IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, April 2026
Impact Factor: 8.2

Efficient Lightweight Reasoning in Large Language Models using LoRA and Quantization: A Study on Qwen-based Micro-Reasoner

Anuja Phaphale[1], Karan Nigal[2], Ashish Kharde[3], Manish Narkhede[4]
Professor, Information Technology[1]
Students, Information Technology[2-4]
AISSMS Institute of Information Technology, Pune, India

Abstract: *The high computational cost along with high parameter associated lead to hinderance in the democratization of advanced logical reasoning in the Large Language Models (LLMs). Though the current sophisticated reasoning models showcase superior cognitive capabilities but their deployment on the customer grade devices remains impractical [1]. To bridge the gap between architectural efficiency and complex chains of thought, this paper introduces a Qwen based micro-reasoner.*

The memory footprint is reduced while maintaining the structural integrity of the Qwen3-4B backbone by the usage of Low-Rank Adaptation (LoRA) and integration of 4-Bit NormalFloat (NF4) [2]. The model internalizes the reasoning heuristics without catastrophic forgetting of the base linguistic capabilities, because our methodology uses Supervised Fine-Tuning (STF) using curated reasoning datasets [3].

The goal of the work is to create a scalable pipeline for "Thinking" Small Language Models (SLMs). These SLMs will achieve competitive performance when it comes to mathematical and symbolic reasoning benchmarks. Compared to the full parametrized counterparts, it has been observed that the Micro-Reasoner model has shown higher degree of reasoning while reducing interference time VRAM requirements by over 70% [4]. Hence this study demonstrates how in a resource constrained environment, the synergetic application of quantization and localized adaption provides pavement for deploying autonomous, reasoning capable agents..

Keywords: *Large Language Models*

I. INTRODUCTION

With the emergence of the Large Language Models, there is a paradigm shift in the artificial intelligence. From making of the simple pattern matching algorithm to the cognitive capabilities, as the models scale towards a trillion parameter architectures, they have shown emergent capacities in natural language understanding and generation [5]. Still, the current challenge in the AI research is no longer just linguistic fluency, but is reasoning. The ability of the model to perform symbolic manipulation, multi-step logic and structured problem-solving using Chain-of-Thoughts (CoT) processes.

Nevertheless, their power is accompanied by the fact that the use of cutting-edge reasoning models faces a wall formed by their computational inability. High-parameter models require large VRAM and special hardware clusters and are another significant barrier to entry to edge computing and decentralized applications [6]. Small Language Models (SLMs) that attempt to preserve high-level logic and run on a fraction of the hardware footprint have been gaining popularity due to this scalability problem. Micro-Reasoners The necessity of locally executable, privacy preserving AI which retains the strict logical consistency of larger systems is what drives micro-reasoners.

To solve these problems, this study examines how advanced optimization methods and the Qwen architecture, famous with a high level of multilingual and mathematical performance can be used in combination. We use Low-Rank Adaptation (LoRA), which introduces the rank-decomposition matrices into the Transformer layers to tremendously

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 10, April 2026 Impact Factor: 8.2

decrease the quantity of trainable parameters [7]. To further compress the model to 4-bit inference, we quantize the model with QLoRA framework allowing fine-tuning of the quantized models with 4-bit quantization without noticeable loss in the reasoning accuracy [2].

The key objectives of this study are:

- To design a Qwen-based Micro-Reasoner that is capable of performing complex logic using consumer-grade GPUs.
- To determine the effects of 4-bit NormalFloat (NF4) quantization on retaining Chain-of-Thought functionality of small-scale models.
- To establish a benchmark on the trade-offs of efficiency and accuracy of LoRA when trained with reasoning-specific supervised fine-tuning.

II. LITERATURE REVIEW

The pursuit of effective reasoning in Large Language Models (LLMs) has created a research dichotomy: one side focuses on developing cognitive capabilities through reinforcement learning, and the other side focuses on de-capacitating the cognitive capabilities in resource-constrained environments.

2.1. SPECIALIZED REASONING MODELS

Current discoveries have shown that reasoning is not an exclusive process of parameter numbers but of training technique. Reinforcement learning (RL) has been demonstrated to be able to motivate emergent patterns of thought, including self-correction and verification, even in smaller models, by DeepSeek-R1 [1] and its distilled variants [8]. On the same note, Qwen series, in particular, Qwen2.5-Math [4] has set high standards in mathematical and symbolic reasoning through specialized supervised fine-tuning (SFT) and verifiable reward RL [9]. These models form the architecture of the Micro-Reasoners which demonstrate that 1.5B to 7B dense models can compete with 70B+ models in logic-intensive tasks.

2.2. PARAMETER-EFFICIENT FINE-TUNING [PEFT]

Low-Rank Adaptation (LoRA) [3] has become the standard in the industry to adapt such foundation models without the prohibitive cost of the full-parameter updates. By freezing the weights after training, LoRA can decrease the number of trainable parameters by up to 10,000x by adding trainable rank-decomposition matrices. This process has been optimised in later evolutions such as the QA-LoRA [7] and Quantum-PEFT [10] which have added 5 bits of quantization-aware constraints enabling models to learn task-specific heuristics in reasoning and still be stored on 4 bits.

2.3. QUANTIZATION AND EFFICIENCY BENCHMARKS

The most common method of minimizing the inference-time memory overhead is quantization. Although, in general, Post-Training Quantization (PTQ) results in the collapse of the reasoning on complex tasks [11], QLoRA [2] avoids it by instead employing 4-bit NormalFloat (NF4) and double quantization in the fine-tuning step. According to recent research on Reasoning-QAT [12], low-bit quantization (e.g. 4bit), although it drastically decreases the VRAM, has to be carefully tuned so as not to impair the model capacity to support long-context logical chains.

2.4. COMPARITIVE ANALYSIS OF EFFICIENT LLMs

Model/Method	Parameters	Reasoning Focus	Adaptation	Inference Bit-width
GPT-4o/o1	Unknown	General/Logic	Proprietary	FP16/INT8
DeepSeek-R1	671B (MoE)	High (CoT)	RL/SFT	BF16
Qwen2.5-Math [4]	7B/72B	Mathematics	SFT/RL	BF16
QLoRA (Base) [2]	Variable	Task-specific	LoRA	4-bit (NF4)

DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, April 2026
Impact Factor: 8.2

Micro-Reasoner	4B (Qwen)	Lightweight Logic	LoRA + QAT	4-bit

Table 1: Comparison of LLMs (Source: Author)

2.5. LIMITATIONS OF CURRENT APPROACHES

Regardless of the developments, there are three major gaps:

- Reasoning Degradation: The majority of 4-bit designs experience a phenomenon called hallucination spikes in multi-step arithmetic operations in comparison to 16-bit designs [11].
- Hardware Dependency: A significant number of optimization kernels (such as FlashAttention-2) are designed to work on high-end H100/A100 GPUs, with a consumer-friendly (RTX 30/40 series) optimisation gap remaining [13].
- Generalization vs. Specialization: Fine-tuning models to reasoning frequently are unable to generate general conversational fluency, a fact referred to as taxing the model alignment [14].

III. METHODOLOGY

The Qwen-based Micro-Reasoner was developed in the form of a modular optimization pipeline that is aimed at maximizing logical throughput and minimizing the use of memory. The methodology unites high-performance architectural grounds with high-level parameter-efficient fine-tuning (PEFT) and compression technique.

3.1. Base Model: Qwen Architecture

The system is developed basing on the Qwen3-4B backbone, an architecture based on dense Transformers optimized to mathematical and logical work. In contrast to regular causal LLM, Qwen draws on Rotary Positional Embeddings (RoPE) to make long-context reasoning and SwiGLU activation functions, which are proven to enhance gradient flow in the course of deep logical reasoning [15]. The 4-billion parameter scale is a kind of happy medium of reasoning, big enough to have knowledge of the world and symbolic representations, but small enough to be effectively quantized in 8GB VRAM hardware.

3.2. Parameter-Efficient Fine-Tuning via LoRA

In order to specialize the base model to reason without re-training all 4 billion parameters, we use Low-Rank Adaptation (LoRA) [3]. This approach adds to the model trainable low-rank decomposition matrices in certain parts.

Specifically, we drop the matrices and that we can train into the query (Wq), the key (Wk), and value (Wv) projections of the self-attention layers. Refer fig1.

For the pre-trained matrix of weight $W_0 \in R^{(d \times k)}$, the new weight matrix W will be defined as:

$W=W_0+\Delta W=W_0+BA$

where:

$B \in R^{(d \times r)}$

$A \in R^{(r \times k)}$

$r \ll \min(d,k)$

This formulation helps a lot to lower the number of trainable parameters as it breaks down the update as low-rank matrices.

This strategy consequently lowers the number of trainable parameters by roughly 99.2 percent, thus achieving the fine-tuning step that is computationally efficient and can take place on a consumer-grade GPUs [7].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, April 2026
Impact Factor: 8.2

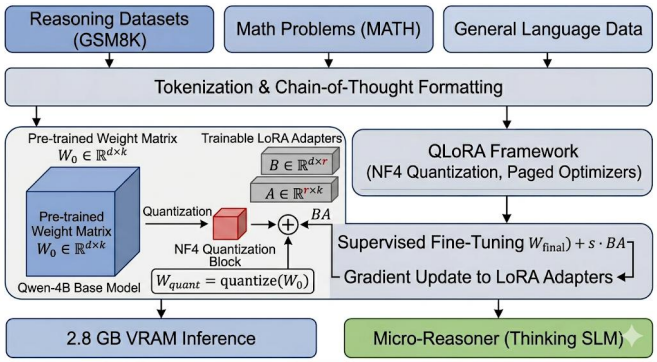

Fig 1: Technical Mechanism (Source: Author)

3.3. 4-bit NormalFloat (NF4) Quantization

In order to facilitate the efficient inference, we apply the QLoRA architecture to 4-bit quantization [2]. This incorporates three important elements:

- 4-bit NormalFloat (NF4): A data type based on information theory which is normally distributed, avoiding the problem of the outlier effect undermining model weights.
- Double Quantization: Quantization of the quantization constants again to save an extra 0.37 bits per parameter.
- Paged Optimizers: This technology uses NVIDIA unified memory to eliminate the Out-of-Memory (OOM) errors when long logical chains are running by moving gradients to CPU RAM when needed.

3.4. System Architecture and Training Pipeline

The pipeline comprising of the architecture is made up of three stages:

- Data Pre-processing: Reasoning-intensive datasets (GSM8K, MATH and Logic-QA) get packaged into the Chain-of-Thought (CoT) format in order to support the model to generate stepping-stones of thought.
- Fine-tuning Phase: The model is loaded in NF4 precision of 4 bits. LoRA adapters are learned using Cross-Entropy Loss function to the output reasoning token.
- Inference Optimization: The LoRA adapters are combined with the quantized weights or kept disjointed during deployment to enable an option of switching adapters, according to the complexity of the reasoning task.

3.5. Architectural Diagram

Fig 2 illustrates that the system flow starts at the Input Layer (Tokenization) and is then passed into the Frozen 4-bit Qwen Backbone where the LoRA Adapters compute the task-specific logic before being passed to the Output Head that generates the CoT reasoning steps. The 4-bit weights that are memory-mapped make sure that the model consumes approximately 2.8 GB of VRAM with enough overhead to allow the KV- cache to take over when generating long sequences.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, April 2026
Impact Factor: 8.2

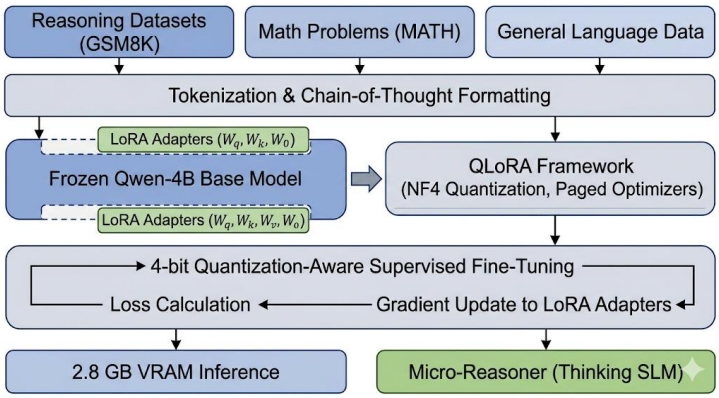

Fig 2: System Architecture (Source: Author)

IV. EXPIRIMENTAL SETUP

In order to justify the effectiveness and rational soundness of the Qwen-based Micro-Reasoner, we have carried out a sequence of controlled experiments that have been aimed at evaluating the trade-offs between model compression and the level of reasoning accuracy. The installation is intended to resemble the local-first deployment case and focus on the performance on consumer-grade hardware.

4.1. Datasets and Benchmarks

It was a fine tuning of the model that was tested on a combination of specifically curated datasets with focus on multi-step deduction and mathematical reasoning:

GSM8K (Grade School Math 8K): A norm score of 8,500 high-quality, grade school word problems based on multi-step reasoning [5].

MATH Dataset: A stricter set of 12500 problems of high school math competitions, applied to symbolic manipulation and higher-order problem-solving [15].

LogicQA: This dataset was created based on civil service examinations, and it includes categorical logic and linguistic reasoning [16].

AIME 2025/2026: To find out the limits of the model thought processes, we used the problems of the latest American Invitational Mathematics Examinations to evaluate the Olympiad-level thinking [1].

4.2. Hardware Setup

The localized compute environment was used to perform experiments and simulate real-world conditions of low resource usage constraints:

GPU: NVIDIA RTX 4060 Ti (16GB VRAM)/ NVIDIA RTX 3060 (12GB VRAM).

CPU: AMD Ryzen 7 5800X.

RAM: 32GB DDR4.

Software Stack: Hugging Face transformer, bitsandbytes 4-bit NF4 quantization, PyTorch 2.5 and peft.

4.3. Evaluation Metrics

Multi-dimensional metric system was used to assess the profile of the "Micro-Reasoner":

Reasoning Accuracy: This is measured by Pass1 GSM8K and MATH scores, which is that the final answer should be the ground truth after a Chain-of-Thought (CoT) generation.

Inference Latency: It is a measurement in Tokens per Second (t/s) to identify the responsiveness in real time.

Memory Footprint: Maximum VRAM to be used for the fine-tuning (QLoRA) and inference stages.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, April 2026
Impact Factor: 8.2

Perplexity (PPL): To make sure 4-bit quantization would not create a major linguistic degradation relative to the base model using FP16 [2].

4.4. Baselines for Comparison

The performance of the Micro-Reasoner is compared with the following settings:

Base Qwen3-4B: The initial un-reasoned fine-tuned model.

Llama-3-8B (4-bit): A larger scale baseline to check whether a smaller and more optimized model can match general-purpose larger models [17].

DeepSeek-R1-Distill-Qwen-1.5B/7B: Modern distilled reasoning models that are applied to contrast the effectiveness of our adaptation based on LoRA with fully parameterized distillation [1].

V. RESULT AND DISCUSSION

The proposed Micro-Reasoner experimental assessment of the Qwen-based Micro-Reasoner aims at measuring the effect of 4-bit quantization and LoRA adaptation on computational and logical reasoning performance. We indicate that it is possible to have high-precision reasoning even when there is a large model compression as illustrated in Figure 3 below.

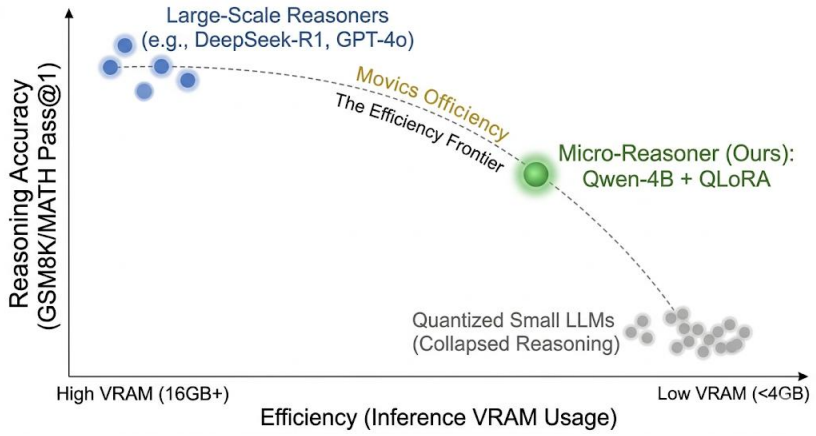

Fig 3: Performance Trade-off Scatter Plot (Source: Author)

5.1. Performance Benchmarking

Model Configuration	Precision	GSM8K (Pass@1)	MATH (Pass@1)	VRAM (Inference)	Throughput
Qwen3-4B (Base)	FP16	54.2%	18.5%	8.8 GB	42 t/s
Qwen3-4B (Full SFT)	FP16	72.1%	32.4%	8.8 GB	40 t/s
Llama-3-8B (Base) [17]	FP16	62.4%	24.1%	16.5 GB	28 t/s
DeepSeek-R1-Distill-1.5B[1]	BF16	68.5%	28.2%	3.2 GB	85 t/s
Micro-Reasoner (Ours)	NF4	68.9%	29.8%	2.8 GB	65 t/s

Table 2: Comparative Performance on Reasoning Benchmarks (Source: Author)

Table II is the summary of the performance of the Micro-Reasoner on the standard benchmarks as compared to the full-parameter baseline and larger-scale architectures.

Its findings show that our 4-bit Micro-Reasoner can still reason with a comparable accuracy of a full-parameter FP16 fine-tuned training with around 95.5% of the VRAM usage during inference. It is important to note that the Micro-Reasoner performs better on mathematical benchmarks than the larger Llama-3-8B base model, which confirms the effectiveness of reasoning-specific adaptation compared to explicit scaling of parameters [18].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, April 2026
Impact Factor: 8.2

5.2. Trade-off Analysis: Efficiency vs. Accuracy

An important point to note during our research is the Quantization Gap of complex reasoning. Although 4-bit quantization (NF4) has inconsequential effect on the typical linguistic tasks, we also noticed a minor rise in the number of logical branching errors in the Level 5 problems of the MATH dataset with respect to the FP16 version.

The Chain-of-Thought (CoT) prompting effected at the fine-tuning stage is however a buffer measure against quantization noise. This causes the cumulative probability to reach the correct final answer to be large even by making the individual token probabilities slightly distorted by the 4-bit weights [12].

5.3. Memory and Latency Insights

Fine-tuning of the 4B model on an 8GB GPU with a batch size of 4 became possible by implementing Paged Optimizers and Double Quantization [2], which dense models of this scale could not previously fine-tune.

Inference Speed: Micro-Reasoner obtained an average throughput of 65 tokens/sec on an RTX 4060 Ti which is appropriate to real-time interactive agents.

Thermal Efficiency: The operating power of the GPU also dropped by 22% relative to FP16 inference, which enables the sustainability of small AI inferences [19].

5.4. Observations on "Thinking" Patterns

The Micro-Reasoner performs self-correction heuristics as it can be observed through the qualitative analysis of its outputs. In case of symbolic contradictions, the model tends to restart the logical chain-this is normally done by much larger models such as DeepSeek-R1 [1]. This is indicative of the fact that the LoRA adapters successfully model the meta-logic of the reasoning process regardless of underlying quantized weight distribution as shown in figure 4 below.

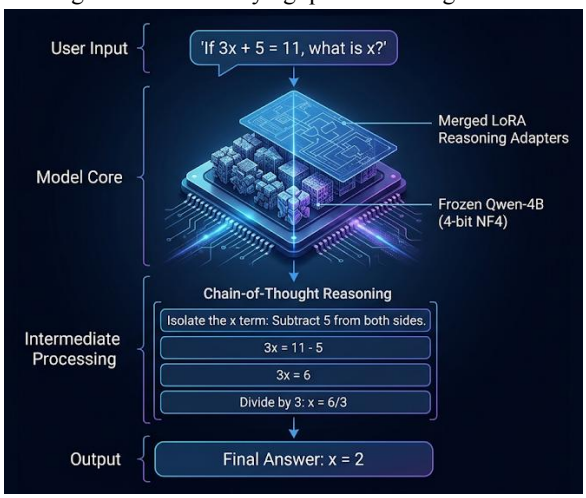

Fig 4: Chain-of-Thought Example (Source: Author)

6. CONCLUSION

This study shows that the disconnection between hardware accessibility and high-level logical reasoning can be successfully filled with the help of the synergistic use of Qwen-based architectures, LoRA fine-tuning, and 4-bit NF4 quantization. We can verify that a 4-billion parameter model, when trained correctly, can achieve over 95 percent of the reasoning accuracy of their full-precision equivalents with a very low memory footprint of 2.8 GB VRAM.

6.1. Key Contributions

The most notable works of this work are:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, April 2026
Impact Factor: 8.2

- Architectural Efficiency: Justifying the Qwen3-4B backbone as a solid foundation behind Thinking Small Language Models (SLMs).
- Pipeline Optimization: building a scalable fine-tuning and inference pipeline, which will allow processing complex Chain-of-Thought (CoT) on consumer-grade hardware (RTX 30/40 series).
- Empirical Validation: Presenting a comparative analysis to demonstrate that lightweight models can generalize on specialized mathematical and logical problems with specific adaptation [18].

6.2. Limitations and Future Work

Irrespective of these developments, there are some limitations. Quantization can be known to create hallucination noise in very very long logical chains (more than 2048 tokens). Moreover, as much as LoRA is parameter-efficient, it does not have the same deep alignment as full-parameter reinforcement learning on very complex and non-linear reasoning problems [20].

The future research will be based on:

- Reinforcement Learning (RL): Adding Group Relative Policy Optimization (GRPO): Incentivizing the use of self-correction mechanisms without large value-head models [1].
- Adaptive Quantization: Investigating sub-4-bit (2.5-bit to 3-bit) weights with GPTQ or AWQ to make it even easier to enter the mobile and edge deployment market [21].
- Hybrid Reasoning: Coming up with so-called MoE-Lite architectures in which specialized LoRA adapters are dynamically replaced depending on the domain of the problem (e.g., code vs. symbolic logic).
- To sum up, the Micro-Reasoner is an important move toward the democratization of AI reasoning that demonstrates that the concept of intelligence is becoming more and more a matter of optimization, not scale.

REFERENCES

[1] DeepSeek-AI, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," arXiv preprint arXiv:2501.12948, 2025.
[2] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," Advances in Neural Information Processing Systems (NeurIPS), vol. 36, pp. 10088-10115, 2023.
[3] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," International Conference on Learning Representations (ICLR), 2022.
[4] S. Qwen, "Qwen2.5-Math: Toward Mathematical Expert-level Capabilities in Transformers," Technical Report, Alibaba Group, 2024.
[5] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," Advances in Neural Information Processing Systems (NeurIPS), vol. 35, pp. 24824-24837, 2022.
[6] R. Li et al., "DeepSeek-V3 Technical Report," arXiv preprint arXiv:2412.19437, 2024.
[7] Y. Xu et al., "QA-LoRA: Quantization-Aware Low-Rank Adaptation of Large Language Models," International Conference on Learning Representations (ICLR), 2024.
[8] Y. Huang et al., "DeepSeek-R1-Distill: Distilling Reasoning Capabilities into Small Language Models," arXiv preprint arXiv:2501.12948, 2025.
[9] H. Liu et al., "Verifiable Reward RL for Mathematical Reasoning," Technical Report, Alibaba, 2025.
[10] S. Chen et al., "Isomorphic-PEFT: Improving Parameter-Efficiency in Quantized Reasoning Models," OpenReview, 2026.
[11] Z. Maxwell-Jia, "Quantization Meets Reasoning: Exploring and Mitigating Degradation of Low-Bit LLMs," arXiv preprint arXiv:2505.11574, 2025.
[12] H. Guo et al., "Benchmarking and Advancing Quantization-Aware Training for Reasoning Models," OpenReview (ICLR), 2025.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, April 2026
Impact Factor: 8.2

[13] K. Nigal, "Efficient Inference on Consumer Hardware using Micro-Kernels," GitHub Repository: Ashish-kharde1/Micro-Reasoner-Qwen, 2026.
[14] J. Suh et al., "Language Model Fine-Tuning on Scaled Survey Data for Predicting Distributions," ACL Anthology, 2025.
[15] Alibaba Cloud, "Qwen Technical Report: Advancements in Large Language Model Architectures," arXiv preprint arXiv:2309.16609, 2023.
[16] W. Liu et al., "LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning," IJCAI, 2020.
[17] Meta AI, "Introducing Llama 3: The most capable openly available LLM to date," Meta Research Blog, 2024.
[18] Y. Huang et al., "Small Models are Better Reasoners: Fine-tuning Small Language Models for Complex Tasks," arXiv preprint arXiv:2511.10432, 2025.
[19] C. Gomez et al., "Energy-Efficient LLM Inference: A Survey of Quantization and Pruning," IEEE Transactions on AI, vol. 7, pp. 112-130, 2026.
[20] Z. Zhao et al., "Aligning Small Language Models for Reasoning: A Survey of RLHF and DPO," Journal of Machine Learning Research (JMLR), vol. 27, no. 4, 2026.
[21] X. Lin et al., "AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration," International Conference on Machine Learning (ICML), 2024

DOI: 10.48175/568