Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.080068

ARTICLE

H-LoRA: Rethinking Rank Selection for Controllable Knowledge Retention in Edge AI

Darren Chai Xin Lun and Lim Tong Ming*

Centre for Business Incubation and Entrepreneurial Ventures, Tunku Abdul Rahman University of Management and Technology, Jalan Genting Kelang, Setapak, Kuala Lumpur, Malaysia

*Corresponding Author: Lim Tong Ming. Email: limtm@tarc.edu.my

Received: 02 February 2026; Accepted: 30 March 2026; Published: 08 May 2026

ABSTRACT: The deployment of specialized language models in resource-constrained edge environments (≤1B parameters, ≤2 GB memory, ≤100 ms latency) faces a critical challenge: Supervised Fine-Tuning (SFT) achieves domain expertise but suffers from irreversible catastrophic forgetting, while traditional Low-Rank Adaptation (LoRA) with conservative ranks ($r \leq 64$) often underperforms due to insufficient adaptation capacity. This work introduces H-LoRA (High-Rank LoRA) for edge-deployable models and establishes a fundamental distinction between destructive forgetting and controllable knowledge retention. Through comprehensive experiments on compact models (0.12B Minimind and Qwen-0.5B) across three domains (Human Resources, Medical, Mathematics) using 29,647 samples, we demonstrate that while both SFT and H-LoRA exhibit general capability degradation, they differ fundamentally: SFT completely destroys the original knowledge structure (1% topic retention), while H-LoRA maintains knowledge integrity with 90% topic retention—an 89 percentage point improvement—enabling post-deployment capability recovery. H-LoRA employs simplified scaling and strategic high-rank adaptation at approximately two-thirds of the model's hidden dimension ($r = 512$ for $d = 768$), achieving SFT-level domain performance (99.81% precision) with 5× greater parameter efficiency (20.35% trainable parameters) and robust cross-domain generalization (93.5 ± 6.8% average precision). In addition, H-LoRA reduces over-the-air (OTA) update size from 1.4 GB to 96 MB (≈93%), enabling practical and frequent deployment of specialized models in bandwidth-limited edge environments. Beyond demonstrating effectiveness, this work establishes the first comprehensive framework for characterizing specialization-retention trade-offs in parameter-efficient fine-tuning, providing practical guidance for method selection in real-world deployments.

KEYWORDS: LoRA; edge AI; knowledge retention; domain adaptation; parameter-efficient fine-tuning; catastrophic forgetting

1 Introduction

1.1 The Imperative for Specialized AI on the Edge

The rapid proliferation of edge computing has created an urgent demand for deploying specialized Artificial Intelligence (AI) models directly on resource-constrained devices, where memory, latency, and communication bandwidth are tightly limited. Unlike cloud-based AI systems that can leverage virtually unlimited computational resources, edge AI applications operate under stringent constraints that demand a fundamental rethinking of model architecture and optimization strategies. These deployment scenarios impose strict resource limitations: memory footprints must typically remain below 2 GB, inference latency

must be constrained to under 100 ms for real-time responsiveness, and energy consumption must be minimized for battery-powered devices [1–3]. Recent surveys on edge AI deployment have consistently identified these constraints as primary barriers to widespread adoption [4,5]. Consequently, the deployment of effective AI systems in these contexts necessitates compact language models with size of parameter typically limited to 1 billion or less.

However, the inherent trade-off between model compactness and task-specific performance presents a significant challenge. While general-purpose compact models can provide broad functionality, they often lack the specialized knowledge and fine-grained understanding required for domain-specific applications such as medical diagnosis, legal document analysis, or technical support systems for a specific company or region-specific knowledge. This limitation creates an urgent need for novel fine-tuning methodologies that can instill deep domain expertise into compact models without compromising the efficiency demanded by edge environments.

While recent large-scale language models demonstrate impressive performance, their computational and memory requirements remain prohibitive for many edge deployments (e.g., Internet of Things (IoT) gateways, embedded systems, and mobile devices). This work therefore focuses on sub-1B models, which represent the practical operating range for resource-constrained edge intelligence.

1.2 The Fine-Tuning Dilemma for Compact Models

Supervised Fine-Tuning (SFT) has long been regarded as the gold standard for achieving domain specialization in language models, demonstrating exceptional capability in adapting general-purpose models to specific application domains. However, when applied to compact models in edge scenarios, SFT reveals a critical vulnerability: it induces irreversible catastrophic forgetting that fundamentally compromises the model's general capabilities. This phenomenon, first systematically studied in connectionist networks [6,7], becomes particularly acute in resource-constrained models where the reduced parameter space amplifies interference effects [8]. This limitation is particularly acute in compact models where the reduced parameter space makes the trade-off between specialization and retention more pronounced, and the resulting knowledge destruction renders the model unsuitable for dynamic deployment scenarios that require adaptability across multiple domains or the ability to recover general capabilities post-deployment.

Traditional Low-Rank Adaptation (LoRA), while successful in mitigating catastrophic forgetting in larger models, faces substantial limitations when applied to compact architectures. Crucially, existing LoRA-based studies treat the rank primarily as a tunable compression hyperparameter, without examining its role in governing the nature of forgetting and knowledge retention under edge deployment constraints. The conventional approach of using conservative rank settings ($r \leq 64$) proves insufficient for achieving SFT-level specialization performance in resource-constrained models. This insufficient adaptation capacity stems from the fundamental mismatch between the reduced parameter space of compact models and the limited expressiveness of low-rank updates, creating a critical performance gap that has hindered the deployment of truly specialized compact models in edge applications. Moreover, the complexity of tuning the alpha (α) scaling parameter in traditional LoRA adds another layer of difficulty for practitioners working under resource constraints [9].

The current state of research presents a fundamental gap: no existing methodology enables compact models to achieve SFT-level domain performance while avoiding the destructive forgetting patterns that compromise their utility in dynamic edge environments. This gap represents not merely a technical limitation, but a barrier to understanding the fundamental trade-offs between specialization and retention that govern parameter-efficient fine-tuning.

1.3 A New Paradigm: Destructive Forgetting vs. Controllable Knowledge Retention

To address this gap, we introduce a new theoretical paradigm centered on the distinction between destructive forgetting and controllable knowledge retention. We observe that existing approaches exhibit two fundamentally different modes of capability change: destructive forgetting, characterized by irreversible degradation of the original knowledge structure, and controllable knowledge retention, where apparent performance degradation masks preserved underlying knowledge structures that remain recoverable and adjustable.

SFT exemplifies destructive forgetting through its direct manipulation of pre-trained weights, which overwrites the carefully learned representations that encode general knowledge. This process fundamentally destroys the model's original knowledge architecture, rendering any capability loss irreversible and eliminating the possibility of post-deployment adaptation. In contrast, our investigation reveals that properly configured high-rank adaptation can achieve controllable knowledge retention, where domain specialization occurs through additive low-rank modifications that maintain the structural integrity of the original knowledge subspace, providing an architectural guarantee for capability recovery and adjustment.

This paradigm shift builds upon recent advances in understanding LoRA's low-rank constraints but fundamentally challenges the conventional wisdom of rank limitations, particularly in the context of compact model deployment. The theoretical foundations of catastrophic forgetting [6] and neural network representation learning [10] provide crucial insights into why existing approaches fail to achieve controllable knowledge retention. To realize this vision, we introduce H-LoRA (High-Rank LoRA), a novel approach that combines strategic high-rank adaptation with simplified scaling mechanisms to unlock the full potential of compact models while maintaining knowledge recoverability and achieving superior specialization performance.

1.4 Our Contributions

This work makes the following key contributions to parameter-efficient fine-tuning for compact models under edge deployment constraints:

1. **H-LoRA for Edge AI Specialization.** We propose H-LoRA, a high-rank adaptation framework with a simplified scaling mechanism, specifically designed to enable efficient and stable specialization of compact models in resource-constrained edge environments.
2. **Destructive Forgetting vs. Controllable Knowledge Retention.** We introduce a clear conceptual distinction between destructive forgetting induced by supervised fine-tuning and controllable knowledge retention achieved through additive adaptation, supported by quantitative retention analysis.
3. **An Empirical Rank Selection Principle.** We identify a consistent empirical guideline showing that the optimal adaptation rank for compact models occurs at approximately two-thirds of the hidden dimension, providing practical guidance across different architectures.
4. **Practical Deployment Feasibility.** We demonstrate that H-LoRA achieves specialization performance comparable to full fine-tuning with significantly improved parameter efficiency, enabling multi-domain adaptation and reliable capability recovery for real-world edge deployments.

2 Related Work

This section critically examines three core areas of literature that situate the contributions of our work: the evolution and limitations of parameter-efficient fine-tuning for compact models, the prevailing understanding of catastrophic forgetting, and the need for a systematic framework to navigate specialization-retention trade-offs in resource-constrained environments.

We deliberately adopt standard LoRA as the primary baseline to isolate the effect of rank selection on knowledge retention. Introducing adaptive rank allocation (e.g., Adaptive Low-Rank Adaptation (AdaLoRA)) or quantization-based methods (e.g., Quantized Low-Rank Adaptation (QLoRA)) would introduce additional confounding factors such as dynamic sparsity patterns or quantization noise, making it difficult to attribute observed effects solely to rank. Therefore, this work does not aim to outperform all Parameter-Efficient Fine-Tuning (PEFT) variants, but to answer a more fundamental question: "How does rank selection affect knowledge retention in compact models?"

2.1 Parameter-Efficient Fine-Tuning for Compact Models

Parameter-Efficient Fine-Tuning (PEFT) has emerged as a computationally efficient alternative to full fine-tuning, building upon early work in transfer learning and domain adaptation [11,12]. Representative PEFT approaches include adapter-based methods such as Houlsby et al. [13], prompt tuning variants [14,15], and selective weight updates [16]. These methods aim to adapt pre-trained models with minimal parameter modifications, as summarized in recent surveys [17–19].

Among these approaches, Low-Rank Adaptation (LoRA) has become a *de facto* standard due to its simplicity and zero-overhead inference. LoRA is based on the hypothesis that task-specific weight updates lie in a low-rank subspace, allowing the update matrix to be decomposed as $\Delta W = BA$ [9], thereby significantly reducing the number of trainable parameters.

Several LoRA variants have been proposed to address different efficiency dimensions, including adaptive rank allocation (AdaLoRA; [20]), quantization-aware fine-tuning (QLoRA; [21]), weight decomposition (DoRA; [22]), vector-based random adaptation (VeRA; [23]), and learning-rate reparameterization (LoRA+; [24]). These approaches target complementary efficiency axes such as parameter budgeting, quantization, optimization stability, or decomposition strategies.

AdaLoRA dynamically redistributes a fixed parameter budget across layers based on importance estimation, focusing on budget allocation efficiency under constrained resources. QLoRA, in contrast, reduces memory consumption by applying low-bit quantization to pretrained weights while maintaining adaptation through low-rank updates.

While highly effective in their respective domains, these methods introduce additional mechanisms (e.g., dynamic sparsity patterns or quantization noise) that are orthogonal to the present study. The objective of this work is not to optimize parameter compression or allocation under a fixed budget, but to investigate a more fundamental question: *How does absolute rank magnitude influence the structural preservation or destruction of pretrained knowledge in compact models?*

To isolate this effect, we deliberately adopt standard LoRA as a controlled baseline. This ensures that observed differences arise from rank scaling itself rather than adaptive allocation strategies or quantization-induced regularization. Therefore, this study does not aim to outperform all PEFT variants in overall efficiency, but to uncover the mechanistic role of rank selection in governing controllable knowledge retention under edge deployment constraints.

2.2 Catastrophic Forgetting in Fine-Tuning

Catastrophic forgetting has long been recognized as a core challenge in neural network training, particularly in continual learning settings where new tasks interfere with previously acquired knowledge [7,25–27]. The phenomenon arises from distributed parameter representations, where gradient updates for new objectives can disrupt existing knowledge structures [6,28].

Prior work shows that forgetting severity is influenced by model scale, with compact models being more vulnerable to interference due to limited representational redundancy [29]. In domain adaptation, PEFT methods such as LoRA have been shown to reduce forgetting relative to full fine-tuning [9]. However, most studies quantify forgetting primarily through performance degradation metrics without distinguishing between fundamentally destructive overwriting and structurally controllable retention.

In contrast, our work emphasizes this distinction: destructive forgetting irreversibly alters pretrained representations, whereas controllable retention preserves the underlying knowledge structure and enables capability recovery. This structural perspective is particularly relevant for compact models deployed under edge constraints, where reliable recovery and update flexibility are essential.

2.3 The Specialization-Retention Trade-off: A Missing Framework

The trade-off between domain specialization and general knowledge retention is a well-documented but poorly formalized aspect of model fine-tuning. While numerous studies report varying performance-efficiency trade-offs for different PEFT methods, the field lacks a unified framework for systematic evaluation and decision-making. Practitioners are often left with *ad-hoc* heuristics for selecting a fine-tuning strategy, making it difficult to optimize for specific deployment scenarios.

This challenge is exacerbated in the context of Edge AI. The stringent resource constraints of edge devices [1,2] create a unique and unforgiving trade-off landscape, as comprehensively analyzed in recent surveys [4,5]. Decisions about parameter allocation, rank selection, and training strategy have direct and significant impacts on latency, memory usage, and energy consumption [30]. A method that offers a slight performance gain at the cost of a large increase in adapter size might be acceptable for a cloud deployment but entirely infeasible on a mobile device. The absence of a systematic framework for characterizing and navigating these fundamental trade-offs in domain adaptation represents a significant barrier to deploying robust and efficient specialized models at the edge.

3 Methodology

This section presents our H-LoRA framework and comprehensive experimental methodology. We first introduce the theoretical foundation and architectural design of H-LoRA in Section 3.1, followed by the detailed experimental setup and training procedures in Section 3.2. Section 3.3 describes the three domain-specific datasets and our domain specialization approach. Section 3.4 outlines our systematic investigation of rank selection principles across different model architectures. Finally, Section 3.5 details our evaluation framework, including metrics for both domain performance and knowledge retention assessment.

Problem Formulation

We consider the problem of adapting a pretrained compact language model $\mathcal{M}_{\text{base}}$ for domain-specific tasks on resource-constrained edge devices. The objective is to maximize domain performance while minimizing deployment cost, subject to strict constraints on memory, latency, and communication bandwidth. Formally, we aim to maximize task performance under a constraint on the size of model updates ΔW, such that the update payload remains feasible for over-the-air (OTA) transmission. This formulation highlights the need for adaptation mechanisms that balance specialization capability and knowledge retention without modifying the base model parameters.

3.1 H-LoRA: High-Rank Adaptation Framework

Traditional LoRA assumes that effective adaptation occurs within a low-dimensional subspace, motivated by the intrinsic dimensionality hypothesis. While this assumption is reasonable for large-scale models,

it can impose a severe adaptation bottleneck in compact architectures with limited parameter capacity. Recent theoretical analyses suggest that the optimal adaptation rank depends on task complexity and model structure rather than a universal low-rank constraint.

Motivated by this insight, H-LoRA explores the high-rank regime to restore sufficient adaptation capacity while preserving the pretrained knowledge structure. To simplify deployment and avoid rank-dependent tuning, we adopt a direct scaling mechanism that maintains the additive form of LoRA without relying on complex α/r relationships:

$$W = W_0 + \text{scale} \cdot (AB), \tag{1}$$

where W_0 denotes the frozen pretrained weights.

H-LoRA is applied uniformly to both attention and feed-forward layers within the Transformer architecture. Specifically, low-rank adaptation matrices $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times d}$ are injected such that the forward pass becomes

$$h = (W_0 + \text{scale} \cdot AB)x, \tag{2}$$

where x and h represent the input and output activations, respectively. This unified formulation preserves the mathematical simplicity of additive adaptation while enabling controllable high-rank updates suitable for resource-constrained edge deployment.

3.2 Experimental Design and Implementation

We conduct controlled experiments on compact decoder-only Transformer models to evaluate the impact of rank selection under edge deployment constraints. Our primary model is MiniMind, a 0.12B-parameter Transformer with a hidden dimension of 768 and 16 layers. To assess cross-architecture generalizability, we additionally evaluate Qwen-0.5B (hidden dimension 896) using the same experimental protocol.

All methods are trained under identical optimization settings to ensure fair comparison. We use the AdamW optimizer with a learning rate of 5×10^{-4}, a batch size of 12, and train all models for 60 epochs. Model selection is based on the best F1 score observed during training.

For H-LoRA, the low-rank adaptation matrices are initialized such that the effective weight update is zero at initialization: $A \sim \mathcal{N}(0, 0.02)$ and $B = 0$. This strategy preserves the pretrained model state at the start of adaptation and ensures stable and comparable optimization across methods.

3.3 Dataset Configuration and Domain Specialization

We evaluate H-LoRA across three heterogeneous domains to assess robustness under diverse specialization requirements: mathematical reasoning, professional knowledge, and rule-based enterprise scenarios. For mathematics, we use GSM8K, a benchmark dataset for multi-step numerical reasoning. For the medical domain, we adopt MedQA-USMLE-4-options, which evaluates domain-specific expert knowledge. In addition, we construct a domain-specific HR dataset comprising approximately 6K question–answer pairs covering employment regulations and workplace policies, which is manually verified to ensure data quality.

All datasets follow a unified training and evaluation protocol to ensure fair comparison across domains and architectures. This multi-domain design reduces the risk of domain-specific bias and enables a comprehensive assessment of controllable knowledge retention in edge deployment settings.

3.4 Rank Selection Investigation Methodology

This subsection outlines our systematic approach to investigating optimal rank selection, including the range of ranks tested and the methodology for identifying the empirical relationship between rank and model hidden dimension.

3.4.1 Systematic Rank Exploration Protocol

We conduct a systematic exploration of rank selection to characterize the performance landscape of H-LoRA. Specifically, we evaluate eight rank configurations, $r \in \{32, 64, 128, 256, 512, 640, 768, 1024\}$, covering low-, medium-, and high-rank regimes. Low ranks ($r \leq 128$) correspond to traditional LoRA assumptions, medium ranks ($128 < r \leq 512$) capture the transition region, and high ranks ($r > 512$) examine potential overfitting effects.

Based on this design, we test the hypothesis that near-optimal performance emerges when the rank is set to approximately two-thirds of the hidden dimension. For the MiniMind model ($d_{\text{model}} = 768$), this corresponds to $r \approx 512$, while for Qwen-0.5B ($d_{\text{model}} = 896$), the expected value is $r \approx 598$. Across experiments, we consistently observe near-optimal performance around this two-thirds ratio.

3.4.2 Cross-Architecture Validation Protocol

To assess cross-architecture generalizability, we validate the identified rank heuristic on Qwen-0.5B ($d_{\text{model}} = 896$). Following the two-thirds guideline, we evaluate a rank of $r = 598$ using the same human resources (HR) dataset as in the MiniMind experiments. This protocol allows us to examine whether the observed rank–retention relationship generalizes beyond the original architecture.

3.5 Evaluation Framework and Metrics

This subsection presents our comprehensive evaluation framework, including domain-specific performance metrics (precision, recall, F1-score) and knowledge retention assessment metrics (topic retention analysis) used to quantify the trade-off between specialization and forgetting.

3.5.1 Domain Specialization Performance Metrics

Primary Performance Indicators.

We evaluate domain specialization performance using standard classification and generation metrics. Specifically, we report precision ($P = \frac{TP}{TP+FP}$), recall ($R = \frac{TP}{TP+FN}$), and F1-score, defined as the harmonic mean of precision and recall, to measure prediction accuracy and coverage. In addition, we report exact match accuracy, which measures the percentage of responses that exactly match the reference answers.

Performance is evaluated separately for the HR, medical, and mathematics domains. We further report the mean and standard deviation across domains to assess aggregate performance and analyze cross-domain consistency.

3.5.2 Controllable Knowledge Retention Assessment

To assess controllable knowledge retention, we adopt a multi-dimensional evaluation framework designed to mitigate known biases in AI-based assessment, following best practices in experimental reporting. For each evaluation, we generate 100 diverse samples drawn from the original training data distribution, ensuring coverage across varying topics and complexity levels. Model outputs are compared

across three states: the pretrained base model, a standard supervised fine-tuned (SFT) model, and the proposed H-LoRA model.

Evaluation is conducted using Gemini 2.5 Pro as a consistent automated evaluator. Responses are scored on a five-point scale ranging from −2 to +2, where negative scores indicate degradation relative to the base model, zero indicates no change, and positive scores indicate quality improvement. Assessment dimensions include topic coherence, factual accuracy, logical reasoning integrity, and overall response quality.

3.5.3 Degradation Pattern Classification

To analyze knowledge degradation patterns, we define a concise taxonomy comprising topic drift, factual errors, logic errors, and other non-specific quality degradation, along with a maintained category indicating preserved or improved performance (score ≥0). Topic drift refers to responses that deviate from the input query, factual errors capture incorrect information, and logic errors denote flawed or inconsistent reasoning.

Degradation categories are identified using a hybrid approach, combining keyword-based automated detection for topic drift with manual expert verification to ensure classification accuracy. To improve reliability, multiple evaluators independently assess samples, and inter-rater consistency is verified across evaluations.

3.6 Computational Efficiency Analysis

3.6.1 Training Time Measurement Protocol

We measure computational efficiency using wall-clock training time, defined as the total duration from training start to convergence. All measurements are conducted on an identical NVIDIA RTX 3090 setup under consistent system load and thermal conditions. To reduce variance, reported results are averaged across multiple runs.

Efficiency is analyzed along three dimensions: training time scalability with respect to rank expansion, peak GPU memory usage across different rank settings, and convergence behavior measured by the number of epochs required to reach optimal performance.

3.6.2 Parameter Efficiency Calculation

Parameter efficiency is quantified by analyzing the number of trainable parameters updated during fine-tuning relative to the total model size, including frozen pretrained weights. We report the parameter ratio, defined as the percentage of parameters requiring gradient updates.

We define the efficiency score as:

$$E = \frac{P}{\rho},$$

where E denotes the efficiency score, P represents the downstream task precision, and ρ denotes the parameter ratio, defined as the proportion of trainable parameters relative to full fine-tuning.

To enable fair comparison across parameter-efficient methods, we compute an efficiency score defined as the ratio between the task performance metric and the parameter ratio. This normalized measure captures the trade-off between performance gains and parameter update cost.

3.7 Experimental Controls and Validation

3.7.1 Controlled Comparison Setup

To ensure a fair and controlled comparison, all methods are evaluated under identical training conditions, including the same hyperparameters, hardware, and software environments. We use consistent training, validation, and test splits across methods, and apply uniform evaluation metrics and assessment protocols. All models are trained with an equal computational budget and the same number of training epochs.

We compare the proposed H-LoRA against three baselines: the unmodified pretrained base model as a reference, a standard supervised fine-tuning (SFT) baseline, and traditional low-rank LoRA using conventional rank settings.

3.7.2 Statistical Validation and Reproducibility

Reproducibility is ensured through deterministic seeding with a fixed random seed of 42 and explicit specification of hardware and software configurations. Key experiments are repeated across multiple independent runs to validate result stability. We further perform statistical validation of performance differences, report confidence intervals for key metrics, and verify cross-domain consistency to ensure that observed trends generalize across tasks.

3.7.3 Ethical Considerations and Limitations

We adopt several measures to mitigate evaluation bias, including diverse domain coverage to reduce task-specific overfitting, automated evaluation to limit human subjectivity, and transparent reporting of all experimental procedures.

We acknowledge several limitations of this study. First, our primary evaluation focuses on compact models (≤1B parameters), which are most relevant for resource-constrained edge settings. Second, domain coverage is limited to three representative tasks. Finally, knowledge retention is assessed using AI-based evaluation, which, while consistent and scalable, may not fully capture all aspects of human judgment.

4 Results and Analysis

This section presents a comprehensive evaluation of H-LoRA, systematically dissecting its performance across different domain specializations, rank sensitivity, computational efficiency, and knowledge retention. The results collectively demonstrate that H-LoRA not only matches the specialization performance of Supervised Fine-Tuning (SFT) but does so with vastly superior knowledge retention and parameter efficiency, establishing it as a viable solution for edge deployment.

4.1 Cross-Domain and Cross-Architecture Performance

H-LoRA demonstrates consistently high performance across diverse domains and model architectures, confirming its robustness and generalizability. Fig. 1 provides a visual summary, while Table 1 details the comprehensive performance metrics.

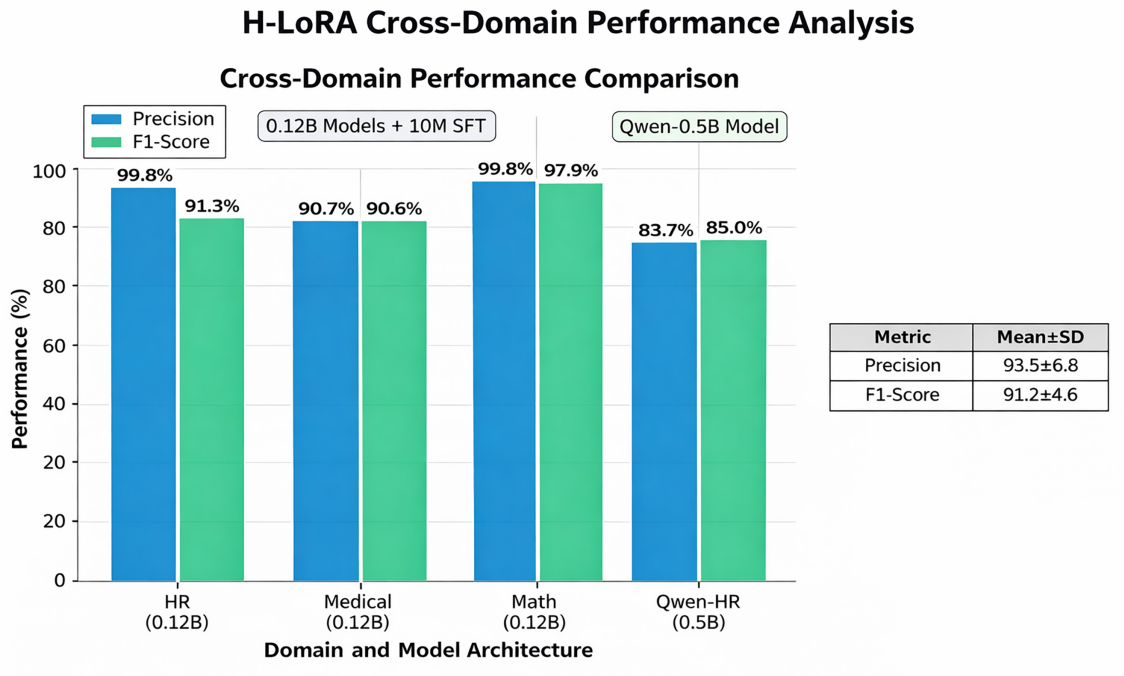

Figure 1: H-LoRA cross-domain performance analysis.

Table 1: Comprehensive performance comparison.

Method	Model	Rank (r)	Params (Trainable)	Precision (%)	Recall (%)	F1-Score (%)	EM (%)
HR H-LoRA	0.12B + 10M	512	25M	99.81	85.72	91.28	42
HR SFT	0.12B + 10M	N/A	123M	99.81	85.72	91.28	42
Medical H-LoRA	0.12B + 10M	512	25M	90.70	91.20	90.56	90
Math H-LoRA	0.12B + 10M	512	25M	99.75	96.55	97.86	84
Qwen-HR H-LoRA	Qwen-0.5B	598	25M	83.72	91.20	84.98	0

Domain-Specific Analysis

H-LoRA's effectiveness is evident across distinct reasoning types. In the Human Resources domain, it achieves an exceptional 99.81% precision, matching SFT's performance with only 20.35% of the trainable parameters. On the Medical (MedQA) dataset, its high exact match rate (90%) underscores its reliability for factual recall tasks. In the Mathematics (GSM8K) domain, an F1-score of 97.86% highlights its proficiency in complex, multi-step logical reasoning.

Cross-Architecture Validation

To test the architectural generalizability of our findings, we applied H-LoRA to Qwen-0.5B. Using the theoretically predicted rank ($r = 598 \approx 2/3 \times 896$), H-LoRA achieved 83.72% precision on the HR dataset. This result confirms that the principles of H-LoRA successfully translate to different model architectures, enabling significant domain adaptation over the base model.

Statistically, H-LoRA maintains an average precision of 93.5 ± 6.8% and an average F1-score of 91.2 ± 4.6% across all configurations, indicating robust and consistent performance with manageable variance across diverse specialization tasks.

4.2 The Effect of Rank: Discovery of the 2/3 Empirical Principle

Our systematic exploration of LoRA rank reveals a predictable, three-phase performance pattern, as illustrated in Fig. 2. This pattern challenges the conventional low-rank assumption and leads to the discovery of a practical empirical principle for rank selection.

Phase I: Rapid Growth ($r \leq 128$). As shown in Fig. 2, performance dramatically improves with rank (from 43.34% to 97.96% precision), confirming that traditional low-rank settings ($r \leq 64$) create a severe adaptation capacity bottleneck in compact models.

Phase II: Diminishing Returns ($128 < r \leq 512$). The model's performance continues to climb but at a decelerating rate, reaching its optimal point at $r = 512$, which achieves 99.81% precision.

Phase III: Overfitting ($r > 512$). Beyond this peak, performance begins to degrade, confirming the existence of an optimal rank rather than a monotonic improvement.

This analysis yields our most significant finding for practitioners: optimal performance is consistently achieved when the rank is approximately two-thirds of the model's hidden dimension ($r \approx 2/3 \times d_{\text{model}}$). This empirical principle was validated across our testbeds (0.12B model: $512 \approx 2/3 \times 768$; Qwen-0.5B: $598 \approx 2/3 \times 896$), providing a robust heuristic that eliminates the need for extensive hyperparameter tuning.

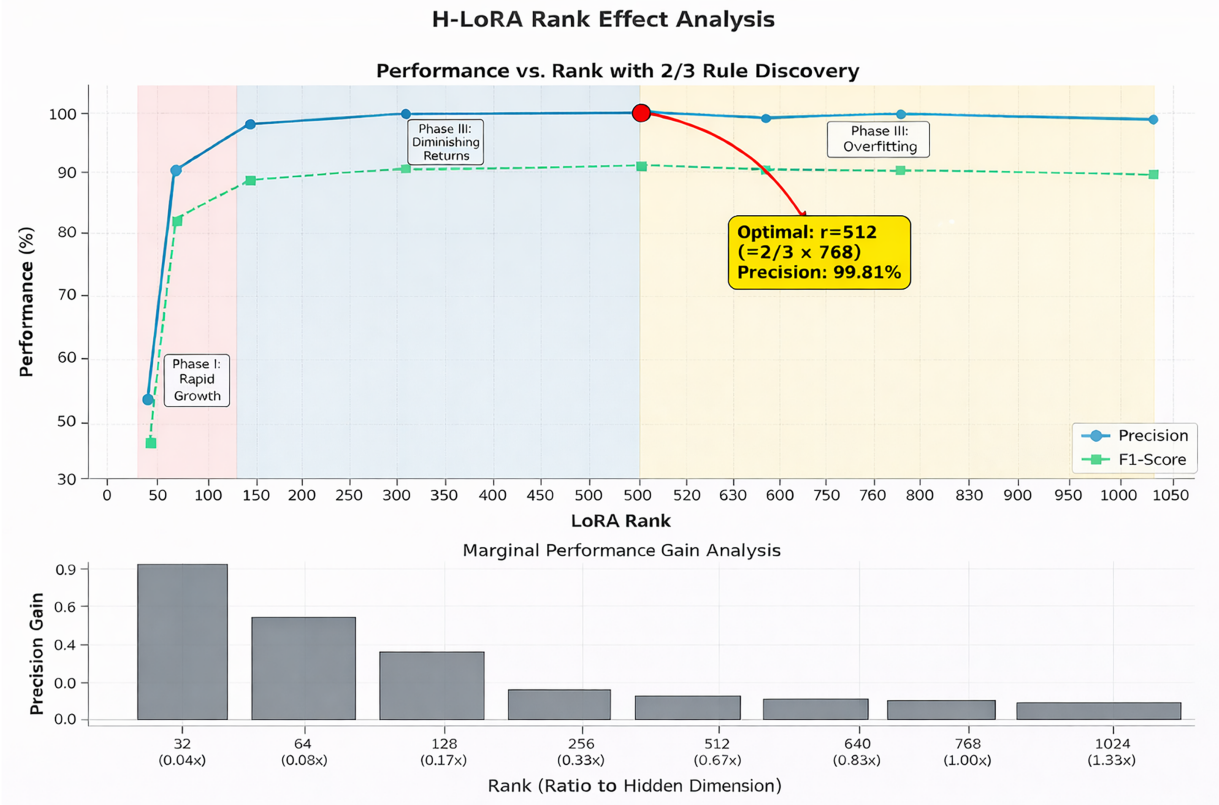

Figure 2: H-LoRA rank effect analysis.

4.3 Computational Efficiency for Edge Deployment

H-LoRA is not only effective but also remarkably efficient, making it ideally suited for resource-constrained edge environments. Despite a 32-fold rank increase (r = 32 to r = 1024), total training time grew by a mere 8.6% (5.23 to 5.68 h). The optimal configuration (r = 512) incurred only a 2.7% training overhead compared to a low-rank baseline, confirming high-rank adaptation is computationally feasible.

As shown in Table 2, H-LoRA achieves a 5× improvement in parameter efficiency over SFT, quantified by an efficiency score of 4.9 vs. 0.998 while precision remains. By requiring only 25M trainable parameters (20.35%) to achieve the same performance as SFT's 123M, H-LoRA provides a better option to deploying highly specialized models without prohibitive memory or storage costs.

Table 2: Parameter efficiency comparison.

Method	Trainable Params	Parameter Ratio (%)	HR Precision (%)	Efficiency Score
SFT	123M	100.00	99.81	0.998
H-LoRA	25M	20.35	99.81	4.900

4.4 Distinguishing Forgetting: Controllable Retention vs. Destructive Forgetting

Our analysis reveals the fundamental mechanistic difference between H-LoRA's and SFT's impact on a model's general knowledge. While both induce performance degradation on out-of-domain tasks, H-LoRA facilitates controllable knowledge retention, whereas SFT causes destructive, irreversible forgetting.

Topic Coherence as the Key Differentiator: The most striking evidence is the 89 percentage point difference in topic retention shown in Fig. 3. H-LoRA maintains topic coherence in 90% of cases, while SFT's coherence collapses to just 1%. This indicates that H-LoRA preserves the model's core knowledge structure, even when its reasoning is flawed.

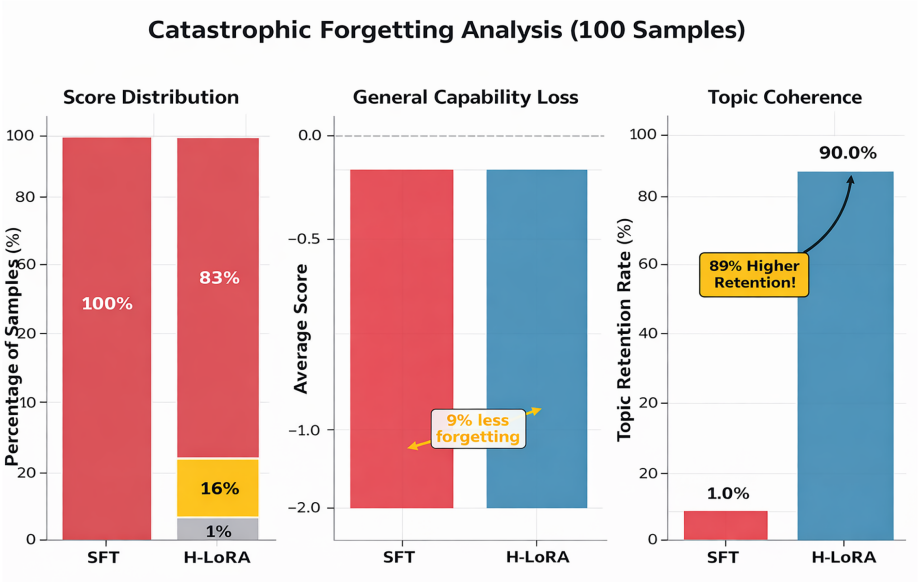

Figure 3: Catastrophic forgetting analysis (100 samples). This figure provides the central evidence for our theoretical distinction, demonstrating that H-LoRA preserves knowledge structure (90% topic coherence) while SFT destructively overwrites it (1% coherence).

Fundamentally Different Degradation Patterns: Table 3 shows that SFT's failure mode is almost exclusively topic drift (99%), indicating a complete disconnection from the query's intent. In contrast, H-LoRA exhibits a diverse error profile, primarily composed of logic (43%) and factual (33%) errors. This pattern signifies that the model is still "on topic" but failing in its reasoning, a functional impairment that is structurally preserved and potentially recoverable, unlike SFT's destructive knowledge overwriting.

Table 3: Catastrophic forgetting – degradation pattern analysis.

Model	Avg Score	Topic Drift (%)	Factual Error (%)	Logic Error (%)	Other Degradation (%)	Maintained (%)	Total
SFT	−2.00	99	0	0	0	1	100%
H-LoRA	−1.85	10	33	43	13	1	100%

Human Validation of Topic Drift

To verify that the observed destructive forgetting pattern is not an artifact of automated evaluation, we conducted an independent human validation study.

Two annotators independently evaluated 40 randomly sampled outputs (80 total judgments). Each response was classified into one of two categories: *Topic Drift* or *On-topic Degradation*.

Table 4 reports the aggregated results.

Table 4: Human validation of topic drift (40 samples, 2 annotators).

Model	Topic Drift (Count)	Topic Drift (%)
Pretrained	26	32.5
SFT	72	90.0
H-LoRA	36	45.0

The findings confirm that SFT exhibits substantially higher topic drift compared to H-LoRA, consistent with the automated evaluation results. Specifically, SFT shows topic drift in 90% of cases, whereas H-LoRA reduces topic drift to 45%, representing a 45-percentage-point improvement in topical coherence.

This human assessment reinforces the conclusion that SFT primarily induces destructive forgetting, while H-LoRA preserves the underlying knowledge structure.

4.5 The Specialization-Retention Trade-off Framework

By synthesizing our findings, we can construct a comprehensive framework to visualize the trade-off landscape between domain specialization and knowledge retention.

As illustrated in Fig. 4, SFT and H-LoRA occupy distinct regions in the performance space. Both achieve the same high level of specialization (~91% F1-score), but SFT does so at the cost of severe forgetting (−2.0 score), placing it in the "Poor Zone." H-LoRA achieves an optimal balance, delivering equivalent specialization with 9% less forgetting, positioning it at the edge of the "Ideal Zone."

Fig. 5 provides a practical demonstration of this framework, showing that applying H-LoRA to the general-purpose Qwen model results in performance improvement for 95% of domain-specific samples, with 73% showing significant gains. This validates H-LoRA's ability to enhance existing models while maintaining their underlying capabilities.

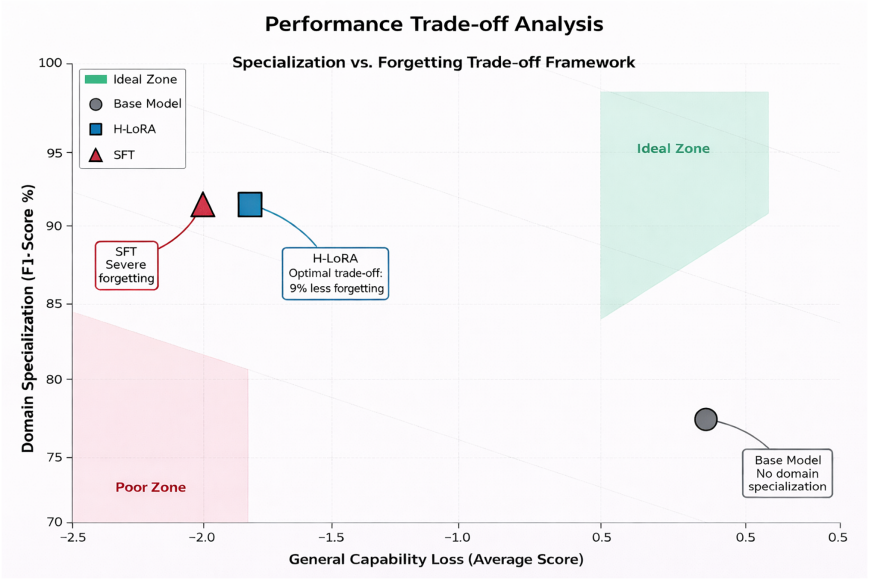

Figure 4: Performance trade-off analysis. The framework visually confirms H-LoRA's optimal positioning, achieving SFT-level specialization while mitigating severe forgetting, thereby placing it at the edge of the 'Ideal Zone'.

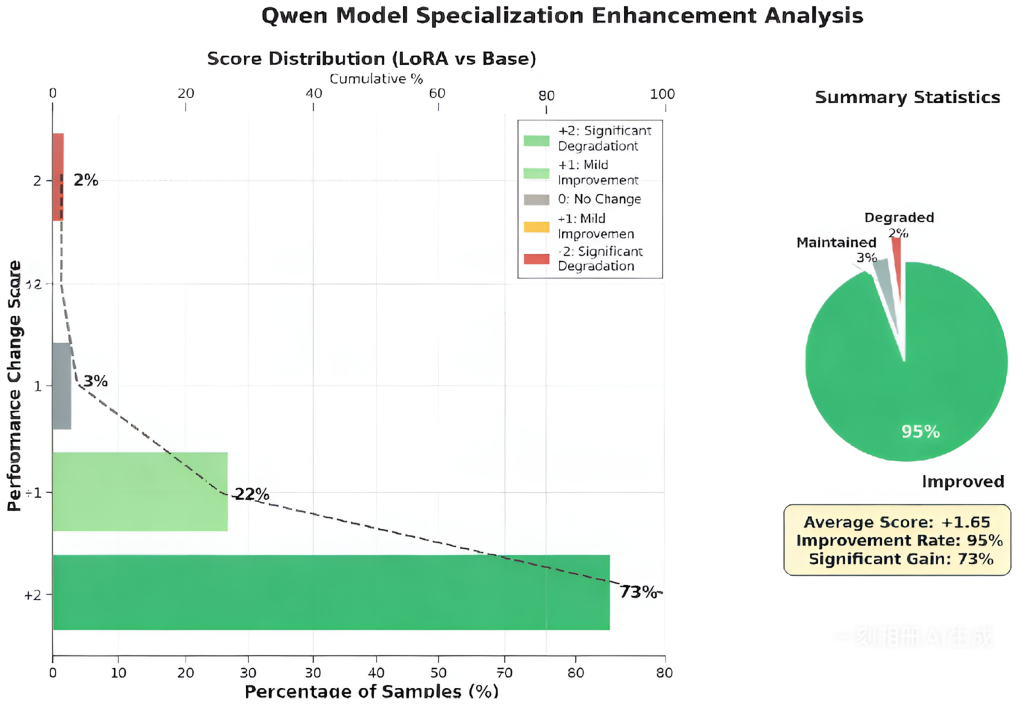

Figure 5: Qwen model specialization enhancement analysis.

4.6 Edge Deployment Analysis: Memory Efficiency and Operational Flexibility

H-LoRA provides quantifiable advantages for practical edge AI deployment by addressing key constraints in memory usage, storage footprint, and update efficiency.

4.6.1 Memory and Storage Efficiency

Inference memory is a primary constraint for edge hardware. As shown in Table 5, activating an H-LoRA adapter (r = 512) on the 0.12B model increases inference Video Random Access Memory (VRAM) from 547.85 to 643.85 MB, corresponding to a modest 17% overhead. This confirms that high-rank adaptation does not impose a prohibitive memory burden during inference.

Table 5: Inference VRAM analysis (0.12B model).

Inference Configuration	Trainable Params	Inference VRAM (MB)	Overhead
Base Model (0.12B)	0	547.85	–
SFT Specialized Model (0.12B)	123M	547.85	0% (replaces base)
Base Model + H-LoRA Adapter	25M	643.85	+17%

More significant gains emerge when considering storage and update requirements for practical-scale models. As summarized in Table 6, domain specialization via SFT requires storing a full 1.4 GB model per domain, whereas H-LoRA encapsulates equivalent domain knowledge in a lightweight 96 MB adapter, yielding a ~93% reduction in storage and update size.

Table 6: Storage & update size analysis (1.4 GB practical-scale model).

Component	File Size (Disk)	Purpose/Update Payload
Base Model (1.4 GB)	~1.4 GB	General capabilities foundation
SFT Specialized Model	~1.4 GB	A single, monolithic domain skill
H-LoRA Adapter	~96 MB	A lightweight, portable domain skill

This dramatic reduction directly impacts over-the-air (OTA) updates in network-constrained environments, where transmitting a 96 MB adapter is substantially more feasible than distributing a full 1.4 GB model.

4.6.2 Operational Flexibility and Deployment Cost Implications

H-LoRA's non-destructive, additive architecture enables flexible and reliable deployment strategies that are difficult to achieve with full fine-tuning. Because the base model weights remain unchanged, unloading an adapter guarantees immediate recovery of the model's general capabilities without empirical validation.

Furthermore, a single base model can support multiple domain-specific adapters stored on disk and dynamically loaded at runtime, enabling true multi-domain capability on resource-constrained devices with minimal VRAM overhead.

To contextualize the practical impact of adapter size, we provide a simple transmission estimate under realistic network conditions. Assuming a conservative edge connectivity bandwidth of 5 Mbps (e.g., constrained LTE or rural deployment scenarios), transmitting a 96 MB adapter requires approximately 150–160 s. In contrast, distributing a full 1.4 GB fine-tuned model would require over 30 min under the same conditions.

This order-of-magnitude difference substantially improves the feasibility of frequent over-the-air (OTA) updates, particularly in bandwidth-limited or intermittently connected environments. While 96 MB is larger

than conventional low-rank LoRA adapters, it remains within a practically deployable range for modern edge networks, whereas full-model replacement becomes operationally prohibitive.

Beyond memory and bandwidth efficiency, this design mitigates hidden operational costs. Unlike SFT, which often requires repeated retraining to balance specialization and retention, H-LoRA's controllable adaptation reduces retraining cycles and simplifies deployment at scale, particularly in large fleets of edge devices with limited connectivity.

4.7 Key Findings and Implications

Our comprehensive evaluation establishes several critical findings that advance both theoretical understanding and practical deployment:

1. **Performance Equivalence with Superior Efficiency:** H-LoRA matches SFT's specialization performance while requiring ~93% less storage for specialized weights (on our 1.4 GB model) and incurring a modest +17% VRAM overhead during inference (on our 0.12B model).
2. **Architectural Generalizability:** The 2/3 principle provides a universal heuristic for rank selection across different model architectures, eliminating hyperparameter search complexity while achieving consistent near-optimal performance.
3. **Controllable vs. Destructive Forgetting:** The 89-point difference in topic retention (90% vs. 1%) reveals fundamentally different adaptation mechanisms. H-LoRA's preservation of knowledge structure enables dynamic capability access, while SFT's destructive approach requires costly mitigation strategies.
4. **Edge Deployment Viability:** With a lightweight 96 MB storage footprint and a ~93% reduction in update bandwidth, H-LoRA directly addresses the most critical hardware and network constraints of Edge AI deployment.
5. **Framework for Trade-off Navigation:** Our specialization-retention framework provides practitioners with quantitative tools for method selection based on deployment requirements, moving beyond ad-hoc heuristics to principled decision-making.
6. **Computational Feasibility of High-Rank Adaptation:** The minimal training overhead (2.7% for optimal rank) challenges conventional assumptions about rank limitations, demonstrating that high-rank adaptation is both effective and efficient for compact models.

These results collectively demonstrate that H-LoRA successfully bridges the critical gap between high-performance domain specialization and knowledge preservation, establishing a new paradigm for parameter-efficient adaptation specifically suited for edge AI deployment scenarios where both specialization performance and operational flexibility are paramount.

5 Discussion

This section examines the broader implications of our findings, positioning H-LoRA within the evolving landscape of parameter-efficient fine-tuning and edge AI deployment. We analyze the theoretical contributions, mechanistic insights, and practical ramifications that extend beyond the immediate experimental results to influence future research directions and deployment strategies.

We evaluate H-LoRA across diverse task types, including logical reasoning (GSM8K), domain-specific expert knowledge (MedQA), and rule-driven enterprise scenarios (Human Resources (HR)), demonstrating robustness across heterogeneous edge application settings.

5.1 Rethinking the Low-Rank Paradigm in Compact Models

Our findings challenge the conventional assumption that effective adaptation necessarily occurs in a strictly low-rank subspace. While low-rank adaptation has proven sufficient for large-scale models where even modest ranks provide ample capacity, our results indicate that this assumption becomes a limiting factor for compact models under edge constraints. In such settings, overly restrictive rank choices can severely hinder domain adaptation and exacerbate catastrophic forgetting.

Importantly, the observed 2/3 hidden-dimension rule should be interpreted as an empirical heuristic rather than a theoretical guarantee. We hypothesize that this phenomenon reflects the minimum representational capacity required to preserve pretrained knowledge while enabling effective specialization under low-rank constraints. This interpretation aligns with general notions of representational capacity and regularization, without assuming a universal optimal rank across architectures or scales.

Implications for PEFT Design under Edge Constraints

The success of H-LoRA suggests that parameter efficiency should not be equated with aggressive parameter reduction. Instead, effective adaptation in compact models requires allocating sufficient capacity to preserve critical subspaces while maintaining deployment feasibility. From this perspective, rank selection becomes a controllable design parameter rather than a fixed hyperparameter.

The proposed heuristic offers a practical guideline for reducing extensive hyperparameter search in edge deployments. Future work may explore integrating such heuristics into automated or adaptive rank selection mechanisms tailored to heterogeneous edge devices.

5.2 Mechanistic Insight: Additive vs. Overwrite Adaptation

The stark contrast in knowledge retention (90% vs. 1% topic coherence) arises from a fundamental difference in how adaptation is implemented at the weight level.

Overwrite Paradigm (SFT). Supervised fine-tuning updates all pretrained parameters directly:

$$W_{\text{final}} = W_{\text{pretrained}} + \Delta W_{\text{full}}.$$

Gradients $\partial L/\partial W$ flow through the entire parameter space, allowing unrestricted modification of pretrained representations. While effective for domain optimization, this unconstrained update can overwrite previously learned knowledge structures, a phenomenon widely associated with catastrophic forgetting [6,7].

Additive Paradigm (H-LoRA). In contrast, H-LoRA constrains adaptation to a dedicated low-rank subspace:

$$W_{\text{final}} = W_{\text{pretrained}} + \text{scale} \cdot (AB).$$

Gradients are restricted to adapter parameters ($\partial L/\partial A$, $\partial L/\partial B$), while pretrained weights remain frozen ($\nabla W_{\text{pretrained}} = 0$). This architectural isolation prevents interference with foundational representations and ensures that general knowledge remains structurally preserved.

From a representational perspective, SFT modifies the original knowledge space itself, whereas H-LoRA augments it with an additional subspace ($\text{span}(AB)$). Because the pretrained weights remain intact, deactivating the adapter restores the original capability without retraining. This structural separation provides a direct mechanistic explanation for the observed difference between destructive forgetting and controllable knowledge retention.

5.3 Gradient Flow Analysis and Optimization Dynamics

The distinction between these paradigms becomes particularly clear when analyzing optimization dynamics:

Unconstrained Gradient Flow (SFT): With gradients flowing through all 123M parameters simultaneously, the optimization process faces conflicting objectives—minimizing domain-specific loss while preserving general capabilities. This multi-objective optimization challenge is well-documented in the optimization literature, where gradient conflicts typically resolve through implicit prioritization of the most recent objective function. This multi-objective optimization typically resolves in favor of the immediate training objective, sacrificing general knowledge for task performance [31].

Constrained Gradient Flow (H-LoRA): By restricting gradients to only 25M adapter parameters, H-LoRA eliminates this optimization conflict [32,33]. The base model's representational capacity remains untouched, while domain-specific knowledge is encoded in dedicated parameters designed specifically for this purpose [13,34].

5.4 Implications for Neural Network Adaptation Theory

This mechanistic understanding establishes a theoretical foundation for controllable neural adaptation where specific capabilities can be modified without disrupting the broader knowledge ecosystem within the model. The additive paradigm suggests that effective adaptation does not require access to all model parameters; instead, it requires architectural guarantees that adaptation occurs in dedicated parameter subspaces.

This insight challenges conventional assumptions in neural network fine-tuning and suggests that future adaptation methods should prioritize architectural preservation over parameter accessibility. The success of high-rank additive updates demonstrates that adaptation capacity is not fundamentally limited by rank constraints, but rather by the intelligent allocation of adaptation parameters within preserved representational structures.

Architectural Adaptation vs. Inference-Time Reflection.

Recent work has explored inference-time self-correction mechanisms, such as self-reflection and dual-loop reasoning strategies, to improve response quality without modifying model parameters. These approaches enhance reasoning reliability by iteratively critiquing and refining generated outputs during inference.

While effective for improving answer accuracy, reflection-based methods operate at the inference level and do not alter the underlying parameter structure of the model. Consequently, they do not address the structural overwriting that occurs during supervised fine-tuning, where pretrained representations may be irreversibly modified.

In contrast, the focus of this work lies in training-time architectural adaptation. H-LoRA prevents destructive forgetting by isolating domain-specific updates within dedicated parameter subspaces, thereby preserving the pretrained representational geometry. Reflection improves reasoning behavior post hoc, whereas architectural isolation governs whether foundational knowledge is structurally preserved during adaptation.

These two directions are therefore complementary rather than competing: architectural preservation ensures recoverability of general knowledge, while inference-time reflection may further enhance reasoning quality on top of a structurally preserved backbone.

Broader Implications: This framework extends beyond LoRA-based methods to inform the design of any parameter-efficient adaptation technique [15,18]. Methods that preserve foundational representational spaces while enabling targeted capability enhancement will likely achieve superior controllability and knowledge retention compared to approaches that modify pre-trained parameters directly [15,18].

5.5 Edge AI Paradigm: From Cloud-Centric to Edge-Native Design

The results of this study highlight the importance of designing adaptation mechanisms that explicitly account for the constraints of edge deployment. Unlike cloud-centric fine-tuning assumptions, edge environments require frequent updates, limited storage, and reliable fallback mechanisms.

By decoupling domain-specific adaptation from the base model, H-LoRA enables lightweight specialization while preserving a shared pretrained backbone. This design facilitates practical multi-domain deployment on resource-constrained devices, where maintaining multiple full models would be infeasible.

Moreover, the ability to activate or deactivate adapters without modifying the base model provides a robust mechanism for capability recovery. Such architectural separation is particularly valuable in autonomous or safety-critical edge scenarios, where preserving general-purpose functionality is essential.

5.6 Broader Implications

While this study focuses on compact models under edge constraints, the findings provide broader insight into the nature of neural adaptation. The distinction between destructive and controllable forgetting suggests that adaptation mechanisms should be evaluated not only by downstream task performance but also by their structural impact on pretrained representations.

This architectural perspective may inform research in continual learning and adaptive systems, where preserving representational geometry can be more valuable than minimizing short-term performance degradation. The results highlight that parameter isolation, rather than aggressive compression, may serve as a foundational principle for controllable model specialization.

More broadly, the findings encourage a shift from purely performance-centric evaluation toward structural analysis of knowledge preservation during adaptation.

5.7 Limitations and Future Research Directions

This study focuses on controllable adaptation in compact models under edge deployment constraints. Our empirical validation is therefore limited to models of at most one billion parameters, which represent the practical operating regime for memory- and bandwidth-constrained edge devices.

Importantly, the observed 2/3 rank heuristic should be interpreted within this compact-model regime. Extremely large-scale models (e.g., 1B–3B parameters and beyond) exhibit substantially higher representational redundancy and over-parameterization, which may alter effective adaptation capacity requirements. In such settings, lower ranks may already provide sufficient expressive flexibility, and the relationship between hidden dimension and optimal rank may follow a different scaling behavior. Therefore, we do not claim universality of the 2/3 principle across all model scales.

While preliminary cross-architecture validation on Qwen-0.5B suggests generality within the compact regime, extending the analysis to larger-scale models would provide valuable insight into potential scaling transitions in rank–retention dynamics and remains an important direction for future research.

In addition, this work isolates the effect of absolute rank magnitude under standard LoRA to study knowledge retention mechanisms. Integrating high-rank adaptation with quantization-aware techniques

such as QLoRA represents a promising future direction to jointly optimize structural knowledge preservation and memory efficiency in practical edge deployment.

Finally, this work intentionally addresses system- and application-level adaptation mechanisms. We do not model physical-layer communication effects or long-term deployment dynamics, as our goal is to study controllable specialization and deployment efficiency rather than wireless transmission or security optimization.

6 Conclusion

This work introduced H-LoRA, a high-rank adaptation framework designed to balance domain specialization and knowledge retention for compact models under edge deployment constraints. By challenging the conventional low-rank assumption in parameter-efficient fine-tuning, we demonstrated that strategically increasing rank enables SFT-level specialization while preserving foundational knowledge structures.

We established a clear distinction between destructive overwrite (SFT) and controllable additive adaptation (H-LoRA). Empirical results show that H-LoRA maintains 90% topic coherence—an 89 percentage point improvement over SFT—while matching SFT's 99.81% precision with 5× greater parameter efficiency. In deployment terms, this translates into a 93% reduction in storage and update size, with only a 17% VRAM overhead during inference.

The proposed 2/3 empirical guideline provides a practical heuristic for rank selection in compact models, reducing the need for extensive hyperparameter tuning. While further validation across larger model scales remains an open direction, our findings demonstrate that high-rank additive adaptation offers a controllable and deployment-ready pathway for edge AI specialization.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Darren Chai Xin Lun; Methodology, Darren Chai Xin Lun; Software, Darren Chai Xin Lun; Validation, Darren Chai Xin Lun; Formal analysis, Darren Chai Xin Lun, Lim Tong Ming; Investigation, Darren Chai Xin Lun; Data curation, Darren Chai Xin Lun; Writing—original draft preparation, Darren Chai Xin Lun; Writing—review and editing, Lim Tong Ming; Visualization, Darren Chai Xin Lun; Supervision, Lim Tong Ming; Project administration, Lim Tong Ming. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The code and training scripts that support the findings of this study are openly available in GitHub at https://github.com/darrencxl0301/Lora-SLM. The Mathematics domain dataset is openly available at Hugging Face (GSM8K) at https://huggingface.co/datasets/openai/gsm8k. The Medical domain dataset is openly available at Hugging Face (MedQA-USMLE-4-options) at https://huggingface.co/datasets/GBaker/MedQA-USMLE-4-options. The Human Resources domain dataset was synthetically generated by the authors for this study and is available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Deng L, Li G, Li S, Liu L, Yu FR. Model compression and hardware acceleration for neural networks: a comprehensive survey. Proc IEEE. 2020;108(4):485–532. doi:10.1109/JPROC.2020.2976475.

2. Li E, Zeng L, Zhou Z, Chen X. Edge AI: on-demand accelerating deep neural network inference via edge computing. IEEE Trans Wirel Commun. 2018;19(1):447–57. doi:10.1109/TWC.2019.2946140.
3. Shi W, Cao J, Zhang Q, Li Y, Xu L. Edge computing: vision and challenges. IEEE Internet Things J. 2016;3(5):637–46. doi:10.1109/JIOT.2016.2579198.
4. Wang X, Han Y, Leung VC, Niyato D, Yan X, Chen X. Convergence of edge computing and deep learning: a comprehensive survey. IEEE Commun Surv Tutor. 2020;22(2):869–904. doi:10.1109/COMST.2020.2970550.
5. Zhou Z, Chen X, Li E, Zeng L, Luo K, Zhang J. Edge intelligence: paving the last mile of artificial intelligence with edge computing. Proc IEEE. 2019;107(8):1738–62. doi:10.1109/JPROC.2019.2918951.
6. French RM. Catastrophic forgetting in connectionist networks. Trends Cogn Sci. 1999;3(4):128–35. doi:10.1016/S1364-6613(99)01294-2.
7. McCloskey M, Cohen NJ. Catastrophic interference in connectionist networks: the sequential learning problem. Psychol Learn Motiv. 1989;24:109–65. doi:10.1016/S0079-7421(08)60536-8.
8. Robins A. Catastrophic forgetting, rehearsal and pseudorehearsal. Conn Sci. 1995;7(2):123–46. doi:10.1080/09540099550039318.
9. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. Low-rank adaptation of large language models. arXiv:2106.09685. 2021.
10. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013;35(8):1798–828. doi:10.1109/TPAMI.2013.50.
11. Ben-David S, Blitzer J, Crammer K, Pereira F. Analysis of representations for domain adaptation. In: NIPS'06: Proceedings of the 20th International Conference on Neural Information Processing Systems; 2006 Dec 4–7; Vancouver, BC, Canada. Cambridge, MA, USA: MIT Press; 2006. p. 137–44. doi:10.5555/2976456.2976474.
12. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng. 2010;22(10):1345–59. doi:10.1109/TKDE.2009.191.
13. Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, De Laroussilhe Q, Gesmundo A, et al. Parameter-efficient transfer learning for NLP. In: Proceedings of the 36th International Conference on Machine Learning (ICML). London, UK: PMLR; 2019. p. 2790–9.
14. Li XL, Liang P. Prefix-tuning: optimizing continuous prompts for generation. arXiv:2101.00190. 2021. doi:10.48550/arXiv.2101.00190.
15. Liu X, Zheng Y, Du Z, Ding M, Qian Y, Yang Z, et al. GPT understands, too. AI Open. 2022;3(120):40–53. doi:10.1016/j.aiopen.2022.10.001.
16. Ben Zaken E, Goldberg Y, Ravfogel S. BitFit: simple parameter-efficient fine-tuning for transformer-based masked language-models. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL). Stroudsburg, PA, USA: ACL; 2022. p. 1–9. doi:10.18653/v1/2022.acl-short.1.
17. Lialin V, Deshpande V, Yao X, Rumshisky A. Scaling down to scale up: a guide to parameter-efficient fine-tuning. arXiv:2303.15647. 2023. doi:10.48550/arXiv.2303.15647.
18. Ding N, Qin Y, Yang G, Wei F, Yang Z, Su Y, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. Nat Mach Intell. 2023;5(3):220–35. doi:10.1038/s42256-023-00626-4.
19. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv. 2023;55(9):1–35. doi:10.1145/3560815.
20. Zhang R, Zhang A, Dettmers T, Zettlemoyer L. AdaLoRA: adaptive budget allocation for parameter-efficient fine-tuning. arXiv:2303.10512. 2023.
21. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: efficient finetuning of quantized LLMs. arXiv:2305.14314. 2023. doi:10.48550/arXiv.2305.14314.
22. Liu S, Han C, Zhao Z, Zhang Z, Zhou H. DoRA: weight-decomposed low-rank adaptation. arXiv:2402.09353. 2024. doi:10.48550/arXiv.2402.09353.
23. Kopiczko D, Blankevoort T, Welling M. VeRA: vector-based random matrix adaptation. arXiv:2310.11454. 2024.
24. Hayou S, Ghosh N, Yu B. LoRA+: efficient low rank adaptation of large models. In: Proceedings of the 41st International Conference on Machine Learning (ICML). London, UK: PMLR; 2024. p. 17783–806.

25. Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu AA, et al. Overcoming catastrophic forgetting in neural networks. Proc Natl Acad Sci USA. 2017;114(13):3521–6. doi:10.1073/pnas.1611835114.
26. Luo Y, Yang Z, Meng F, Li Y, Zhou J, Zhang Y. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. arXiv:2308.08747. 2023. doi:10.48550/arXiv.2308.08747.
27. Parisi GI, Kemker R, Part JL, Kanan C, Wermter S. Continual lifelong learning with neural networks: a review. Neural Netw. 2019;113(2):54–71. doi:10.1016/j.neunet.2019.01.012.
28. Goodfellow IJ, Mirza M, Xiao D, Courville A, Bengio Y. An empirical investigation of catastrophic forgetting in gradient-based neural networks. arXiv:1312.6211. 2013. doi:10.48550/arXiv.1312.6211.
29. Ramasesh VV, Lewkowycz A, Dyer E. Effect of scale on catastrophic forgetting in neural networks. 2022 [cited 2026 Jan 23]. Available from: https://openreview.net/forum?id=GhVS8_yPeEa.
30. Xu J, Chen Z, Quach TT, Yang K, Chiang P. EdgeBERT: optimizing on-device deep NLP inference. arXiv:1907.00019. 2019. doi:10.48550/arXiv.1907.00019.
31. Yu T, Kumar S, Gupta A, Hausman K, Levine S, Finn C. Gradient surgery for multi-task learning. arXiv:2001.06782. 2020.
32. He J, Zhou C, Ma X, Berg-Kirkpatrick T, Neubig G. Towards a unified view of parameter-efficient transfer learning. arXiv:2110.04366. 2022.
33. Pfeiffer J, Vulić I, Gurevych I, Ruder S. MAD-X: an adapter-based framework for multi-task cross-lingual transfer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: ACL; 2020. p. 7654–73. doi:10.18653/v1/2020.emnlp-main.617.
34. Kalajdzievski D. A rank stabilization scaling factor for fine-tuning with LoRA. arXiv:2312.03732. 2023. doi:10.48550/arXiv.2312.03732.