal that input-dependent parameter allocation achieves superior performance with fewer parameters by optimally matching rank configurations to question complexity. Furthermore, task-specific sensitivity to rank dynamics varies, with mathematical reasoning tasks exhibiting higher sensitivity than QA tasks. Successful adaptation manifests not only in correctness but also in reasoning quality and instruction adherence. Flexi-LoRA consistently outperforms static LoRA while using fewer parameters, with performance gains more pronounced on tasks requiring strict reasoning chains. Our approach realizes key benefits of mixture-of-experts frameworks through a more streamlined implementation, reducing parameter redundancy while enhancing model capabilities. We provide comprehensive empirical studies across diverse tasks, establishing a foundation for future work in input-adaptive and efficient fine-tuning approaches. [1]
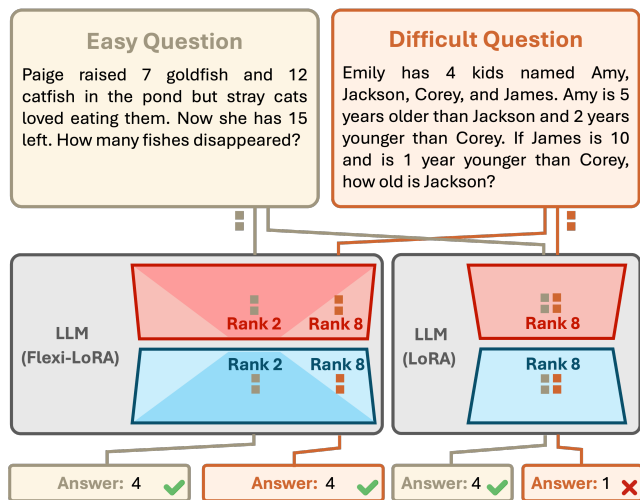
Figure 1. Flexi-LoRA's input-adaptive rank allocation versus static normal LoRA. Flexi-LoRA (left) dynamically assigns rank 2 (dark trapezoid) for simple problems and rank 8 (light trapezoid) for complex ones, successfully solving both. LoRA (right) uses fixed rank 8 (light trapezoid) regardless of complexity, failing on difficult problems. This demonstrates the necessity of input-adaptive parameter allocation for handling varying question complexity.

1. Introduction

As large language models grow in size, efficient fine-tuning methods like LoRA (Hu et al., 2022) have become essential for applications. However, their static parameter allocation remains suboptimal for questions of varying complexity, suggesting the need for input-adaptive approaches in parameter-efficient fine-tuning (Jiang et al., 2025).

Through empirical analysis, we observe two key **phenomena** in LoRA-based fine-tuning. **First**, there exists a notable performance gap when using static ranks during inference for models trained with dynamic ranks at fine-grained level, particularly in their ability to follow instructions precisely (DyLoRA vs DyLoRA+, Table 2 and 3, Figure 4). **Second**, while model performance generally saturates with increasing ranks, the optimal rank varies across different inputs: simple questions can be effectively handled with small ranks, while complex problems benefit substantially from larger ranks (Rank 4 vs 8, Table 2 and 3, Figure 3). These observa-

---

[1]University of Cambridge. Correspondence to: Zongqian Li, Nigel Collier <zl510@cam.ac.uk, nhc30@cam.ac.uk>.

[1]https://github.com/ZongqianLi/Flexi-LoRA

tions indicate that a one-size-fits-all approach to rank selection is suboptimal, motivating the need for input-adaptive rank allocation.

Inspired by these observations, we propose **Flexi-LoRA**, a finetuning framework that dynamically adjusts LoRA ranks based on input complexity. Our approach not only achieves superior performance to high-rank LoRA while using fewer parameters, but also successfully solves some complex problems that static LoRA fails to handle even with equivalent rank, as shown in Figure 1.

Our **contributions** are threefold:

- **Novel Framework:** We introduce the first input-adaptive LoRA framework that maintains dynamic ranks during both training and inference, achieving higher performance with reduced parameter count compared to static LoRA.
- **Insights:** We demonstrate that (1) maintaining consistency between training and inference dynamics is important for LoRA adaptation, particularly for sequential reasoning tasks; (2) input-dependent parameter allocation achieves superior performance with fewer parameters by optimally matching rank configurations to question complexity; (3) task-specific sensitivity to rank dynamics varies, with mathematical reasoning tasks exhibiting higher sensitivity than QA tasks; (4) successful adaptation manifests not only in correctness but also in reasoning quality and instruction adherence; and (5) our approach realizes benefits of mixture-of-experts through a more streamlined implementation, reducing parameter redundancy while enhancing model capabilities.
- **Comprehensive Analysis:** We provide comprehensive empirical studies across diverse tasks, establishing a foundation for future work in input-adaptive and efficient finetuning approaches.

2. Related Work

LoRA with dynamic ranks. Recent works have explored dynamic rank adaptation in LoRA, with differences shown in Table 1. AdaLoRA (Zhang et al., 2023) performs importance-based parameter pruning at training checkpoints to gradually reduce ranks to a fixed target. DyLoRA (Valipour et al., 2023) randomly samples ranks from a predefined range for each training batch, with all samples in the batch sharing the same rank. Both approaches, while improving rank flexibility, are limited by either steps-level pruning or random batch-level assignment, and neither supports dynamic rank selection at inference. On the other hand, Flexi-LoRA enables true sample-level rank selection by learning to map input complexity to appropriate ranks, maintaining this adaptive behavior during both training and inference.

Method	Train	Level	Inference
LoRA	Fixed	All	Fixed
AdaLoRA	Selective	Steps	Fixed
DyLoRA	Random	Batch	Fixed
Flexi-LoRA	**Router**	**Sample**	**Router**

Table 1. Comparison of rank adaptation strategies across different LoRA variants. "Train" indicates how ranks are determined during training, "Level" shows the level of rank assignment, and "Inference" specifies the rank selection mechanism at test time. Only Flexi-LoRA maintains consistent router-based sample-level dynamic rank allocation across both training and inference stages, while existing methods use fixed ranks during inference regardless of training dynamics. "O" is our method.

3. Methods

Building upon previous work, we first introduce **DyLoRA+**, an enhanced variant of DyLoRA that maintains consistent rank dynamics by employing random batch-level rank selection during both training and inference stages. While DyLoRA+ demonstrates improved performance over the original DyLoRA, its random rank allocation remains suboptimal as it fails to account for input-specific complexity differences. We therefore propose **Flexi-LoRA**, a framework that automatically adjusts the rank based on input complexity. Our method consists of two key components: a difficulty-aware router that maps inputs to appropriate rank assignments and a flexible-rank LoRA training framework that maintains consistent dynamic rank allocation during both training and inference, as shown in Figure 2.

Router focuses on learning an optimal mapping $R(h)$: $\mathbb{R}^d \rightarrow r_i$ from input embeddings to rank assignments. Given an input sequence x with mask m, we first compute its token embeddings $H \in \mathbb{R}^{n \times d}$ and obtain a pooled embedding $h = \sum_i (m_i H_i) / \sum_i m_i$, where m_i masks padding tokens. We categorize training samples into difficulty classes based on task-specific metrics: F1 scores for MRQA datasets and accuracy for mathematical reasoning tasks. The router is then optimized using a noise-added cross-entropy objective: $\mathcal{L}(\theta) = -\sum_i y_i \log(R(h_i + \epsilon))$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is Gaussian noise and y_i denotes the ground-truth difficulty label. The training data is balanced between easy and hard samples to ensure uniform difficulty evaluation.

Input-adaptive LoRA freezes the base model parameters and optimize only the LoRA matrices. For input x, we first obtain its token embeddings H^0 and pooled embedding h following the same procedure as router training. The router then predicts rank $r = R(h)$, which is applied consistently across all transformer layers. Within each batch, different samples can be assigned different ranks based on their predicted difficulty, enabling dynamic resource allocation. For each transformer layer l, we compute the LoRA update as
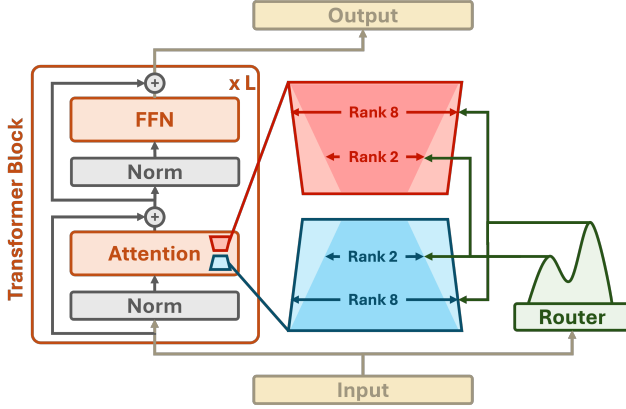
Figure 2. Flexi-LoRA framework with input-adaptive rank selection. Router analyzes input embeddings and outputs rank assignments (green arrows) for transformer layers. Red and blue trapezoids are LoRA's A and B matrices, with color darkness indicating rank magnitude (darker = rank 2, lighter = rank 8). The router enables dynamic rank allocation based on input complexity while maintaining efficient gradient flow through residual connections.

$\Delta W_l = B_{l,r} A_{l,r}$, where $A_{l,r} \in \mathbb{R}^{r \times d}$ and $B_{l,r} \in \mathbb{R}^{d \times r}$ are dynamically reduced to the first r rows/columns. The layer output is computed as $H^l = W_l H^{l-1} + \alpha_r (B_{l,r} A_{l,r} H^{l-1})$, where α_r is a rank-specific scaling variable and H^{l-1} is the output from the previous layer. The model is trained to minimize the task-specific loss $\mathcal{L}_{\text{task}} = -\sum_i \log p(y_i|x_i)$, where x_i is the input sequence and y_i is the corresponding ground truth outputs. This design enables efficient batch processing while allowing for input-dependent rank adaptation.

4. Experimental Design

Datasets. We evaluate Flexi-LoRA on both QA and mathematical reasoning tasks. For QA tasks, we conduct training on datasets from the MRQA (Fisch et al., 2019) training set, which unifies QA samples from SQuAD (Rajpurkar et al., 2016), TriviaQA (Joshi et al., 2017), NewsQA (Trischler et al., 2017), SearchQA (Dunn et al., 2017), HotpotQA (Yang et al., 2018), and NaturalQuestions (Kwiatkowski et al., 2019). Evaluation is performed on the MRQA test set consisting of BioASQ (Partalas et al., 2013), DROP (Dua et al., 2019), DuoRC (Saha et al., 2018), RACE (Lai et al., 2017), RelationExtraction (Levy et al., 2017), and TextbookQA (Kembhavi et al., 2017). For mathematical reasoning, we train on the GSM8K (Cobbe et al., 2021) subset of the MetaMathQA (Yu et al., 2024) dataset and evaluate on a diverse set of math benchmarks including GSM8K, SVAMP (Patel et al., 2021), MultiArith (Roy & Roth, 2015), and MAWPS (Koncel-Kedziorski et al., 2016). This design allows us to evaluate both in-distribution and out-of-

distribution generalization capabilities of our method.

Evaluation Metrices. We evaluate **QA** performance using **F1** and **Exact Match (EM)** scores. F1 computes the balanced average of precision and recall between prediction and ground truth, while EM measures exact string match. For **mathematical** reasoning tasks, we use **accuracy** for evaluation.

Gold Standard & Baselines. We compare against two **gold** standards: **full** model fine-tuning and standard **LoRA** with fixed rank. For **baselines**, we include **AdaLoRA**, which adapts ranks through importance-based parameter pruning while maintaining fixed inference ranks, and **DyLoRA**, which randomly samples ranks from a predefined range for each training batch but uses fixed ranks during inference. Our Flexi-LoRA differs by enabling input-adaptive rank selection during both training and inference.

Models. We employ **LLaMA-3.2-1B-Instruct** (Grattafiori et al., 2024) as the base model for our main results and include **LLaMA-3.2-3B-Instruct** to analyze model size in ablation studies.

5. Results

5.1. Overview

Figure 3 illustrates the performance-efficiency trade-offs across different parameter-efficient fine-tuning methods on QA and mathematical reasoning tasks. Flexi-LoRA consistently achieves superior performance while requiring fewer parameters than competing approaches. Notably, DyLoRA results are not visible from mathematical reasoning figures due to substantially decreased performance, highlighting the importance of maintaining consistency between training and inference dynamics. AdaLoRA shows competitive results on specific tasks but fails to achieve a consistent advantage across different domains. From a Pareto optimality perspective, Flexi-LoRA dominates all baselines and gold standards by offering better performance at lower parameter counts, positioning itself closest to the full fine-tuning performance while maintaining parameter efficiency below 0.1% of total model parameters. The following sections analyze these results in detail across different task categories and provide ablation studies to analyze the contributions of individual components.

5.2. Question Answering

Table 2 presents the performance of different parameter-efficient fine-tuning methods on six out-of-domain QA datasets from the MRQA benchmark. We analyze these results from multiple perspectives: (1) **Overall Performance**: Flexi-LoRA (2,8) achieves the highest average F1 score (52.37%) and EM score (37.41%), outperforming
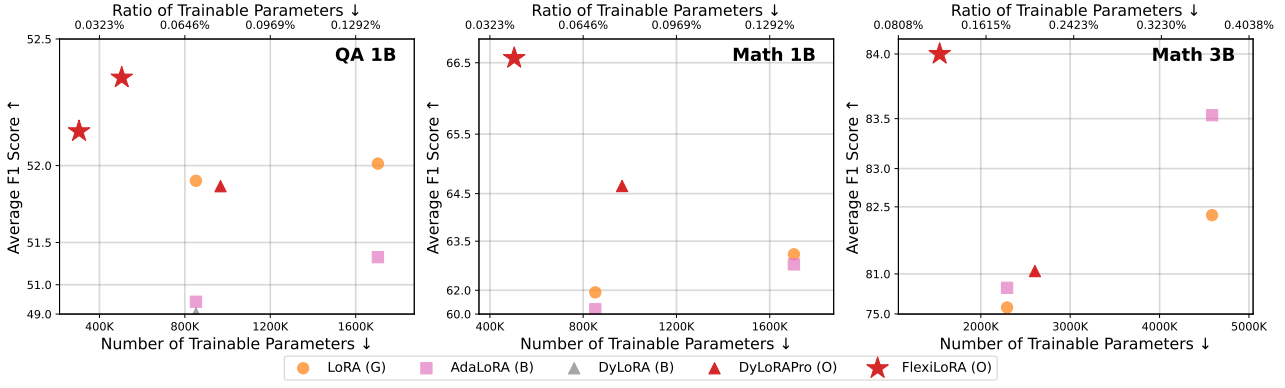
Figure 3. Performance-efficiency trade-off across different parameter-efficient fine-tuning methods. Our methods (O) achieve superior performance with fewer trainable parameters compared to both baseline methods (B) as well as gold standards (G). Results are shown for QA tasks and mathematical tasks using LLaMA-3.2-1B-Instruct and 3B.

	Gold Standard			Baselines				Ours		
	Full	LoRA		AdaLoRA		DyLoRA		DyLoRA+	Flexi-LoRA	
Rank	-	4	8	4	8	4	8	1-8	1, 8	2, 8
#	1.2B	851K	1703K	851K	1703K	851K	1703K	966K	**304K**	**504K**
F1 Score										
BioASQ	69.81	64.85	66.22	63.62	65.12	60.01	52.40	65.21	65.75	65.82
DROP	47.15	37.88	36.03	32.79	34.35	43.59	38.26	36.32	37.27	37.52
DuoRC	45.21	43.85	43.72	44.00	42.49	39.10	36.89	43.86	42.92	43.22
RACE	41.49	38.64	37.49	34.58	35.81	36.89	33.24	39.26	38.83	39.10
RE	84.11	74.97	76.41	74.51	75.73	81.02	78.70	75.47	76.47	76.83
TextbookQA	49.56	51.35	52.20	54.14	54.66	37.90	32.94	51.22	51.74	51.75
Average	56.22	51.92	52.01	50.61	51.36	49.75	45.40	51.89	52.16	**52.37**
Exact Match										
BioASQ	49.13	42.02	42.61	40.29	40.35	34.24	27.52	41.88	41.42	41.48
DROP	35.46	25.81	23.08	20.15	21.49	30.07	24.01	23.61	24.55	25.01
DuoRC	35.57	32.37	31.51	32.37	30.77	27.71	24.58	32.31	30.44	30.71
RACE	29.37	24.03	22.99	20.62	22.10	21.81	16.61	24.92	24.48	24.77
RE	72.01	57.73	60.27	57.25	58.44	69.13	66.31	59.15	59.90	60.44
TextbookQA	40.98	41.91	42.38	44.97	45.10	27.01	21.29	41.91	42.04	42.04
Average	43.75	37.31	37.14	35.94	36.38	34.99	30.05	37.30	37.14	**37.41**

Table 2. Performance comparison on out-of-domain **QA** tasks from the MRQA benchmark using LLaMA-3.2-1B-Instruct. F1 and Exact Match (EM) scores are reported, comparing our proposed methods (Flexi-LoRA and DyLoRA+) against gold standards and baselines. Flexi-LoRA (2,8) achieves the best average performance on both metrics while using fewer parameters than standard approaches. The "#" row indicates the number of trainable parameters. Green (teal) and red (maroon) cell coloring is higher and lower scores respectively, with deeper colors indicating larger performance differences.

both LoRA-8 with only 29.59% of LoRA-8's parameters. The performance gap between Flexi-LoRA and full fine-tuning is considerably smaller than that of other parameter-efficient methods, demonstrating its effectiveness in approaching full fine-tuning. (2) **Stability Across Metrics**: Flexi-LoRA demonstrates consistency by achieving the best average performance on both F1 and EM metrics simultaneously, unlike other methods that typically excel in one metric. This dual-metric superiority indicates that Flexi-LoRA produces outputs that are both semantically close to ground truth (high F1) and syntactically precise (high EM). (3) **Dataset-Specific Analysis**: Flexi-LoRA performs particularly well on domain-specific and knowledge-based tasks, while showing moderate improvements on information extraction tasks. The adaptive rank allocation proves beneficial for datasets requiring diverse reasoning capabili-

ties, suggesting Flexi-LoRA's ability to assign appropriate computational resources based on input complexity. (4) **Comparison with Other Dynamic Approaches**: Flexi-LoRA consistently outperforms DyLoRA+, despite both using dynamic ranks during inference, demonstrating the importance of learned input-adaptive rank assignment versus the random assignment in DyLoRA+. DyLoRA shows high variance across datasets (from 81.02% F1 on RE to 32.94% F1 on TextbookQA), highlighting instability issues when training and inference dynamics are inconsistent. (5) **Rank Configuration Influence**: The comparison between Flexi-LoRA (1,8) and (2,8) configurations reveals that a slight increase in the minimum rank (from 1 to 2) provides modest but consistent performance improvements (from 52.16% to 52.37% F1), with a reasonable parameter increase (from 304K to 504K). This suggests that while low

ranks can handle simpler questions, maintaining a slightly higher minimum rank improves robustness across diverse question types. (6) **Cross-Domain Generalization**: On out-of-domain test datasets, baseline methods exhibit inconsistent performance, excelling on specific domains while decreasing on others. On the other hand, Flexi-LoRA maintains strong performance across all test datasets, suggesting its input-adaptive parameter allocation learns more generalized knowledge than static approaches that could overfit to training domain characteristics. (7) **Key Insights**: These results demonstrate that input-adaptive parameter allocation provides dual benefits in QA tasks: improved performance through parameter allocation and enhanced parameter efficiency through optimization of rank selection. The consistent improvement across diverse datasets suggests that question complexity varies substantially even within the same task category, validating our input-adaptive approach.

5.3. Math Problems

Table 3 presents the performance of different fine-tuning methods on mathematical reasoning tasks across two model sizes. We analyze these results for task-specific characteristics: (1) **Overall Performance**: Flexi-LoRA achieves the highest average accuracy on both LLaMA-3.2-1B-Instruct (66.56%) and LLaMA-3.2-3B-Instruct (84.00%), outperforming LoRA-8 (63.17% and 82.37%) while using only 31.29% and 33.40% of its parameters, respectively. The performance gap between Flexi-LoRA and full fine-tuning is narrower than other parameter-efficient methods, particularly on the larger model. (2) **Model Size Influence**: Increasing model size from 1B to 3B improves performance across all methods, with Flexi-LoRA maintaining its advantage. Notably, the absolute performance gap between Flexi-LoRA and full fine-tuning decreases from 5.56% to 1.84%, suggesting that input-adaptive approaches become even more effective with larger models. (3) **In-Domain vs. Out-of-Domain**: Flexi-LoRA shows strong generalization from GSM8K (in-domain) to out-of-domain datasets. On the 1B model, Flexi-LoRA achieves an average accuracy of 74.65% on out-of-domain tasks (SVAMP, MultiArith, MAWPS), compared to 42.30% on in-domain GSM8K, demonstrating cross-domain robustness. This is consistent across both model sizes. (4) **Dataset Complexity**: Flexi-LoRA performs exceptionally well on both elementary arithmetic (MultiArith) and complex multi-step reasoning (GSM8K), indicating its ability to effectively handle varying levels of mathematical complexity. The exceptional performance on MultiArith (92.22%) approaches full fine-tuning (93.88%), showcasing the method's ceiling capability on reasoning tasks. (5) **DyLoRA Performance Decrease**: DyLoRA exhibits performance decrease on mathematical tasks (average 26.50% on 1B model). The 40.06% performance gap between DyLoRA and Flexi-LoRA on math tasks high-

lights the influence of training-inference inconsistency on sequential reasoning tasks. This decrease is substantially more pronounced than in QA tasks, suggesting that mathematical reasoning is particularly sensitive to dynamic rank consistency. (6) **Key Insights**: These results demonstrate that maintaining consistent training-inference dynamics is important for mathematical reasoning tasks. The substantial performance improvements (Flexi-LoRA outperforms LoRA-8 by 3.39% on 1B) illustrate that input-adaptive parameter allocation provides greater benefits for problems with higher complexity variance and stricter evaluation criteria, compared to the modest gains observed on QA tasks.

5.4. Cross-Task Analysis

Comparing Flexi-LoRA's performance on QA and mathematical reasoning tasks reveals important insights about the relationship between task characteristics and parameter-efficient fine-tuning approaches: (1) **Task Nature Influence**: Task characteristics influence adaptation strategy efficacy. QA primarily involves information extraction, whereas mathematical reasoning demands coherent computational chains where early errors go through solutions. This sequential dependency in math tasks necessitates consistency between training and inference dynamics, unlike the more information retrieval in QA. (2) **Performance Gain Difference**: Flexi-LoRA's improvement over LoRA-8 is substantially larger on mathematical tasks than on QA tasks. This difference suggests that input-adaptive parameter allocation yields greater benefits for sequential reasoning tasks where capacity requirements vary between simple and complex problems. (3) **Error Analysis**: QA and mathematical reasoning tasks demonstrate distinct error behaviors. QA permits partial correctness (evidenced by F1 scores exceeding EM scores), whereas math tasks produce binary outcomes. This evaluation in mathematical reasoning highlights the benefits of input-adaptive approaches that allocate capacity proportionally to problem complexity. (4) **Input Complexity Distribution**: Flexi-LoRA's superior performance on mathematical tasks indicates more pronounced complexity variations compared to QA. Math problems encompass a broader difficulty distribution, from elementary arithmetic to multi-step reasoning, making them optimal candidates for adaptive parameter allocation. (5) **Cross-Task Generalization**: Flexi-LoRA's consistent superior performance across both QA and mathematical reasoning tasks, despite their distinct nature and evaluation metrics, validates the robustness and generality of our input-adaptive approach. This multi-domain effectiveness indicates that the underlying complexity-aware parameter allocation mechanism learns fundamental aspects of language model adaptation rather than utilizing task-specific characteristics. These cross-task comparisons highlight that the benefits of input-adaptive parameter allocation are not uniform across tasks,

	Gold Standard			Baselines				Ours	
	Full	LoRA		AdaLoRA		DyLoRA		DyLoRA+	Flexi-LoRA
Rank	-	4	8	4	8	4	8	1-8	2, 8
				LaMA-3.2-1B-Instruct					
#	1.2B	851K	1703K	851K	1703K	851K	1703K	953K	**533K**
GSM8K	57.31	42.15	41.31	45.71	42.22	19.78	21.00	41.77	42.30
SVAMP	57.29	52.87	51.18	53.37	50.52	20.08	19.57	56.03	52.02
MultiArith	93.88	82.77	85.00	78.88	84.44	43.88	42.77	85.55	92.22
MAWPS	80.00	69.85	75.21	65.63	74.36	22.25	17.46	75.21	79.71
Average	72.12	61.91	63.17	60.90	62.89	26.50	25.20	64.64	**66.56**
				LaMA-3.2-3B-Instruct					
#	1.2B	2.29M	4.58M	2.29M	4.58M	2.29M	4.58M	2.60M	**1.53M**
GSM8K	75.20	65.95	69.37	71.49	72.32	60.04	60.87	70.43	69.90
SVAMP	78.30	66.29	74.47	75.91	77.80	67.48	64.87	74.71	77.09
MultiArith	100	97.77	99.44	90.55	95.55	75.00	82.22	92.77	100
MAWPS	89.85	84.50	86.19	83.66	88.45	69.57	65.63	86.47	89.01
Average	85.84	78.63	82.37	80.40	83.53	68.02	68.40	81.10	**84.00**

Table 3. Performance comparison on **mathematical tasks** using LLaMA-3.2-1B-Instruct and LLaMA-3.2-3B-Instruct models. Accuracy scores (%) are reported across four benchmark datasets (GSM8K as in-domain and the others as out-of-domain), comparing our proposed methods against gold standards and baselines. Flexi-LoRA achieves the best average performance on both model sizes while using fewer parameters than standard approaches. The "#" row indicates the number of trainable parameters. Green (teal) and red (maroon) cell coloring is higher and lower scores respectively, with deeper colors indicating larger performance differences.

but instead depend on the task's inherent complexity distribution and error passing dynamics. The more sequential the reasoning required by a task, the greater the advantage offered by Flexi-LoRA's dynamic rank allocation.

5.5. Case Study

We present a qualitative analysis of model outputs to understand the influence of different fine-tuning approaches on mathematical reasoning tasks. Table 4 showcases two typical cases that show key differences between our proposed methods and existing baselines. (1) **Training-Inference Consistency Influence**: Table 4 reveals DyLoRA's outputs mirror the Original model, indicating finetuning failure due to training-inference inconsistency (random rank sampling during training versus fixed rank at inference). On the other hand, DyLoRA+ successfully fine-tunes by maintaining consistent random rank selection across both stages. This difference demonstrates that dynamic consistency between training and inference is important for effective parameter-efficient finetuning. (2) **Fine-tuning Quality Indicators**: Beyond correctness, fine-tuned models show qualitative improvements in reasoning. DyLoRA+ and Flexi-LoRA produce detailed step-by-step solutions and follow output formatting ("The answer is:"), while Original and DyLoRA generate abbreviated responses that ignore instruction requirements. These differences, detailed reasoning and instruction adherence, serve as reliable indicators of successful finetuning. (3) **Error Analysis**: The cases reveal three error types in mathematical reasoning: (a) conceptual errors (Case 1: treating $4/carton as $4/serving, leading to $3,600 vs $16), (b) arithmetic misunderstanding (Case 2: "2/5 times more" computed as "2 times more"), and (c) process errors (correct intermediate steps but incomplete final calculations). These errors go through multi-step reasoning, showing initial mistakes. (4) **Reasoning Quality Comparison**: Successful methods maintain accuracy throughout

reasoning chains. Case 2 illustrates this: LoRA correctly understands proportional relationships but fails in final answer, while Flexi-LoRA completes all computational steps correctly. Early-stage errors in baseline approaches go through the entire solution.

5.6. Ablation Studies

Figure 4 presents a comprehensive analysis analyzing key variables that contribute to Flexi-LoRA's effectiveness. We examine four aspects of our approach: (1) **Training-Inference Rank Consistency** is essential for performance, as evidenced by DyLoRA's decrease when using fixed ranks at inference despite dynamic training. This confirms our hypothesis that maintaining consistent rank dynamics is important, especially for sequential reasoning tasks. (2) **Input-Adaptive vs. Random Selection** comparison between Flexi-LoRA and DyLoRA+ demonstrates that learned complexity-aware allocation consistently outperforms random selection, validating the effectiveness of our router-based approach. (3) **Model Size Influences** show that while scaling from 1B to 3B parameters improves overall performance across all methods, Flexi-LoRA maintains its relative advantage, indicating that input-adaptive allocation remains beneficial regardless of model size. (4) **Task-Dependent Sensitivity** analysis reveals that mathematical reasoning tasks exhibit higher sensitivity to rank dynamics than QA tasks, with up to 39.44% performance gap between consistent and inconsistent methods on math compared to 6.49% on QA, illustrating how tasks with stricter evaluation criteria and error passing benefit more from adaptive parameter allocation.

5.7. Efficiency Analysis

Our analysis demonstrates Flexi-LoRA's superior parameter efficiency across multiple aspects. Figure 3 and Tables 2

6

Question: Cynthia eats one serving of ice cream every night. She buys cartons of ice cream with 15 servings of ice cream per carton at a cost of $4.00 per carton. After 60 days, how much will she spend on ice cream? **Standard Answer:** 16		
Original: To find the total cost of ice cream, multiply the number of servings per carton by the number of cartons Cynthia buys. 15 servings per carton * 60 cartons = 900 servings. 900 servings * $4.00 per serving = $3,600. The answer is: $3,600.	**DyLoRA (Baseline):** To find the total cost of ice cream, multiply the number of servings per carton by the number of cartons Cynthia buys. 15 servings per carton * 60 cartons = 900 servings. 900 servings * $4.00 per serving = $3600. The answer is: $3600.	**DyLoRA+ (Ours):** Cynthia eats one serving of ice cream every night, so in 60 days, she will eat 60 servings of ice cream. Each carton of ice cream has 15 servings, so she will buy 60/15 = 4 cartons of ice cream. Each carton costs $4.00, so she will spend 4 * $4.00 = $16.00 on ice cream. #### 16 The answer is: 16
Question: Meredith is a freelance blogger who writes about health topics and submits to clients each day as her permanent job. A blog article takes an average of 4 hours to research and write about. Last week, she wrote 5 articles on Monday and 2/5 times more articles on Tuesday than on Monday. On Wednesday, she wrote twice the number of articles she wrote on Tuesday. Calculate the total number of hours she spent writing articles in the three days. **Standard Answer:** 104		
Original: Meredith wrote 5 articles on Monday, 5 + 2(5) = 15 articles on Tuesday, and 2(15) = 30 articles on Wednesday. The total number of hours she spent writing articles is 4(5) + 4(15) + 4(30) = 20 + 60 + 120 = 200 hours.	**LoRA (Gold Standard):** Meredith wrote 5 articles on Monday, which took 4 hours to research and write, so she spent 5 * 4 = 20 hours writing on Monday. On Tuesday, she wrote 2/5 times more articles than on Monday, so she wrote 5 + (2/5 * 5) = 5 + 2 = 7 articles. On Wednesday, she wrote twice the number of articles she wrote on Tuesday, so she wrote 7 * 2 = 14 articles. In total, Meredith spent 20 + 7 + 14 = 41 hours writing articles in the three days. #### 41 The answer is: 41	**Flexi-LoRA (Ours):** Meredith wrote 5 articles on Monday, and 2/5 times more articles on Tuesday, so she wrote 5 + (2/5)*5 = 5 + 2 = 7 articles on Tuesday. On Wednesday, she wrote twice the number of articles she wrote on Tuesday, so she wrote 2*7 = 14 articles on Wednesday. In total, Meredith wrote 5 + 7 + 14 = 26 articles in the three days. Each article takes an average of 4 hours to research and write about, so she spent 26*4 = 104 hours writing articles in the three days. #### 104 The answer is: 104

Table 4. **Case study** comparing reasoning performance across different LoRA variants on mathematical problems for LLaMA-3.2-1B-Instruct. Examples demonstrate how training-inference LoRA dynamics consistency influences problem solving: DyLoRA+ and Flexi-LoRA correctly solve problems that confound baseline methods, with errors highlighted in red and correct steps in green. Flexi-LoRA's input-adaptive parameter allocation enables more accurate calculations and multi-step reasoning compared to static LoRA.
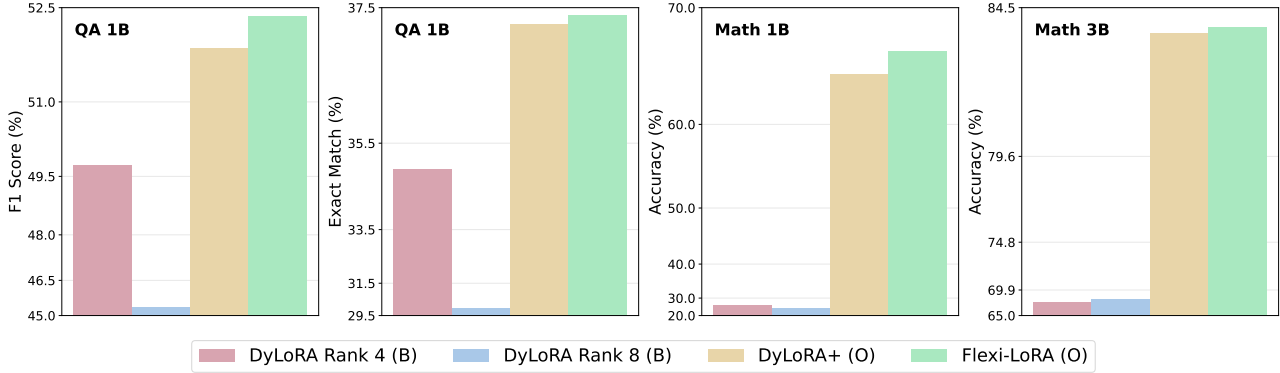

Figure 4. **Ablation study** comparing performance influence of rank selection strategies across model sizes. Charts show performance on QA tasks and mathematical reasoning tasks for LLaMA-3.2-1B-Instruct and 3B models. The methods are labeled as (B) for baseline methods and (O) for our proposed methods. All methods use dynamic ranks during training but differ in inference approach: DyLoRA uses fixed ranks (4 or 8), DyLoRA+ employs random rank selection, and Flexi-LoRA utilizes input-adaptive rank allocation. Results demonstrate that Flexi-LoRA maintains its advantage regardless of model size and task type, confirming the benefits of input-adaptive rank allocation and the consistency between training and inference rank dynamics.

and 3 quantify this advantage: Flexi-LoRA (2,8) achieves the highest QA performance with only 504K trainable parameters, compared to 1703K for LoRA-8, a 70.40% parameter reduction while improving performance. Flexi-LoRA (1,8) further reduces parameter count to 304K (17.85% of LoRA-8) while maintaining competitive performance. This efficiency improvement extends to mathematical reasoning tasks, where Flexi-LoRA outperforms LoRA-8 on 1B models (66.56% vs. 63.17% accuracy) using only 533K

parameters (31.29% of LoRA-8's 1703K). Regarding computational overhead, Flexi-LoRA's router consists of only two layers that process the pooled input embedding once per sequence, introducing negligible additional computation compared to the base model computation. This minimal overhead is substantially balanced by the parameter efficiency gains, resulting in an overall more efficient finetuning framework that demonstrates consistent advantages across model sizes and tasks.

6. Conclusions

This paper introduces Flexi-LoRA, an input-adaptive framework that dynamically adjusts LoRA ranks based on question complexity. We demonstrate that maintaining consistent rank dynamics between training and inference is important for finetuning models, particularly for sequential reasoning tasks. Flexi-LoRA outperforms static LoRA while using fewer parameters (29.6% for QA and 31.3% for math reasoning), with performance gains more pronounced on mathematical tasks requiring reasoning chains. These results confirm that input-dependent parameter allocation enables efficient capacity allocation while reducing parameter redundancy, achieving benefits similar to mixture-of-experts frameworks through a more streamlined method.

Future work could have several directions: (1) layer-specific dynamic ranks to optimize parameter utilization at finer level; (2) router frameworks learning hierarchical aspects of input complexity; (3) integration with other parameter-efficient techniques such as sparse fine-tuning.

Ethics Statement

No ethical approval was required for this study.

Availability Statement

The codes and models related to this paper are uploaded to the open-source community at https://github.com/ZongqianLi/Flexi-LoRA.

References

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., and Gardner, M. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In Burstein, J., Doran, C., and Solorio, T. (eds.), *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. 2368–2378, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1246. URL https://aclanthology.org/N19-1246.

Dunn, M., Sagun, L., Higgins, M., Guney, V. U., Cirik, V., and Cho, K. Searchqa: A new q&a dataset augmented with context from a search engine, 2017. URL https://arxiv.org/abs/1704.05179.

Fisch, A., Talmor, A., Jia, R., Seo, M., Choi, E., and Chen, D. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of 2nd Machine Reading for Reading Comprehension (MRQA) Workshop at EMNLP*, 2019.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., ..., and Ma, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jiang, S., Liao, Y., Zhang, Y., Wang, Y., and Wang, Y. Fine-tuning with reserved majority for noise reduction. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ZV7CLf0RHK.

Joshi, M., Choi, E., Weld, D., and Zettlemoyer, L. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Barzilay, R. and Kan, M.-Y. (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147.

Kembhavi, A., Seo, M., Schwenk, D., Choi, J., Farhadi, A., and Hajishirzi, H. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5376–5384, 2017. doi: 10.1109/CVPR.2017.571.

Koncel-Kedziorski, R., Roy, S., Amini, A., Kushman, N., and Hajishirzi, H. MAWPS: A math word problem repository. In Knight, K., Nenkova, A., and Rambow, O. (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1152–1157, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1136. URL https://aclanthology.org/N16-1136/.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.

1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026.

Lai, G., Xie, Q., Liu, H., Yang, Y., and Hovy, E. RACE: Large-scale ReAding comprehension dataset from examinations. In Palmer, M., Hwa, R., and Riedel, S. (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1082. URL https://aclanthology.org/D17-1082.

Levy, O., Seo, M., Choi, E., and Zettlemoyer, L. Zero-shot relation extraction via reading comprehension. In Levy, R. and Specia, L. (eds.), *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pp. 333–342, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/K17-1034. URL https://aclanthology.org/K17-1034.

Partalas, I., Gaussier, E., Ngomo, A.-C. N., et al. Results of the first bioasq workshop. In *BioASQ@ CLEF*, pp. 1–8, 2013.

Patel, A., Bhattamishra, S., and Goyal, N. Are NLP models really able to solve simple math word problems? In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.168. URL https://aclanthology.org/2021.naacl-main.168/.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. SQuAD: 100,000+ questions for machine comprehension of text. In Su, J., Duh, K., and Carreras, X. (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264.

Roy, S. and Roth, D. Solving general arithmetic word problems. In Màrquez, L., Callison-Burch, C., and Su, J. (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1743–1752, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1202. URL https://aclanthology.org/D15-1202/.

Saha, A., Aralikatte, R., Khapra, M. M., and Sankaranarayanan, K. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In Gurevych, I. and Miyao, Y. (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1683–1693, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1156. URL https://aclanthology.org/P18-1156.

Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., and Suleman, K. NewsQA: A machine comprehension dataset. In Blunsom, P., Bordes, A., Cho, K., Cohen, S., Dyer, C., Grefenstette, E., Hermann, K. M., Rimell, L., Weston, J., and Yih, S. (eds.), *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 191–200, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2623. URL https://aclanthology.org/W17-2623/.

Valipour, M., Rezagholizadeh, M., Kobyzev, I., and Ghodsi, A. DyLoRA: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In Vlachos, A. and Augenstein, I. (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3274–3287, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.239. URL https://aclanthology.org/2023.eacl-main.239/.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., and Manning, C. D. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259.

Yu, L., Jiang, W., Shi, H., YU, J., Liu, Z., Zhang, Y., Kwok, J., Li, Z., Weller, A., and Liu, W. Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=N8N0hgNDRt.

Zhang, Q., Chen, M., Bukharin, A., He, P., Cheng, Y., Chen, W., and Zhao, T. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=lq62uWRJjiY.