Quantized LLM Reasoning: A Comprehensive Study of Post-Training Quantization in LLMs

Anonymous ACL submission

Abstract

Quantization plays a crucial role in enabling the efficient deployment of large language models (LLMs) on memory-constrained hardware, significantly reducing memory usage and computational costs. However, extreme low-bit quantization methods often impair essential capabilities such as complex reasoning, memory retention, and adherence to instructions. In this work, we systematically evaluate state-of-the-art quantization techniques on tasks involving chain-of-thought reasoning, instruction following, and multi-agent simulations. Additionally, we investigate partial and stochastic 1-bit quantization, i.e., binarization strategies that aim to preserve key reasoning capabilities, achieving a balance between model compression and performance retention. To evaluate the effectiveness of low-bit LLMs in advanced scenarios like multi-agent simulations, we curated two novel datasets for multi-agent doctor-patient simulation, USMLE and NHS, to overcome the challenge of data scarcity in the domain of medical simulation and reasoning. Our experiments on LLaMA, LLaMA-3.1, LLaMA-3.2 and LLaMA-3.3, as well as a reasoning-centric benchmark, demonstrate the potential of quantized models in maintaining functional integrity under extreme compression. The code for our work will be publicly available.

1 Introduction

Large Language Models (LLMs) have transformed the landscape of artificial intelligence, unlocking new possibilities in natural language understanding and complex reasoning (Huang and Chang, 2022). Their ability to process vast amounts of data and generate human-like responses has made them indispensable across various domains, from automated assistants to scientific discovery. However, their immense size comes at a cost—LLMs require substantial computational resources and memory capacity, making real-world deployment challenging, particularly on resource-constrained devices (Alizadeh et al., 2023).

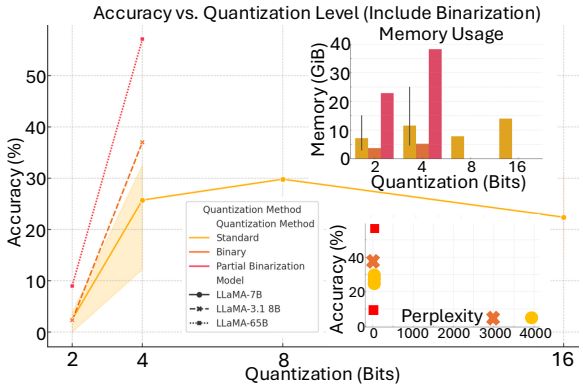

Figure 1: Impact of quantization and binarization on accuracy, memory, and perplexity. The main graph shows accuracy dropping below 4-bit, mitigated by binarization. The top-right bar graph highlights memory savings with minimal accuracy loss, while the bottom-right scatter plot shows severe degradation at 2-bit, with partial binarization preserving accuracy and low perplexity.

Quantization reduces LLM storage and computation, enabling deployment on lower-end hardware while maintaining performance (Lin et al., 2024). However, it degrades accuracy, particularly in reasoning, memory retention, and instruction-following (Jin et al., 2024), raising the question: **how can we improve efficiency without sacrificing core capabilities?** A key concern is preserving emergent abilities, such as in-context learning (ICL), chain-of-thought reasoning (CoT), instruction-following (IF), multi-agent collaboration, and agentic simulation. While prior work examines quantization's general effects (Huang et al., 2024), its impact on emergent behaviors under extreme compression remains underexplored, necessitating a **systematic evaluation of how quantized models retain or lose these abilities.** Recent advancements in extreme quantization, including ZeroQuant (Yao et al., 2022a), GPTQ (Frantar et al.,

2022a), and Partial Binarization (Yuan et al.), improve memory efficiency by reducing model size while preserving functionality. However, studies (Liu et al., 2023a; Wei et al., 2022a; Li et al., 2024; Dong et al., 2023) indicate that extreme low-bit quantization degrades memory retention, complex reasoning, and instruction-following, especially under aggressive compression. Addressing these limitations requires new strategies to mitigate performance loss.

In this work, We investigate the impact of quantization (See Figure 1) on the emergent abilities of LLMs, both in general reasoning tasks and domain-specific applications requiring complex reasoning and multi-agent interaction, such as medical simulations. While quantization enhances efficiency, its effect on instruction-following and reasoning capabilities remains underexplored, particularly in high-stakes applications. To address this gap, we curated two novel datasets (employing USMLE (Jin et al., 2020a) and NHS (National Health Service (NHS))) designed to systematically evaluate LLM performance in multi-agent medical reasoning, providing a benchmark for assessing robustness and generalization under extremely low-bit LLMs.

The main contributions of this work are summarized as follows:

1. We designed a multi-agent evaluation framework and curated a simulated medical dataset of over 13,000 cases to assess reasoning, memory retention, and instruction-following capabilities in quantized LLMs under extreme compression.

2. We analyze a hybrid approach – partial binarization – where only a subset of weights is binarized, while the remaining—critical for reasoning—are maintained at higher precision. This addresses the main limitation of uniform low-bit quantization in retaining complex reasoning skills. Additionally, we also implement a strategy for non-uniform partial binarization where the ratio of binarization varies across layers. We analyze this strategy to study the contribution of shallow and deeper layers in maintaining aformentioned abilities under partial binarazation.

3. We conducted extensive experiments on quantized LLMs, analyzing structured reasoning, performance trade-offs, and efficiency gains in extreme quantization settings.

2 Related Work

2.1 Quantization of Large Language Models

Quantization compresses large language models (LLMs) by reducing numerical precision, enabling deployment on resource-constrained hardware. Methods like ZeroQuant (Yao et al., 2022b), GPTQ (Frantar et al., 2022b), XTC (Wu et al., 2022), BitNet (Wang et al., 2023), and T-MAC (Wei et al., 2024) reduce model size and computational overhead, making LLMs viable on lower-end GPUs and CPUs. However, **aggressive quantization degrades reasoning accuracy, memory retention, and instruction-following**, limiting effectiveness in complex tasks.

Existing methods mitigate these trade-offs but remain insufficient for reasoning-intensive applications. GPTQ maintains general functionality at 3- or 4-bit precision but struggles with complex reasoning (Frantar et al., 2022b). BitNet (Wang et al., 2023) optimizes memory but lacks activation quantization, limiting efficiency. T-MAC improves matrix computation but sacrifices precision (Wei et al., 2024). ZeroQuant enhances post-training quantization but suffers in extreme settings like INT4, especially for generative tasks (Yao et al., 2022b). These limitations highlight the need for new quantization strategies that balance extreme compression with robust reasoning performance.

2.2 Emergent Abilities in Quantized LLMs

Scaling LLMs unlocks emergent abilities like in-context learning (ICL), chain-of-thought reasoning (CoT), instruction-following (IF), multi-agent coordination, and agentic simulation (Wei et al., 2022b), essential for complex reasoning and decision-making.

Quantization disrupts these abilities, affecting structured reasoning and multi-agent communication (Wei et al., 2022b), yet its impact remains underexplored. Existing research focuses on simple benchmarks, lacking systematic evaluation of high-level reasoning. To address this, we explore partial binarization to retain reasoning structures under extreme compression.

2.3 Binarization Techniques

Binarization improves efficiency but degrades accuracy, particularly in reasoning tasks. Partial bi-

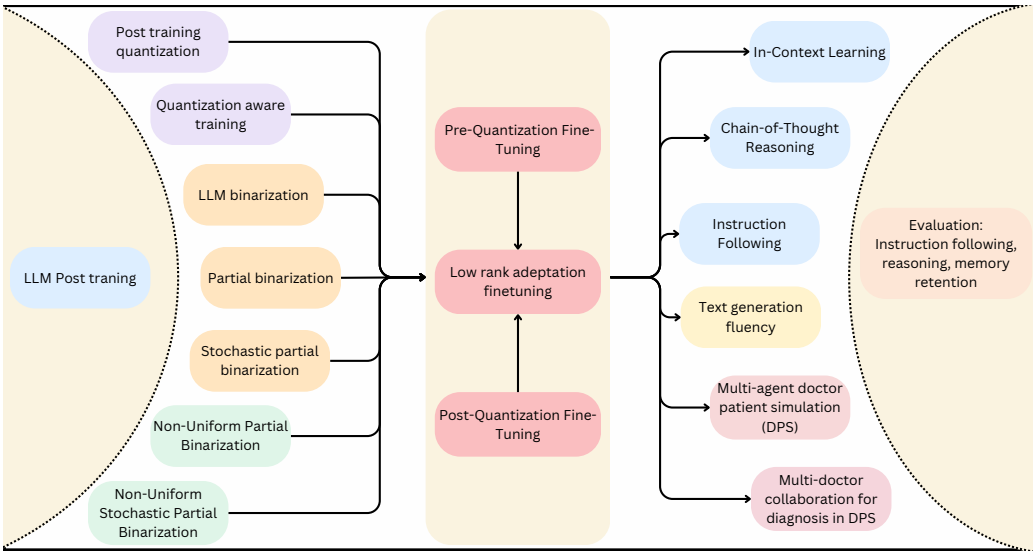

Figure 2: Our pipeline evaluates and optimizes LLMs under extreme low-bit quantization. It consists of post-training quantization, quantization-aware training, and binarization, extending to partial, stochastic, and non-uniform layer-based strategies. Fine-tuning is applied via pre- and post-quantization adaptation using LoRA for improved learning. The evaluation phase examines key emergent abilities, including in-context learning, chain-of-thought reasoning, instruction following, text generation fluency, and memory retention. A multi-agent doctor-patient simulation (DPS) assesses diagnostic reasoning, extending to multi-doctor collaboration for complex cases.

narization addresses this by retaining key weights in full precision while binarizing others. (Bamba et al., 2024) optimized this balance, improving reasoning accuracy, while (Shang et al., 2023) introduced PB-LLM to selectively preserve critical parameters. Stochastic quantization reduces bias but lacks adaptation for structured reasoning (Jin et al., 2024). Our framework integrates partial binarization with structured quantization to retain emergent abilities while maximizing efficiency.

3 Methodology

Our approach systematically evaluates quantization strategies for large language models (LLMs) while preserving their reasoning capabilities. Figure 2 outlines our pipeline, which consists of three main stages: quantization, fine-tuning, and evaluation. First, LLMs are quantized using various techniques to reduce memory consumption and improve efficiency. Next, fine-tuning strategies are applied to mitigate accuracy loss introduced by lower-bit representations. Finally, we assess the impact of quantization on emergent abilities using a multi-agent simulation framework.

3.1 Preliminary and Analysis

Quantization of Large Language Models. Quantization reduces numerical precision to lower memory and computational costs (Foundation, 2024; Frantar et al., 2022b), but aggressive quantization can degrade reasoning performance. To mitigate this, we employ post-training quantization (PTQ) and quantization-aware fine-tuning.

PTQ applies quantization to trained models using a small calibration dataset to minimize precision loss. GPTQ achieves state-of-the-art results through layer-wise reconstruction (Frantar et al., 2022a), preserving accuracy.

Quantization-aware fine-tuning restores lost capabilities by retraining after quantization. Pre-quantization fine-tuning improves performance in higher-bit models but is less effective for extreme quantization. Post-quantization fine-tuning optimizes quantized parameters, leveraging low-rank adaptation (LoRA) for efficient tuning on large models.

Preserving Reasoning Capabilities in Quantized Models. Preserving emergent abilities in quantized LLMs is key to maintaining reasoning. In-context learning enables zero-shot and few-shot adaptation but weakens under aggressive quantization, reducing long-range dependency retention. Chain-of-thought reasoning, crucial for tasks like math and medical diagnosis, is disrupted by lower-bit quantization, requiring fine-tuning. Instruction following, essential for structured tasks, suffers from token distortion, demanding targeted optimization.

Sensitivity of Model Components to Quantiza-

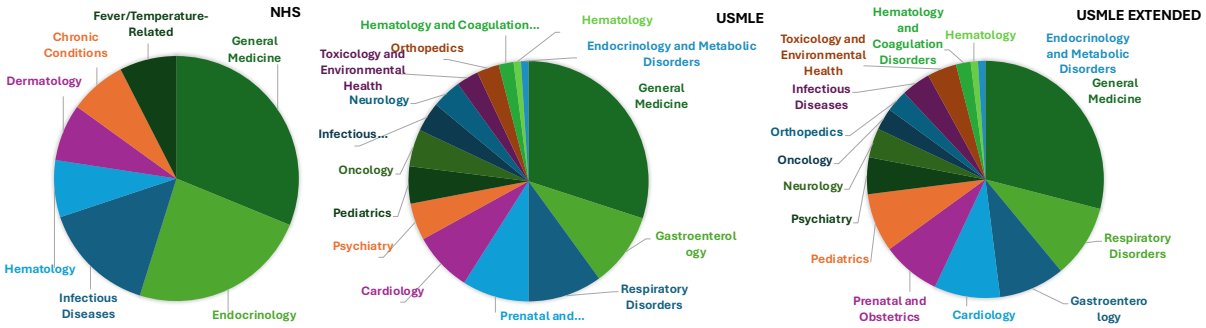

Figure 3: Percentage Distribution of Diagnostic Categories.

tion. Different components of LLMs exhibit varying sensitivity to quantization. Feedforward networks (FFN) are particularly prone to degradation under extreme quantization, especially at 2-bit precision, where reduced numerical representation affects inference stability. The presence of outlier activations—where specific feature dimensions hold disproportionately high values—also plays a crucial role. Standard quantization often fails to preserve these values, leading to information loss. To address this, we explore selective quantization strategies that retain high-precision representations for critical dimensions while binarizing less essential components.

3.2 Partial Binarization: A Hybrid Precision Approach

Partial binarization (PB-LLM) offers a compromise between full quantization and precision retention by selectively applying binarization to specific model components. Instead of binarizing all weights, PB-LLM retains high-precision parameters critical for structured reasoning. This approach follows a hybrid representation:

$$w \approx \alpha \cdot w_b \tag{1}$$

where $w_b \in \{-1, +1\}$ are the binarized weights, and α is a scaling factor computed as:

$$\alpha = \frac{\|w\|_1}{n} \tag{2}$$

where $\|w\|_1$ represents the $L1$-norm of the full-precision weights and n is the number of weights. Activations are binarized using a sign function:

$$a_b = \text{Sign}(a) = \begin{cases} +1 & \text{if } a \geq 0 \\ -1 & \text{otherwise} \end{cases} \tag{3}$$

Gradient computation is handled via the Straight-Through Estimator (STE), ensuring gradients flow through binarized layers without significant loss of training signal.

3.3 Stochastic Binarization for Robustness

Stochastic binarization introduces controlled randomness into the quantization process, improving robustness by preventing systematic errors from fixed threshold binarization. Each weight is binarized probabilistically:

$$w_b = \begin{cases} +1 & \text{with probability } \sigma(w) \\ -1 & \text{otherwise} \end{cases} \tag{4}$$

where $\sigma(w)$ is a clipped hard sigmoid function:

$$\sigma(w) = \max(0, \min(1, \frac{w+1}{2})) \tag{5}$$

The same stochastic approach is applied to activations, reducing information loss while improving the stability of extreme quantization settings.

3.4 Multi-Agent Simulation for Evaluating Quantized Models

We developed a multi-agent simulation framework inspired by AgentClinic (Schmidgall et al., 2024) to assess the reasoning capabilities of quantized models in clinical decision-making. The system includes four agents: doctor, patient, measurement, and moderator. The doctor agent gathers patient history, requests diagnostic tests, and formulates diagnoses. The patient agent provides symptom descriptions and medical history, while the measurement agent supplies multimodal test results. The moderator agent evaluates diagnostic accuracy by comparing predictions against ground truth.

To enhance diagnostic reliability, we implemented a multi-agent debate framework where multiple doctor agents engage in iterative discussions. A majority voting mechanism determines the final diagnosis, ensuring robust clinical reasoning despite quantization constraints.

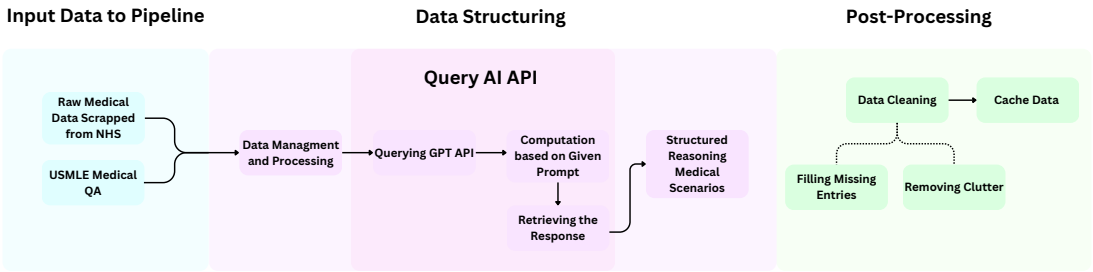

Figure 4: Overview of our medical scenario processing pipeline. Raw medical data from NHS and USMLE Medical QA is processed and sent to query AI API. The model analyzes the input, generates structured reasoning scenarios based on the given prompts, and returns a structured response. Post-processing steps ensure data cleanliness and caching data.

3.5 Dataset Pipeline for Medical Reasoning

We developed an automated pipeline (Figure 4) that transforms raw medical data into structured, instruction-driven reasoning datasets for evaluating quantized models in medical diagnosis.

Two datasets were curated from primary sources. The first dataset was scraped from NHS UK, covering 923 distinct diseases. The second dataset consists of medical case studies from the United States Medical Licensing Examination (USMLE), with a standard version containing 1,273 cases and an extended version comprising 10,178 cases. These datasets provide clinically relevant benchmarks for evaluating LLMs in medical reasoning tasks. Figure 3 presents the distribution of diagnoses across the Extended USMLE, USMLE, and NHS datasets, categorized into key medical fields.

The data structuring module organizes raw medical information into a structured format for model training. Using OpenAI, it extracts key components, including task objectives, patient demographics, symptoms, physical examination findings, diagnostic test results, and expected diagnoses. The post-processing module refines the dataset by filtering irrelevant information, filling missing data with a generative AI model, and conducting manual reviews to ensure accuracy. The final dataset is stored for analysis and fine-tuning.

4 Experiments

4.1 Experimental Setup

We begin with the unquantized LLaMa 3 (Touvron et al., 2023), LlaMA 3.1 (Grattafiori et al., 2024), LLaMA LLaMA 3.2[1] and LLaMA 3.3[1] family models and progressively apply each quantization method. We ensure a fair comparison by keeping all hyperparameters, training settings, and evaluation protocols consistent across experiments. In the training process of our quantized network, we commence with a pretrained model for initialization. The optimization of the model is facilitated through the AdamW optimizer (Loshchilov and Hutter, 2019), applied with zero weight decay. We assign a batch size of 1 to each GPU and implement a learning rate of 2e-5, adhering to a cosine learning rate decay strategy. All experiments in this study were conducted on 4 NVIDIA A100 40GB GPUs and 1 NVIDIA A100 80GB GPUs, ensuring a consistent and high-performance computing environment for evaluating quantized LLMs.

4.2 Datasets and Benchmarks

To evaluate the impact of quantization on logical inference and decision-making, we utilize a diverse set of reasoning benchmarks: MMLU (Hendrycks et al., 2021) to assess performance across multiple domains, including STEM, humanities, and social sciences. BBH (Suzgun et al., 2022) to evaluate complex reasoning abilities that require deep logical deduction. GSM8K (Cobbe et al., 2021) to measure CoT reasoning (Kim et al., 2023) in mathematical problem-solving. AlpacaFarm (Dubois et al., 2024) tests instruction-following capabilities through preference-based evaluations. WikiText (Merity et al., 2016) evaluates language modeling performance using perplexity metrics. For medical reasoning multi-agent simulation, we used MedQA (Jin et al., 2020b) and our curated USMLE and NHS datasets.

4.3 Quantization and Binarization Strategies

We apply GPTQ (Frantar et al., 2022b), Bitsandbytes (Foundation, 2024), ZeroQuant (Yao et al.,

[1] https://github.com/meta-llama/llama-models/blob/main/models/

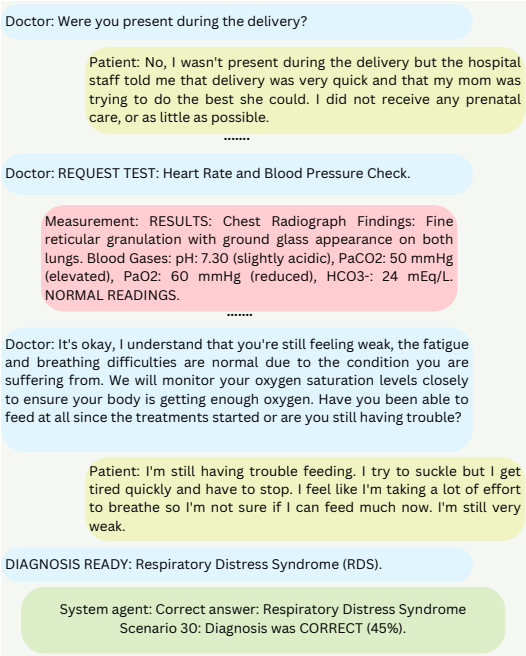

Figure 5: Multi-Agent Interaction in Medical Diagnosis Simulation: This figure illustrates the Doctor Agent collecting symptoms and history from the Patient Agent while requesting test results from the Measurement Agent. Using structured reasoning, the Doctor Agent formulates a diagnosis. The system employs LLaMA 3.3 70B 4-bit quantized for realistic doctor-patient interactions and diagnostic assessment.

2022b), and PB-LLM (Yuan et al., 2024) for efficient quantization. For binarization, we explore Binary-0.5, non-uniform binarization, and stochastic partial binarization (Bamba et al., 2023) to balance compression with reasoning performance.

Pre-Quantization Fine-Tuning Before applying quantization, fine-tuning is leveraged to optimize model performance, ensuring robustness in ICL, CoT reasoning, and IF. Following best practices from prior studies, LLaMA models are fine-tuned on the Alpaca dataset (for instruction tuning) and CoT-annotated datasets (for logical reasoning). Additionally, LoRA-based parameter-efficient fine-tuning (Hu et al., 2021) is explored to maintain adaptability without significantly increasing computational costs.

Post-Quantization Fine-Tuning To counteract performance degradation caused by quantization, a specialized fine-tuning framework is introduced for post-quantized LLaMA models. This allows direct optimization of 2-bit larger models on a single A100 GPU (Liu et al., 2023b), achieving better performance than a 16-bit LLaMA-c13B in MMLU (5-shot evaluation). Inspired by QAT (Ashkboos et al., 2024) and parameter-efficient tuning methods (Hu et al., 2021), the approach modifies LoRA to incorporate GPTQ-generated quantized weights, drastically reducing memory overhead. The bigger models like LLaMa 3.3 70B model at 2-bit precision requires only 23.2 GiB, making it a highly efficient fine-tuning strategy for extreme low-bit LLM quantization.

4.4 Performance on Reasoning Tasks

To assess the impact of quantization on reasoning abilities, we evaluate LLaMA models across multiple benchmarks, as detailed in Tables 1, 2, 3. Our analysis highlights how extreme quantization affects logical inference, structured reasoning, and memory retention.

Overall Performance Trends: Table 1 presents the zero-shot and few-shot accuracy of LLaMA-7B and LLaMA 3.1 8B at different quantization levels. While 16-bit and 8-bit models maintain strong reasoning performance, accuracy drops significantly below 4-bit, especially in MMLU and BBH benchmarks. Notably, partial binarization (PB-LLM) preserves performance better than uniform 4-bit or 2-bit quantization, demonstrating its ability to balance compression and accuracy.

Chain-of-Thought and Commonsense Reasoning: Table 3 shows the performance across different reasoning categories. For arithmetic reasoning (GSM8K), accuracy remains stable under 4-bit quantization but collapses at 2-bit. Commonsense reasoning (HellaSwag) follows a similar trend, with PB-LLM outperforming standard 4-bit models due to selective precision retention. Multi-hop reasoning, which requires combining multiple facts, suffers the most under aggressive quantization, reinforcing the need for fine-tuning.

Instruction-Tuned Models and Prompt-Level Accuracy: Table 2 further examines performance on an instruction-following dataset. Instruction-tuned models perform significantly better than non-tuned models, with 50% binarization retaining more accuracy than full quantization methods like GPTQ-INT4. This suggests that partial binarization mitigates degradation in structured reasoning.

Overall, our findings emphasize that while extreme quantization reduces memory and computational costs, it comes at the expense of logical coherence. However, PB-LLM effectively preserves reasoning abilities, making it a promising approach for deploying LLMs in resource-constrained environments.

Table 1: Zero-shot and few-shot performance comparison of LLaMA-7B and LLaMA 3.1 8B across different precisions, including 10% partial binarization (4-bit setting) on various reasoning benchmarks.

Models	Precision	MMLU		BBH		GSM8K	WikiText	Mem.
		0-shot	5-shot	0-shot	3-shot		(PPL)	(GiB)
LLaMA-7B	16-bit	30.9	36.8	18.4	32.1	13.9	5.7	14.0
	8-bit	29.8	35.5	17.6	32.7	14.7	5.7	7.8
	4-bit	32.5	35.8	19.9	32.2	13.8	5.8	4.8
	2-bit	3.7	5.5	1.8	4.0	0.0	3939.1	3.1
LLaMA 3.1 8B	4-bit	37.0	38.2	20.8	37.8	15.2	5.0	5.2
	2-bit	2.3	3.8	0.4	2.7	1.5	2968.0	3.7
LLaMA 3.1 8B (PB 10%)	4-bit	29.2	35.2	17.3	31.0	13.1	5.7	13.9

Table 2: Zero-shot performance on `ifeval` (Zhou et al., 2023) dataset using LLaMA 3.1 8B with various binarization ratios. The metric compares instance-level and prompt-level accuracy in both loose and strict criteria.

Precision	Instance-Level Acc. (Loose)	Instance-Level Acc. (Strict)	Prompt-Level Acc. (Loose)	Prompt-Level Acc. (Strict)
Non-Instruction-Tuned Models				
Binarization-30%	21.43	21.30	19.24±1.51	18.50±1.10
Binarization-20%	21.10	19.90	10.17±1.30	8.69±1.21
Binarization-10%	18.47	17.75	10.54±1.32	9.80±1.28
Instruction-Tuned Models				
GPTQ-Binarization-50%	59.83	56.71	47.13±2.15	44.36±2.10
bnb-4-bit	61.03	57.19	46.77±2.15	42.14±2.12
GPTQ-INT4	57.43	54.20	44.77±2.14	40.48±2.11
GPTQ-16-bit	60.55	57.19	46.77±2.15	42.51±2.13

Table 3: Reasoning Breakdown Across Task Types for LLaMA-7B under different quantization settings. Arithmetic is evaluated on GSM8K, commonsense on HellaSwag, and multi-hop on an internal QA dataset.

Model	Precision	Arithmetic (GSM8K)	Commonsense (HellaSwag)	Multi-Hop QA
LLaMA-7B	16-bit	13.9	69.2	48.5
LLaMA-7B	4-bit	13.8	66.0	46.2
LLaMA-7B	PB (4-bit, 10%)	13.1	67.8	47.1
LLaMA-7B	2-bit	0.0	10.5	2.2

4.5 Multi-Agent Simulation

We evaluate quantized models in a multi-agent clinical simulation where a Doctor Agent interacts with a Patient Agent and requests diagnostic tests from a Measurement Agent. This setup assesses reasoning retention under quantization.

Table 4 presents the accuracy comparison of various models across simulated and non-simulated medical environments. Figure 5 shows a sample simulation of the Multi-Agent simulation The non-simulated environment, representing real-world medical decision-making scenarios, shows the highest accuracy of 73.2%, achieved by the LLaMA 3.1 (70B) model. In contrast, within the simulated environment, performance varies significantly across models. Among full-precision models, GPT-4 Turbo achieves the best accuracy at 53.4%, followed by Mixtral 7B*8B and GPT-4 Vision Preview, scoring 37.6% and 35.7%, respectively. The smaller LLaMA models (3B, 3.2B, and 2 (70B)) demonstrate notably lower performance, with accuracy ranging from 4.3% to 8.5%.

Furthermore, when applying quantization techniques, such as Binary 4-bit and GPTQ 4-bit, accuracy slightly improves compared to some smaller full-precision models. Specifically, the LLaMA 3.1 (70B) model in Binary 4-bit format reaches 20.8%, whereas GPTQ 4-bit quantization yields a higher accuracy of 26.3%. These results suggest that precision reduction impacts accuracy but can still maintain competitive performance depending on the model and task requirements.

Table 5 shows that lower-bit models struggle with diagnostic accuracy due to information loss, particularly at 2-bit precision. Partial binarization mitigates this by preserving critical reasoning pathways, with PB-LLM at 4-bit performing comparably to 8-bit models.

Table 4: Performance of Different Models in Simulated (agentic) and Non-Simulated (direct QnA) Medical Environments. "Accuracy (%)" is based on correct diagnostic or management decisions.

Precision	Model	Accuracy (%)
Non-Simulated Environment		
Full	LLaMA 3.1 (70B)	**73.0**
Simulated Environment		
Full	GPT-4 Turbo	53.4
	GPT-4 Vision Preview	35.6
	Mixtral 7B×8B	37.6
	LLaMA 3.2 (3B)	5.8
	LLaMA 3.0 (3B)	4.4
	LLaMA 2 (70B)	8.5
Binary 4-bit	LLaMA 3.1 70B	20.8
GPTQ 4-bit	LLaMA 3.1 70B	26.2

Table 5: Multi-Agent Simulation Performance on Medical Reasoning Datasets for LLaMA 3.3 70B.

Model	Quantization	MedQA	NHS	USMLE	USMLE ext.
LLaMA 3.3 70B	16-bit (FP16)	58.5	60.2	56.9	55.4
LLaMA 3.3 70B	8-bit	56.3	58.5	54.8	53.9
LLaMA 3.3 70B	4-bit (GPTQ)	53.9	55.2	51.5	59.8
LLaMA 3.3 70B	4-bit (PB-LLM)	55.8	57.4	53.2	52.1
LLaMA 3.3 70B	2-bit	6.2	9.5	9.1	7.8

4.6 Ablation Studies

Multi-agent Collaboration for Final Disease Diagnosis: To evaluate the effectiveness of multi-agent collaboration in the end of the question answering, we analyze diagnostic accuracy under different levels of agent interaction. Table 6 summarizes the performance improvements from single-agent diagnosis to collaborative decision-making with structured debate and voting.

Table 6: This experiment evaluates the effectiveness of multi-agent collaboration in enhancing diagnostic accuracy when using LLaMA 3.3 70B with 4-bit PB-LLM binarization. The results demonstrate that despite aggressive quantization, PB-LLM preserves logical reasoning abilities, enabling reliable disease diagnosis through structured agent interactions.

Collaboration Level	MedQA(%)	NHS(%)	USMLE (%)
Single Doctor Agent	55.8	57.5	52.1
Two Doctors (Majority Vote)	56.1	57.8	53.4
Multi-Doctor (n=5)	58.4	62.7	58.2
Multi-Doctor (n=5) + Feedback	59.0	65.9	61.1

Impact of varying binarized layers in LlaMA Table 7 provides insights into the effect of varying binarization levels on model perplexity. It demonstrates that retaining full-precision weights in deep layers is crucial for preserving structured reasoning. A balanced binarization ratio (e.g., 50–60%) ensures stability, while excessive binarization (e.g., 80%) significantly degrades performance. The results highlight the effectiveness of non-uniform binarization strategies, where deeper layers maintain more precision to support long-range dependencies, improving efficiency without compromising coherence.

Table 8 extends this analysis to stochastic partial binarization. Unlike deterministic binarization, stochastic methods introduce variability but do not yield significant improvements in perplexity. While lower binarization levels maintain model stability, increasing binarization beyond 60% drastically impacts performance. These findings reinforce the necessity of adaptive quantization strategies, ensuring optimal trade-offs between efficiency and reasoning capabilities.

Table 7: Effective PB Quantization Results: Impact of varying binarized layers on perplexity and improvement. "# Deep Layers" refers to the number of last layers excluded from binarization.

Effective % Binar.	# Deep Layers	% Binarization		Perplexity	%↑
		Shallow Layers	Deep Layers		
50%	–	–	–	18.71	–
50%	4	48%	55%	18.57	0.75%
50%	2	48%	60%	18.59	0.65%
60%	–	–	–	24.12	–
60%	4	56.60%	70%	22.19	8%
60%	2	57%	80%	22.76	5.60%
70%	–	–	–	45.90	–
70%	4	63%	90%	33.47	27.10%
70%	2	69%	80%	45.62	0.60%
80%	–	–	–	84.33	–
80%	4	76.60%	90%	82.24	2.40%
80%	2	78.50%	90%	79.15	6.14%

Table 8: Uniform and Non-Uniform Stochastic Partial Binarization of LLaMA-3.2-1B model in a Post-Training Quantization Framework. The columns with '–' mean that each layer $l \in \{1, 2...L\}$ was stochastically partially binarized with the binarization percentage shown in the first column. Other columns show when shallow layers (L-4 or L-2) and deep layers (4 or 2) are binarized at different rates.

Effective Binarization (%)	Deep Layers	Shallow Layers (%)	Deep Layers (%)	Perplexity
50%	–	–	–	43.72
	4	48	55	40.61
	2	48	60	42.37
60%	–	–	–	150.47
	4	56.60	70	156.75
	2	57	80	174.89
70%	–	–	–	1114.24
	4	63	90	1529.14
	2	69	80	1118.32
80%	–	–	–	45849.47
	4	76.60	90	123991.02
	2	78.50	90	40236.45

5 Conclusion

We have conducted a comprehensive study on the reasoning capabilities of post-training quantized large language models. Our results indicate that while extreme quantization can degrade performance on complex reasoning tasks, a hybrid precision strategy like partial binarization can maintain a surprising amount of these emergent abilities. Future work will explore more dynamic quantization strategies and extend the reasoning-centric dataset to other complex modalities and multi-step reasoning domains.

6 Limitations

Quantization at extremely low bit-widths (e.g., 2-bit) results in severe performance degradation due to the loss of representational capacity, affecting both logical consistency and reasoning accuracy. While PB-LLM mitigates this to some extent by improving weight binarization, it does not fully restore model precision, particularly in complex inference tasks. Multi-agent simulations face limitations in adaptive reasoning due to the rigid nature of scripted interactions, failing to capture real-world uncertainty and variability in doctor-patient dialogues. Generalization remains a critical issue for quantized models, as reduced precision leads to brittleness across diverse reasoning benchmarks, making them unreliable in complex, unseen scenarios. Additionally, existing binarization techniques optimize for efficiency at the cost of accuracy, lacking a universal approach to balance computational constraints with robust performance.

References

Keivan Alizadeh, Iman Mirzadeh, Dmitry Belenko, Karen Khatamifard, Minsik Cho, Carlo C Del Mundo, Mohammad Rastegari, and Mehrdad Farajtabar. 2023. Llm in a flash: Efficient large language model inference with limited memory. *arXiv preprint arXiv:2312.11514*.

Saleh Ashkboos, Bram Verhoef, Torsten Hoefler, Evangelos Eleftheriou, and Martino Dazzi. 2024. Efqat: An efficient framework for quantization-aware training. *Preprint*, arXiv:2411.11038.

Udbhav Bamba, Neeraj Anand, Saksham Aggarwal, Dilip K. Prasad, and Deepak K. Gupta. 2023. Partial binarization of neural networks for budget-aware efficient learning. *Preprint*, arXiv:2211.06739.

Udbhav Bamba, Neeraj Anand, Saksham Aggarwal, Dilip K Prasad, and Deepak K Gupta. 2024. Partial binarization of neural networks for budget-aware efficient learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2336–2345.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023. How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback. *Preprint*, arXiv:2305.14387.

BitsAndBytes Foundation. 2024. Bits and bytes library. https://github.com/bitsandbytes-foundation/bitsandbytes.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022a. Gptq: Accurate post-training compression for generative pretrained transformers. *arXiv preprint arXiv:2210.17323*, 1.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022b. Gptq: Accurate post-training compression for generative pretrained transformers. *arXiv preprint arXiv:2210.17323*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and llama3 team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

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

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Wei Huang, Xudong Ma, Haotong Qin, Xingyu Zheng, Chengtao Lv, Hong Chen, Jie Luo, Xiaojuan Qi, Xianglong Liu, and Michele Magno. 2024. How good are low-bit quantized llama3 models? an empirical study. *arXiv preprint arXiv:2404.14047*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020a. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020b. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

Renren Jin, Jiangcun Du, Wuwei Huang, Wei Liu, Jian Luan, Bin Wang, and Deyi Xiong. 2024. A comprehensive evaluation of quantization strategies for large language models. *arXiv preprint arXiv:2402.16775*.

Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. *Preprint*, arXiv:2305.14045.

Shiyao Li, Xuefei Ning, Luning Wang, Tengxuan Liu, Xiangsheng Shi, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. 2024. Evaluating quantized large language models. *arXiv preprint arXiv:2402.18158*.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of Machine Learning and Systems*, 6:87–100.

Peiyu Liu, Zikang Liu, Ze-Feng Gao, Dawei Gao, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2023a. Do emergent abilities exist in quantized large language models: An empirical study. *arXiv preprint arXiv:2307.08072*.

Peiyu Liu, Zikang Liu, Ze-Feng Gao, Dawei Gao, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2023b. Do emergent abilities exist in quantized large language models: An empirical study. *Preprint*, arXiv:2307.08072.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

National Health Service (NHS). Conditions and treatments.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

Yuzhang Shang, Zhihang Yuan, Qiang Wu, and Zhen Dong. 2023. Pb-llm: Partially binarized large language models. *arXiv preprint arXiv:2310.00034*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *Preprint*, arXiv:2210.09261.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Huaijie Wang, Lingxiao Ma, Fan Yang, Ruiping Wang, Yi Wu, and Furu Wei. 2023. Bitnet: Scaling 1-bit transformers for large language models. *arXiv preprint arXiv:2310.11453*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei et al. 2022b. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jianyu Wei et al. 2024. T-mac: Cpu renaissance via table lookup for low-bit llm deployment on edge.

Xiaoxia Wu et al. 2022. Xtc: Extreme compression for pre-trained transformers made simple and efficient. *Advances in Neural Information Processing Systems*, 35:3217–3231.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. 2022a. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. *Advances in Neural Information Processing Systems*, 35:27168–27183.

Zhewei Yao et al. 2022b. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. *Advances in Neural Information Processing Systems*, 35:27168–27183.

Zhihang Yuan, Yuzhang Shang, and Zhen Dong. Pb-llm: Partially binarized large language models. In *The Twelfth International Conference on Learning Representations*.

Zhihang Yuan, Yuzhang Shang, and Zhen Dong. 2024. Pb-llm: Partially binarized large language models. In *The Twelfth International Conference on Learning Representations*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

Appendix

A Summary of Experiments

Table A9 presents an overview of the experiments conducted, including quantization techniques, reasoning tasks, multi-agent medical simulations, and ablation studies.

Table A9: Summary of Experiments Conducted and Models Used.

Experiment	Objective	Models Used
Quantization Evaluation	Comparison of GPTQ, bitsandbytes, ZeroQuant, PB-LLM.	LLaMA-7B (16, 8, 4, 2-bit), LLaMA-3.1 8B, LLaMA-3.2 (3B), LLaMA-2 (70B)
Reasoning Tasks	Evaluation of CoT reasoning, ICL, Multi-agent Simulation.	LLaMA-7B, LLaMA-3.1 8B, GPT-4 Turbo, Mixtral 7B, GPT-4 Vision
Benchmarks Used	Datasets: GSM8K, MMLU, BBH, WikiText.	LLaMA-7B, LLaMA-3.1 8B, GPT-4 Turbo, Mixtral 7B
Multi-Agent Simulation	Doctor-patient interactions, USMLE/NHS datasets.	LLaMA-3.1 8B, LLaMA-3.2 (3B), GPT-4 Turbo
Ablation Studies	Binarization ratios, quantization backends, task complexity.	LLaMA-7B, LLaMA-3.1 8B, LLaMA-3.2 (3B), LLaMA-2 (70B)

B Performance on Reasoning Benchmarks

Table A10 provides accuracy results for reasoning tasks: BoolQ, PIQA, HellaSwag, Winogrande, ARC-Easy, ARC-Challenge, and OpenBookQA.

Task	Accuracy (%) ± Std.
BoolQ	37.83 ± 0.85
PIQA	49.51 ± 1.17
HellaSwag	25.04 ± 0.43
Winogrande	49.57 ± 1.41
ARC-Easy	25.08 ± 0.89
ARC-Challenge	22.70 ± 1.22
OpenBookQA	27.60 ± 2.00
Mean	**33.90**

Table A10: Accuracy (%) and standard deviation for reasoning benchmarks using LLaMA.