CMedBench: A Comprehensive Benchmark for Efficient Medical Large Language Models

Shengbo Gao[1], Jinyang Guo[1,3*], Lixian Su[4], Yifu Ding[1,2], Shiqiao Gu[5], Aishan Liu[1,2], Yuqing Ma[1,3], Zhiwang Zhang[6], Xianglong Liu[1,2]

[1]State Key Laboratory of Complex & Critical Software Environment, Beihang University, Beijing, China
[2]School of Computer Science and Engineering, Beihang University, Beijing, China
[3]School of Artifcial Intelligence, Beihang University, Beijing, China
[4]Western University, Ontario, Canada
[5]SenseTime Research, Beijing, China
[6]NingboTech University, Ningbo, China
{gtabris, jinyangguo, yifuding, liuaishan, mayuqing, xlliu}@buaa.edu.cn, 16ls19@queensu.ca, gushiqiao@sensetime, zhiwang.zhang@nbt.edu.cn

Abstract

Large Language Models (LLMs) hold significant potential for enhancing healthcare applications, yet their deployment is hindered by high computational and memory demands. Model compression techniques offer solutions to reduce these demands, but their impact on medical LLMs remains underexplored. In this paper, we introduce CMedBench, the first comprehensive benchmark for evaluating compressed LLMs in medical contexts. CMedBench assesses five core dimensions: Medical Knowledge Ability, Medical Application Ability, Trustworthiness Maintenance, Compression Cross Combination, and Computational Efficiency. Through extensive empirical studies, we analyze the trade-offs between model efficiency and clinical performance across diverse models, datasets, and compression strategies. Our findings highlight critical limitations in current evaluation practices and provide a robust framework for aligning compression strategies with medical requirements. CMedBench serves as a vital resource for researchers and practitioners, guiding the development of efficient, trustworthy, and clinically effective LLMs for healthcare applications.

Introduction

Large Language Models (LLMs) have rapidly advanced the state of artificial intelligence(Achiam et al. 2023), showing exceptional capabilities across various domains. In healthcare, they hold particular promise for enhancing diagnostic accuracy and supporting clinical decision-making(Singhal et al. 2025). However, deploying LLMs in healthcare is hindered by their high computational and memory demands, which often exceed the capabilities of resource-constrained and privacy-sensitive clinical settings(Guo et al. 2024; Liu et al. 2024a).

Model compression techniques notably training free algorithms such as quantization and sparsification, have emerged as critical solutions to mitigate these medical-context challenges by reducing the computational footprint of LLMs(Guo, Ouyang, and Xu 2020; He et al. 2025; Lv et al. 2024; J. Guo, W. Ouyang, and D. Xu 2020; Wang et al. 2025b; Guo et al. 2020; Rang et al. 2025; Liu et al. 2024b). However, the impact of compression on LLMs' medical knowledge, application, trustworthiness, and efficiency remains underexplored across diverse healthcare scenarios(Wang et al. 2024). The primary challenges of evaluating compressed LLMs in real clinical medical contexts can be divided into two folds:

***Challenge-1*: Fragmented Medical Context Evaluation:** Current evaluation for LLMs are fragmented, focusing on isolated medical tasks without comprehensively addressing the diverse demands of medical applications (Jin et al. 2019), such as advanced reasoning and interdisciplinary communication. These demands encompass a wide range of capabilities, including foundational medical knowledge recall, advanced diagnostic reasoning for complex or rare conditions, etc. Additionally, they inadequately assess trustworthiness attributes like truthfulness, privacy, and robustness, which are also critical for safe medical deployment(Wang et al. 2021; Lin, Hilton, and Evans 2022). This dual gap risks deploying unreliable or inconsistent models in real-world healthcare scenarios.

***Challenge-2*: Vacancy of Compressed Medical LLM Practical Evaluation:** Model compression is essential for deploying LLMs in medical applications, yet standardized evaluation frameworks for assessing compressed LLMs remain undeveloped(Yang et al. 2024a; Guo, Xu, and Ouyang 2023). This gap impedes the identification of optimal model-compression configurations, limiting performance-efficiency trade-off in diverse medical contexts, including clinical diagnosis, telemedicine, and real-time health monitoring. This vacancy is critical in resource-constrained settings like rural clinics, where models must balance efficiency (e.g., low latency, minimal memory) with domain-specific accuracy and privacy compliance(Guo, Liu, and Xu 2022; Guo, Xu, and Lu 2023). A framework for practical, comparative evaluations is essential to guide effective model de-

*Corresponding author: jinyangguo@buaa.edu.cn

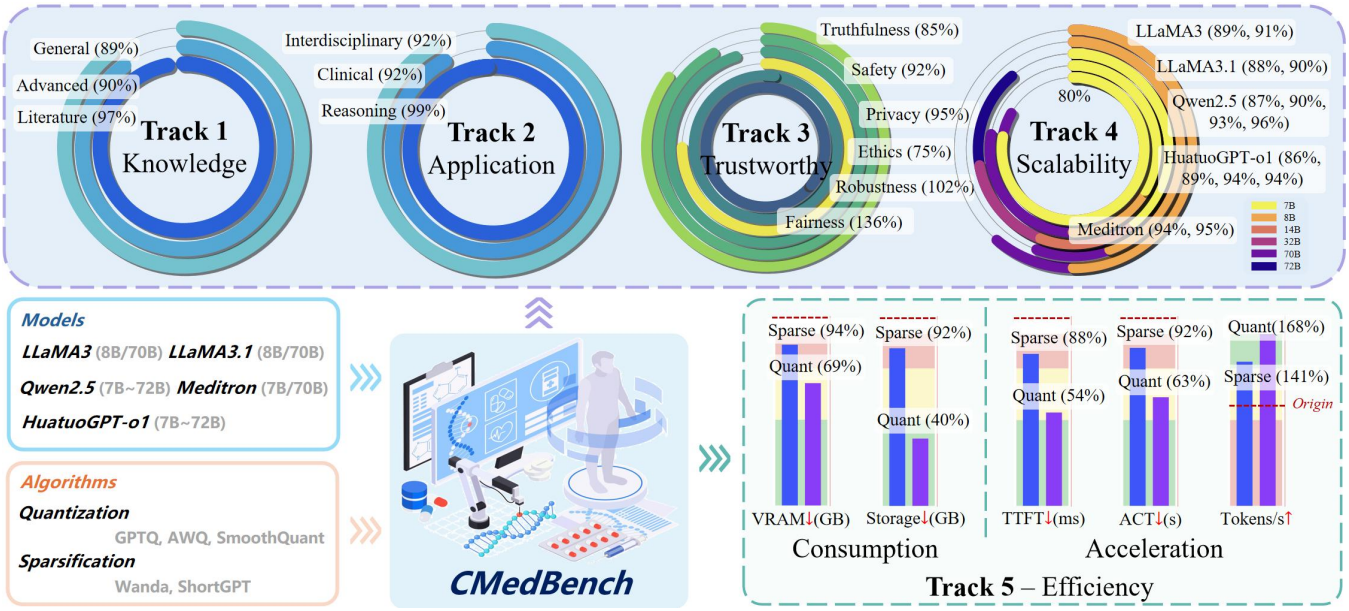

Figure 1: Overview of CMedBench, which benchmarks the performance of training free compression algorithms on a range of comprehensive real medical tracks, including Medical knowledge, Application, Trustworthy, Scalability and Efficiency.

ployment.

In this work, we present CMedBench, the first comprehensive benchmark designed to systematically evaluate the effects of compressed medical LLMs across five core dimensions of medical contexts. Our contributions are threefold: **(1)** we identify the critical limitations of current evaluation practices for compressed LLMs in medical contexts and articulate the unique challenges posed by healthcare LLM deployment; **(2)** we introduce a robust benchmark that aligns compression evaluation with interpretable criteria, emphasizing both model quality and deployment feasibility; and **(3)** we conduct extensive empirical studies across **14** model architectures, **31** datasets, and **11 training free compression settings** in **5** critical domains, offering new insights into the principle of trade-offs between model efficiency and clinical readiness.

CMedBench encompasses the direction of efficient LLM development and deployment for researchers and healthcare practitioners. enabling the principled development and selection of compression strategies that support the versatile, trustworthy, and efficient adoption of LLMs in healthcare.

CMedBench: Tracks and Metrics

This section introduces CMedBench's comprehensive framework to evaluate compressed LLMs, focusing on performance impact and efficiency gains.

Track 1: Medical Knowledge Ability

The exceptional capabilities of large language models (LLMs) in medical tasks, such as consultation, diagnosis, treatment planning, and patient education, fuel their growing adoption in healthcare. Thus, preserving these capabilities is essential for LLM deployment under resource-constrained medical contexts. CMedBench firstly evaluates compression performance across three fundamental dimensions:

General Medical Knowledge (GMK): Tasks requiring the recall and application of foundational medical knowledge, such as answering questions on anatomy, pharmacology, or pathology. **Advanced Medical Knowledge(AMK)**: Tasks demanding professional expertise, evaluating the medical knowledge that needs to be mastered in the practising physician examination.

Biomedical Literature Comprehension (BLC): Tasks focused on extracting and synthesizing information from complex medical texts, simulating evidence-based decision-making in complicated scenarios.

These task dimensions ensure that CMedBench captures the multifaceted demands of medical knowledges, where compressed models must maintain performance across diverse subjects. Unlike existing benchmarks that often focus on isolated tasks, CMedBench evaluates compression models holistically.

To quantitatively measure the performance of each compression models, we define a ***Compression Performance Score(CPS)*** that evaluates the relative performance of compression models across all task dimensions. The score for each task and the overall metric is computed as follows:

$$CPS = \frac{1}{NT}\sum_{n,t}^{N,T} log_2\left(1 + \mathbf{P}^{c}_{\text{task}_{n,t}} \oslash \mathbf{P}_{\text{task}_{n,t}}\right), \quad (1)$$

where $\mathcal{N}$ is the number of model types, $\mathcal{T}$ is the tasks dimensions. $\mathbf{P}_{\text{task}_{n,t}}$ is the performance (e.g., accuracy) vectors of the n-th pretrained model on the t-th task, while $\mathbf{P}^{c}_{\text{task}_{n,j}}$ is the corresponding performance of the compressed model, $\oslash$ denotes the Hadamard division. To ensure a robust measure of performance, we apply an element-wise $\log_{\mathbf{2}}(\mathbf{1}+\mathbf{x})$

transformation to the performance ratios. This transformation effectively compresses larger ratio values and inherently guarantees non-negative results, thereby preventing extreme outliers from disproportionately skewing the overall score. This metric provides a comprehensive and interpretable measure of compression performance, highlighting how well a compressed model preserves the capabilities of its pre-trained counterpart across diverse medical tasks.

Track 2: Medical Application Ability

Beyond preserving foundational medical knowledge, a critical measure of compressed LLMs' performance is their ability to flexibly integrate and apply acquired knowledge effectively. CMedBench assesses compression performance across three critical task dimensions, comprehensively capturing the multifaceted demands of medical applications:

Allied Medical Explanation(AME) Tasks that demand generating clear, concise, and accurate explanations of medical concepts in multidisciplinary domains, such as biomedical engineering, clinical psychology, or occupational therapy. These tasks emulate patient education or interdisciplinary communication, ensuring accessibility and precision in complex medical scenes.

Clinical Diagnostic Assistant(CDA): Tasks involving the application of medical knowledge in clinical scenarios, such as diagnosis, treatment planning, and patient management. These tasks evaluate the model's ability to support clinicians in real-world healthcare settings.

Expert-level Understanding&Reasoning(EUR): Tasks that require integrating medical experiences with complex logical reasoning to solve realistic medical problems, such as expert-level diagnostic reasoning and nuanced understanding of clinical cases.

These task dimensions, encompassing comprehensive clinical diagnostic scenarios as well as critical medical treatment processes, including the proficient use of medical equipment and effective patient communication, ensure that CMedBench thoroughly captures the multifaceted demands of medical applications. This holistic approach evaluates whether the compressed models remain capable of delivering precise, contextually appropriate, and clinically actionable outputs, facilitating reliable deployment in real-world healthcare settings. Following Track 1, Track 2 adopts the ***CPS*** metric to evaluate the medical application capabilities of LLMs across AME, CDA, and EUR tasks.

Track 3: Trustworthiness Maintenance

Deploying LLMs in medical settings demands rigorous trustworthiness to ensure safe and reliable clinical outcomes. Unlike less critical domains, errors in medical LLMs can jeopardize patient care and decision-making. Thus, beyond task-specific performance, our study evaluates the trustworthiness of compressed LLMs across six key trustworthy dimensions to guide effective and secure deployment in healthcare.

Truthfulness: Truthfulness refers to the accuracy and factual correctness of the information generated by the LLM. Ensuring outputs align with established medical knowledge and clinical guidelines, and patient-specific data to prevent harmful errors.

Safety: Overall safety encompasses minimizing the risk of the LLM causing harm to patients, clinicians, or the healthcare system. This is a broad category includes preventing the generation of dangerous advice, avoiding actions that could lead to physical or psychological harm, and ensuring models operate reliably within their intended scope of use.

Privacy: Given the highly sensitive nature of medical data, privacy protection is critical. This involves ensuring that LLMs handle patient information securely, comply with relevant regulations (like HIPAA(Gostin, Levit, and Nass 2009)), and do not inadvertently reveal protected health information through their outputs.

Ethical: This concerns the alignment of the LLM's behavior and decision-making processes with ethical principles relevant to healthcare. Including aspects like not promoting harmful practices, respecting patient autonomy, maintaining professional boundaries, and ensuring accountability for the model's actions or recommendations.

Robustness: Robustness refers to the LLM's ability to maintain performance when faced with variations, noise, or adversarial attacks. In medical context, this is vital because real-world clinical data can be messy, incomplete, or presented in unexpected ways. A non-robust model could fail or produce unreliable outputs under such conditions.

Fairness: Fairness in medical LLMs means that the model's outputs and performance are not biased towards or against any particular group of individuals based on sensitive attributes such as race and sex. Unfair models could lead to disparities in healthcare access, quality, or outcomes.

To provide a comprehensive evaluation across all relevant aspects, we define the metric for each dimension of this Trustworthiness track as:

$$TWY = \frac{1}{NT} \sum_{n,t}^{N,T} \exp\left[\mathbf{d}_{\text{sub}_\mathbf{n}} \cdot \left(\mathbf{P}^{\text{c}}_{\text{sub}_{n,t}} - \mathbf{P}_{\text{sub}_{n,t}}\right)\right]. \quad (2)$$

N is the number of model types and T is the index of trustworthy dimensions. $\mathbf{P}_{\text{sub}_{n,t}}$ represents the reliability (e.g. accuracy) of the original pretrained model on sub-task t, while $\mathbf{P}^{\text{c}}_{\text{sub}_{n,t}}$ is the corresponding performance of the compressed model. Besides, we apply an element-wise $\exp(x)$ transformation to the performance ratios with task-specific exponent $\mathbf{d}_{\text{sub}_t}$ to ensure non-negative metric values for smooth assessment.

Track 4: Compression Cross Combination

Modern hospital environments feature diverse computational infrastructures and clinical demands, requiring varied LLM architectures and scales (Singhal et al. 2023; Gilson et al. 2023). Global disparities in healthcare resources further amplify this heterogeneity, with resource-rich institutions deploying advanced models and under-resourced areas relying on smaller systems. Effective compression algorithms are thus essential to ensure performance, efficiency, and clinical utility across diverse settings.

In this section, we evaluate compression strategies across model scales and architectures, quantifying performance

variability and degradation via the Cross Combination Score (CCS), providing insights for selecting optimal model–compression pairings in heterogeneous medical contexts:

$$CCS = \frac{1}{NM} \sum_{n,m}^{N,M} log_2 \left(1 + \mathbf{P}^{\mathrm{c}}_{\mathrm{set}_{n,m}} \oslash \mathbf{P}_{\mathrm{set}_n}\right), \quad (3)$$

where $\mathbf{P}_{\mathrm{set}_n}$ and $\mathbf{P}^{\mathrm{c}}_{\mathrm{set}_{n,m}}$ represents the accuracy of original model and compressed models for the combination of m-th compression algorithm and the n-th model type, respectively. M and N denotes the total number of compression algorithms and models, respectively.

Track 5: Computational Efficiency

A central motivation for compressing LLMs is to reduce the substantial computational and storage demands associated with their large-scale architectures, thereby improving deployment efficiency.

Computational Resource Consumption: This is particularly critical in resource-constrained environments such as isolated medical systems, where strict privacy regulations and customization requirements often preclude the use of external computational resources. Unlike professional data centers, these systems typically have limited processing power and memory, necessitating advanced compression techniques to enable efficient and secure local inference without compromising model performance. We evaluate computational efficiency across three key dimensions:

VRAM: The video random-access memory usage is an essential metric for deployment on memory-limited hardware, such as edge devices or local medical servers.

Bits: The number of model parameters, influencing storage requirements and inference latency.

Inference Acceleration: The efficiency of medical LLMs is another critical factor, especially in isolated medical settings where real-time responsiveness is essential for applications such as clinical decision support, patient communication, and diagnostic assistance. To assess acceleration performance, we consider three primary metrics:

Time to First Token (TTFT): The latency until the first token is produced, a key indicator for interactive applications requiring fast initial feedback.

Average Completion Time (ACT): The total time to generate a complete response, important for use cases like medical reporting where both speed and completeness matter.

Tokens Per Second (TPS): The model's token generation throughput, relevant for batch processing and high-volume medical tasks such as automated triaging.

Building upon the aforementioned dimensions, we introduce a unified metric—the *Computational Efficiency Score (CES)*— to quantitatively assess the computational performance of compressed LLMs under the multifaceted constraints of medical deployment scenarios.

$$CES = \frac{1}{NT} \sum_{n,t}^{N,T} \mathbf{P}^{\mathrm{c}}_{\mathrm{eff}_{n,t}} \oslash \mathbf{P}_{\mathrm{eff}_{n,t}} \quad (4)$$

Here, N denotes the number of model variants evaluated, and T corresponds to the computational efficiency dimensions discussed previously. $\mathbf{P}_{\mathrm{eff}_{n,t}}$ and $\mathbf{P}^{\mathrm{c}}_{\mathrm{eff}_{n,t}}$ represent the performance of the original and compressed models, respectively, on dimension t.

CMedBench Implementation Details

Implementation Framework

Owing to their ease of implementation and widespread adoption, this study focuses on evaluating the efficacy of training-free model compression techniques, specifically post-training quantization/sparsity. To this end, we investigate multiple compression strategies, including weight-only quantization: (GPTQ(Frantar et al. 2023) and AWQ(Lin et al. 2023)). and weight-activation quantization SmoothQuant(Xiao et al. 2023). The unstructured sparsity algorithm Wanda(Sun et al. 2023) is configured with sparsity ratios of 50% to investigate performance sensitivity to sparsity levels. For training free structured sparsity, ShortGPT(Men et al. 2024) is applied with pruning set to approximately 25% layers for the highest compression ratio while maintaining LLM's performance. All hyperparameters adhere to the official configurations provided by the respective compression methods to ensure fidelity to their original implementations(Gong et al. 2024).

Evaluation Protocol

The evaluation protocol is structured across five distinct tracks to comprehensively assess the capabilities of compressed LLMs in medical contexts. For most tracks, we adopt four LLMs for the evaluations: Meditron-7B(Chen et al. 2024c), HuatuoGPT-o1-8B(Chen et al. 2024b), LLaMA3-8B(Dubey et al. 2024), and Qwen2.5-7B(Yang et al. 2024b), with the exception of track 4, to cover various LLM structures and medical tuning methods.

We evaluate compressed LLMs across five tracks: (1) **Medical Knowledge** uses MMLU-Health(Hendrycks et al. 2021), MedQA(Jin et al. 2021), and PubMedQA(Jin et al. 2019) to assess factual accuracy; (2) **Medical Application** tests complex queries with MedexQA(Kim et al. 2024), MedMCQA(Pal, Umapathi, and Sankarasubbu 2022), CareQA(Arias-Duart et al. 2025), MedBullets(Chen et al. 2024a), Jmed(Wang et al. 2025a), and MedXpertQA(Zuo et al. 2025); (3) **Trustworthiness** examines truthfulness, safety, privacy, ethics, robustness, and fairness per the TrustLLM framework(Sun et al. 2024); (4) **Compression Cross Combination** benchmarks LLaMA3(Dubey et al. 2024), LLaMA3.1(Dubey et al. 2024), Qwen2.5(Yang et al. 2024b), HuatuoGPT-o1(Chen et al. 2024b), and Meditron (7B–72B parameters)(Chen et al. 2024c) on MMLU-Health(Hendrycks et al. 2021); and (5) **Computational Efficiency** measures inference performance on A800 using

Each method received over one hundred citations in one year, and publicly available.

Owing to space constraints, we present selected representative results herein; for a more detailed and comprehensive set of experiment setting and results, please refer to the Supplementary.

Method	Sparsity/ #Bits	Model	GMK						AMK	BLC	CPS
			$MMLU_{AN}$	$MMLU_{CK}$	$MMLU_{CB}$	$MMLU_{CM}$	$MMLU_{MG}$	$MMLU_{PM}$	MedQA	PubMedQA	
Dense	N/A	Meditron-7B	$37.04_{1.00}$	$44.53_{1.00}$	$35.42_{1.00}$	$30.06_{1.00}$	$47.00_{1.00}$	$38.24_{1.00}$	$36.84_{1.00}$	$55.30_{1.00}$	1.00
		HtGPTo1-8B	$47.41_{1.00}$	$47.17_{1.00}$	$48.61_{1.00}$	$38.73_{1.00}$	$51.00_{1.00}$	$42.65_{1.00}$	$42.11_{1.00}$	$58.00_{1.00}$	
		LLaMA3-8B	$41.48_{1.00}$	$50.19_{1.00}$	$45.83_{1.00}$	$35.26_{1.00}$	$54.00_{1.00}$	$41.91_{1.00}$	$40.38_{1.00}$	$58.20_{1.00}$	
		Qwen2.5-7B	$48.89_{1.00}$	$50.57_{1.00}$	$46.53_{1.00}$	$43.93_{1.00}$	$56.00_{1.00}$	$48.53_{1.00}$	$39.91_{1.00}$	$52.00_{1.00}$	
AWQ	4	Meditron-7B	$37.04_{1.00}$	$43.40_{0.98}$	$36.11_{1.01}$	$30.06_{1.00}$	$48.00_{1.02}$	$38.24_{1.00}$	$35.66_{0.98}$	$54.80_{0.99}$	0.99
		HtGPTo1-8B	$46.67_{0.99}$	$47.17_{1.00}$	$46.53_{0.97}$	$41.04_{1.04}$	$47.00_{0.94}$	$45.22_{1.04}$	$41.08_{0.98}$	$58.40_{1.00}$	
		LLaMA3-8B	$41.48_{1.00}$	$47.17_{0.96}$	$48.61_{1.04}$	$34.68_{0.99}$	$52.00_{0.97}$	$43.01_{1.02}$	$39.20_{0.98}$	$57.30_{0.99}$	
		Qwen2.5-7B	$45.19_{0.94}$	$49.06_{0.98}$	$45.83_{0.99}$	$42.77_{0.98}$	$53.00_{0.96}$	$45.96_{0.96}$	$38.88_{0.98}$	$53.60_{1.02}$	
GPTQ	4	Meditron-7B	$39.26_{1.04}$	$44.91_{1.01}$	$37.50_{1.04}$	$32.95_{1.07}$	$45.00_{0.97}$	$37.50_{0.99}$	$36.21_{0.99}$	$55.00_{1.00}$	0.97
		HtGPTo1-8B	$41.48_{0.91}$	$46.42_{0.99}$	$43.75_{0.93}$	$36.99_{0.97}$	$44.00_{0.90}$	$40.07_{0.96}$	$39.28_{0.95}$	$58.20_{1.00}$	
		LLaMA3-8B	$37.78_{0.93}$	$46.04_{0.94}$	$43.75_{0.97}$	$32.37_{0.94}$	$48.00_{0.92}$	$40.44_{0.97}$	$36.45_{0.93}$	$56.20_{0.97}$	
		Qwen2.5-7B	$48.89_{1.00}$	$47.92_{0.96}$	$43.06_{0.95}$	$39.31_{0.92}$	$50.00_{0.92}$	$43.01_{0.92}$	$38.88_{0.98}$	$53.90_{1.03}$	
SMQU	4	Meditron-7B	$28.89_{0.83}$	$40.00_{0.92}$	$30.56_{0.90}$	$28.32_{0.96}$	$36.00_{0.82}$	$31.99_{0.88}$	$30.40_{0.87}$	$47.20_{0.89}$	0.80
		HtGPTo1-8B	$28.89_{0.69}$	$33.58_{0.78}$	$22.92_{0.56}$	$26.59_{0.75}$	$34.00_{0.74}$	$28.68_{0.74}$	$24.74_{0.67}$	$47.20_{0.86}$	
		LLaMA3-8B	$30.37_{0.79}$	$32.08_{0.71}$	$34.03_{0.80}$	$22.54_{0.71}$	$33.00_{0.69}$	$25.74_{0.69}$	$27.57_{0.75}$	$52.40_{0.93}$	
		Qwen2.5-7B	$34.07_{0.76}$	$38.87_{0.82}$	$36.81_{0.84}$	$34.68_{0.84}$	$39.00_{0.76}$	$36.03_{0.80}$	$31.11_{0.83}$	$54.00_{1.03}$	
Wanda	50%	Meditron-7B	$31.85_{0.90}$	$38.11_{0.89}$	$34.03_{0.97}$	$26.01_{0.90}$	$44.00_{0.95}$	$37.50_{0.99}$	$33.07_{0.92}$	$55.40_{1.00}$	0.90
		HtGPTo1-8B	$32.59_{0.75}$	$40.75_{0.90}$	$40.28_{0.87}$	$29.48_{0.82}$	$47.00_{0.94}$	$38.60_{0.93}$	$35.82_{0.89}$	$57.20_{0.99}$	
		LLaMA3-8B	$34.07_{0.87}$	$41.13_{0.86}$	$39.58_{0.90}$	$26.59_{0.81}$	$40.00_{0.80}$	$34.19_{0.86}$	$34.33_{0.89}$	$56.80_{0.98}$	
		Qwen2.5-7B	$40.74_{0.87}$	$46.04_{0.93}$	$45.14_{0.98}$	$36.99_{0.88}$	$45.00_{0.85}$	$41.54_{0.89}$	$35.35_{0.92}$	$55.40_{1.05}$	
ShortGPT	25%	Meditron-7B	$26.67_{0.78}$	$29.06_{0.72}$	$31.25_{0.91}$	$27.17_{0.93}$	$34.00_{0.79}$	$20.59_{0.62}$	$27.73_{0.81}$	$55.20_{1.00}$	0.74
		HtGPTo1-8B	$27.41_{0.66}$	$30.19_{0.71}$	$26.39_{0.63}$	$23.70_{0.69}$	$31.00_{0.69}$	$26.84_{0.70}$	$25.84_{0.69}$	$55.30_{0.97}$	
		LLaMA3-8B	$25.93_{0.70}$	$28.30_{0.65}$	$28.47_{0.70}$	$25.43_{0.78}$	$35.00_{0.72}$	$23.90_{0.65}$	$27.89_{0.76}$	$55.50_{0.97}$	
		Qwen2.5-7B	$21.48_{0.53}$	$24.91_{0.58}$	$20.83_{0.53}$	$26.59_{0.68}$	$35.00_{0.70}$	$29.78_{0.69}$	$24.90_{0.70}$	$55.20_{1.04}$	

Table 1: Performance of Compressed Models on Medical Dimensions: General Medical Knowledge (GMK), Advanced Medical Knowledge (AMK), and Biomedical Literature Comprehension (BLC). Accuracy is shown in **black**; Compression Performance Score (CPS) in **gray**. Cell colors indicate CPS: **blue** (< 1, worse than original model), **red** (> 1, better than original model).

vLLM(Kwon et al. 2023) frameworks. This structured evaluation framework ensures a rigorous and comprehensive analysis of compressed LLMs in medical contexts, addressing both performance and practical deployment considerations.

Evaluation and Analysis

Track 1: Medical Knowledge Ability

Quantization configurations effectively preserve the medical knowledge capabilities. Table 1 presents the performance metrics for Track 1. The results demonstrate that most quantization configurations nearly losslessly compress LLMs across diverse model architectures, with the notable exception of the 4-bit SmoothQuant method, which exhibits suboptimal performance. In contrast, unstructural sparsity technique: Wanda, generally underperform in this domain. At a 50% pruning ratio, model performance declines notably across all evaluated models. The structured pruning algorithm, ShortGPT, exhibits particularly pronounced degradation. Across all four evaluated models, rendering these settings unsuitable for medical applications.

Most medical knowledge dimensions exhibit similar performance change. From the perspective of medical knowledge preservation, the *CPS* of LLMs are greater than 0.8, indicating that most compression settings successfully maintain the medical knowledge proficiency of LLMs. Notably, the *CPS* for BLC remains consistent in most settings, underscoring its robustness relative to other evaluated tasks. This suggests that BLC is less susceptible to performance degradation under various compression settings.

Track 2: Medical Application Ability

Compression has a similar impact on medical applications as on medical knowledge. Table 2 evaluating the accuracy of compressed LLMs on medical application-related datasets. Quantization methods mainly maintain robust medical capabilities across models. The 4-bit SmoothQuant setting consistently underperforms, especially for the LLaMA3-8B and HuatuoGPT-o1-8B models, with notable accuracy declines. Unstructured sparsity Wanda shows stable performance at 50% pruning. The structured pruning algorithm ShortGPT similarly demonstrates severe performance degradation.

The sensitivity varies among medical application abilities Notably, the *CPS* of $\texttt{MedexQA}_{\texttt{SP}}$ and $\texttt{MedXpert}_{\texttt{R}}$ remain consistent across various compressed models. suggesting that these tasks are particularly resilient to compression-induced degradation. This phenomenon shows the robust metrics in deploying compressed LLMs, but they could not individually reveal the gap between compression settings. While the Jmed dataset appears to be most sensitive to sparsity, which results in a devastating decline in various sparsity settings.

Track 3: Trustworthiness Maintenance

Truthfulness, Safety, and Ethics Mirror Performance Trends As shown in Figure 2, Truthfulness, Safety, and Ethics exhibit trends analogous to those observed in

Method	Sparsity /#Bits	Model	AME					CDA				EUR		CPS
			$\text{MedexQA}_{\text{BE}}$	$\text{MedexQA}_{\text{CLS}}$	$\text{MedexQA}_{\text{CP}}$	$\text{MedexQA}_{\text{OT}}$	$\text{MedexQA}_{\text{SP}}$	MedMCQA	CareQA	MedBullets	Jmed	$\text{MedXpert}_{\text{R}}$	$\text{MedXpert}_{\text{U}}$	
Dense	N/A	Meditron-7B	$37.76_{1.00}$	$43.28_{1.00}$	$33.96_{1.00}$	$44.97_{1.00}$	$25.38_{1.00}$	$33.56_{1.00}$	$35.10_{1.00}$	$25.00_{1.00}$	$8.01_{1.00}$	$7.90_{1.00}$	$10.53_{1.00}$	1.00
		HtGPTo1-8B	$51.75_{1.00}$	$59.14_{1.00}$	$46.23_{1.00}$	$57.14_{1.00}$	$29.23_{1.00}$	$38.35_{1.00}$	$39.55_{1.00}$	$32.47_{1.00}$	$16.22_{1.00}$	$7.20_{1.00}$	$9.68_{1.00}$	
		LLaMA3-8B	$47.55_{1.00}$	$53.76_{1.00}$	$44.34_{1.00}$	$50.26_{1.00}$	$34.62_{1.00}$	$37.84_{1.00}$	$40.06_{1.00}$	$27.27_{1.00}$	$16.82_{1.00}$	$8.65_{1.00}$	$11.04_{1.00}$	
		Qwen2.5-7B	$44.06_{1.00}$	$55.38_{1.00}$	$44.34_{1.00}$	$52.91_{1.00}$	$25.38_{1.00}$	$36.03_{1.00}$	$39.53_{1.00}$	$29.22_{1.00}$	$15.92_{1.00}$	$7.85_{1.00}$	$11.21_{1.00}$	
AWQ	4	Meditron-7B	$38.46_{1.01}$	$45.43_{1.04}$	$33.02_{0.98}$	$48.68_{1.06}$	$26.15_{1.02}$	$33.40_{1.00}$	$35.31_{1.00}$	$23.38_{0.95}$	$9.21_{1.10}$	$8.01_{1.01}$	$9.34_{0.92}$	1.00
		HtGPTo1-8B	$47.55_{0.94}$	$55.65_{0.96}$	$45.28_{0.99}$	$54.50_{0.97}$	$29.23_{1.00}$	$37.39_{0.98}$	$38.71_{0.98}$	$28.57_{0.91}$	$14.51_{0.92}$	$7.47_{1.03}$	$10.19_{1.04}$	
		LLaMA3-8B	$46.85_{0.99}$	$54.03_{1.00}$	$43.40_{0.98}$	$52.91_{1.04}$	$33.08_{0.97}$	$37.20_{0.99}$	$39.53_{0.99}$	$28.90_{1.04}$	$18.52_{1.07}$	$8.38_{0.98}$	$12.39_{1.09}$	
		Qwen2.5-7B	$44.06_{1.00}$	$55.11_{1.00}$	$47.17_{1.05}$	$52.91_{1.00}$	$25.38_{1.00}$	$35.67_{0.99}$	$38.36_{0.98}$	$26.95_{0.94}$	$17.12_{1.05}$	$8.01_{1.01}$	$11.04_{0.99}$	
GPTQ	4	Meditron-7B	$39.86_{1.04}$	$46.51_{1.05}$	$34.91_{1.02}$	$47.09_{1.03}$	$24.62_{0.98}$	$33.54_{1.00}$	$35.01_{1.00}$	$23.70_{0.96}$	$10.01_{1.17}$	$7.68_{0.98}$	$10.70_{1.01}$	0.99
		HtGPTo1-8B	$46.85_{0.93}$	$53.76_{0.93}$	$43.40_{0.96}$	$52.38_{0.94}$	$27.69_{0.96}$	$35.33_{0.94}$	$37.61_{0.96}$	$27.92_{0.90}$	$17.22_{1.04}$	$8.17_{1.09}$	$10.02_{1.03}$	
		LLaMA3-8B	$42.66_{0.92}$	$50.27_{0.95}$	$44.34_{1.00}$	$47.09_{0.95}$	$33.08_{0.97}$	$35.60_{0.96}$	$37.38_{0.95}$	$25.32_{0.95}$	$20.52_{1.15}$	$8.60_{1.00}$	$10.02_{0.93}$	
		Qwen2.5-7B	$41.96_{0.97}$	$52.69_{0.96}$	$41.51_{0.95}$	$49.74_{0.96}$	$27.69_{1.06}$	$35.60_{0.99}$	$37.47_{0.96}$	$24.03_{0.87}$	$16.12_{1.01}$	$8.06_{1.02}$	$10.87_{0.98}$	
SMQU	4	Meditron-7B	$36.36_{0.97}$	$35.75_{0.87}$	$29.25_{0.90}$	$38.10_{0.89}$	$27.69_{1.06}$	$29.48_{0.91}$	$30.40_{0.90}$	$20.13_{0.85}$	$9.51_{1.13}$	$7.90_{1.00}$	$8.83_{0.88}$	0.88
		HtGPTo1-8B	$32.87_{0.71}$	$33.06_{0.64}$	$31.13_{0.74}$	$33.86_{0.67}$	$26.92_{0.94}$	$26.37_{0.75}$	$26.92_{0.75}$	$20.78_{0.71}$	$15.42_{0.96}$	$8.44_{1.12}$	$10.36_{1.05}$	
		LLaMA3-8B	$32.87_{0.76}$	$32.26_{0.68}$	$29.25_{0.73}$	$37.57_{0.81}$	$33.85_{0.98}$	$29.05_{0.82}$	$28.52_{0.78}$	$20.78_{0.82}$	$14.71_{0.91}$	$8.01_{0.95}$	$9.51_{0.90}$	
		Qwen2.5-7B	$36.36_{0.87}$	$45.43_{0.86}$	$34.91_{0.84}$	$46.03_{0.90}$	$28.46_{1.08}$	$32.15_{0.92}$	$33.61_{0.89}$	$25.97_{0.92}$	$12.41_{0.83}$	$7.79_{0.99}$	$9.68_{0.90}$	
Wanda	50%	Meditron-7B	$39.16_{1.03}$	$41.13_{0.96}$	$28.30_{0.87}$	$40.74_{0.93}$	$23.08_{0.93}$	$30.96_{0.94}$	$32.61_{0.95}$	$23.70_{0.96}$	$14.41_{1.48}$	$7.25_{0.94}$	$10.02_{0.96}$	0.96
		HtGPTo1-8B	$40.56_{0.83}$	$45.97_{0.83}$	$44.34_{0.97}$	$50.26_{0.91}$	$29.23_{1.00}$	$32.90_{0.89}$	$35.12_{0.92}$	$24.03_{0.80}$	$20.42_{1.18}$	$6.99_{0.98}$	$9.85_{1.01}$	
		LLaMA3-8B	$41.26_{0.90}$	$44.89_{0.88}$	$43.40_{0.98}$	$46.56_{0.95}$	$30.00_{0.90}$	$31.99_{0.88}$	$34.07_{0.89}$	$25.00_{0.94}$	$18.42_{1.07}$	$8.54_{0.99}$	$11.54_{1.03}$	
		Qwen2.5-7B	$41.26_{0.95}$	$49.73_{0.92}$	$38.68_{0.90}$	$49.21_{0.95}$	$30.00_{1.13}$	$33.16_{0.94}$	$36.11_{0.94}$	$24.03_{0.87}$	$18.02_{1.09}$	$7.79_{0.99}$	$11.04_{0.99}$	
ShortGPT	25%	Meditron-7B	$33.57_{0.92}$	$37.90_{0.91}$	$26.42_{0.83}$	$34.39_{0.82}$	$30.00_{1.13}$	$27.32_{0.86}$	$28.61_{0.86}$	$20.45_{0.86}$	$12.11_{1.33}$	$6.66_{0.88}$	$6.28_{0.67}$	0.77
		HtGPTo1-8B	$27.97_{0.62}$	$31.72_{0.62}$	$26.42_{0.65}$	$30.16_{0.61}$	$26.92_{0.94}$	$26.42_{0.76}$	$28.86_{0.79}$	$17.53_{0.62}$	$5.91_{0.45}$	$6.77_{0.96}$	$6.79_{0.77}$	
		LLaMA3-8B	$28.67_{0.68}$	$36.56_{0.75}$	$29.25_{0.73}$	$40.74_{0.86}$	$27.69_{0.85}$	$27.35_{0.78}$	$30.31_{0.81}$	$16.88_{0.70}$	$6.01_{0.44}$	$7.15_{0.87}$	$7.30_{0.73}$	
		Qwen2.5-7B	$20.98_{0.56}$	$26.34_{0.56}$	$32.08_{0.79}$	$29.10_{0.63}$	$23.85_{0.96}$	$25.32_{0.77}$	$24.05_{0.69}$	$19.81_{0.75}$	$5.01_{0.39}$	$8.60_{1.07}$	$7.47_{0.74}$	

Table 2: Performance of Compressed Models on Medical Tasks: Allied Medical Explanation (AME), Clinical Diagnostic Assistant (CDA), and Expert-level Understanding & Reasoning (EUR).

Model	Dense	GPTQ		AWQ		SMQU		WANDA			ShortGPT		CCS
		4bit	8bit	4bit	8bit	4bit	8bit	25%	50%	75%	15%	25%	
LLaMA3-8B	$44.78_{1.00}$	$41.40_{0.94}$	$44.50_{1.00}$	$44.49_{1.00}$	$44.54_{1.00}$	$29.63_{0.73}$	$44.31_{0.99}$	$43.88_{0.99}$	$35.93_{0.85}$	$26.41_{0.67}$	$30.74_{0.75}$	$27.84_{0.70}$	0.89
LLaMA3-70B	$52.42_{1.00}$	$43.34_{0.87}$	$52.73_{1.00}$	$51.75_{0.99}$	$52.86_{1.01}$	$26.21_{0.58}$	$52.67_{1.00}$	$51.83_{0.99}$	$48.19_{0.94}$	$33.17_{0.71}$	$49.62_{0.96}$	$44.27_{0.88}$	0.91
LLaMA3.1-8B	$45.48_{1.00}$	$42.85_{0.96}$	$45.23_{1.00}$	$43.73_{0.97}$	$45.35_{1.00}$	$29.57_{0.72}$	$45.06_{0.99}$	$43.06_{0.96}$	$36.82_{0.86}$	$27.56_{0.68}$	$31.09_{0.75}$	$28.82_{0.71}$	0.88
LLaMA3.1-70B	$53.05_{1.00}$	$46.28_{0.90}$	$50.29_{0.96}$	$46.28_{0.90}$	$53.33_{1.00}$	$29.76_{0.64}$	$52.82_{1.00}$	$51.35_{0.98}$	$48.43_{0.94}$	$31.37_{0.67}$	$48.84_{0.94}$	$44.65_{0.88}$	0.90
Qwen2.5-7B	$49.08_{1.00}$	$45.37_{0.94}$	$49.12_{1.00}$	$46.97_{0.97}$	$49.26_{1.00}$	$36.58_{0.80}$	$48.70_{0.99}$	$48.23_{0.99}$	$42.58_{0.90}$	$25.71_{0.61}$	$28.44_{0.66}$	$26.43_{0.62}$	0.87
Qwen2.5-14B	$48.18_{1.00}$	$46.65_{0.98}$	$48.24_{1.00}$	$44.57_{0.94}$	$48.52_{1.01}$	$35.02_{0.79}$	$48.65_{1.01}$	$46.35_{0.97}$	$41.18_{0.89}$	$24.24_{0.59}$	$35.58_{0.80}$	$36.41_{0.81}$	0.90
Qwen2.5-32B	$49.73_{1.00}$	$46.86_{0.96}$	$49.23_{0.99}$	$48.51_{0.98}$	$49.77_{1.00}$	$41.28_{0.87}$	$49.58_{1.00}$	$49.32_{0.99}$	$47.76_{0.97}$	$30.50_{0.69}$	$43.41_{0.91}$	$36.43_{0.79}$	0.93
Qwen2.5-72B	$48.44_{1.00}$	$48.38_{1.00}$	$48.89_{1.01}$	$48.75_{1.00}$	$48.63_{1.00}$	$41.03_{0.89}$	$49.06_{1.01}$	$48.58_{1.00}$	$49.48_{1.02}$	$34.93_{0.78}$	$43.72_{0.93}$	$38.92_{0.85}$	0.96
HuatuoGPT-o1-8B	$45.93_{1.00}$	$42.12_{0.94}$	$45.71_{1.00}$	$45.60_{0.99}$	$45.80_{1.00}$	$29.11_{0.71}$	$45.21_{0.99}$	$47.78_{1.03}$	$38.12_{0.87}$	$25.82_{0.64}$	$32.39_{0.77}$	$27.59_{0.68}$	0.89
HuatuoGPT-o1-70B	$49.06_{1.00}$	$47.75_{0.98}$	$50.25_{1.02}$	$49.26_{1.00}$	$48.97_{1.00}$	$31.12_{0.71}$	$49.21_{1.00}$	$49.91_{1.01}$	$46.97_{0.97}$	$32.16_{0.73}$	$44.79_{0.94}$	$41.83_{0.89}$	0.94
HuatuoGPT-o1-7B	$50.25_{1.00}$	$46.86_{0.95}$	$50.70_{1.01}$	$45.84_{0.94}$	$50.43_{1.00}$	$35.42_{0.77}$	$49.66_{0.99}$	$49.52_{0.99}$	$43.14_{0.89}$	$26.23_{0.61}$	$25.16_{0.59}$	$24.84_{0.58}$	0.86
HuatuoGPT-o1-72B	$51.75_{1.00}$	$52.22_{1.01}$	$52.06_{1.00}$	$53.92_{1.03}$	$51.99_{1.00}$	$39.86_{0.82}$	$51.87_{1.00}$	$52.06_{1.00}$	$52.51_{1.01}$	$37.64_{0.79}$	$44.77_{0.90}$	$34.29_{0.73}$	0.94
Meditron-7B	$38.72_{1.00}$	$39.52_{1.01}$	$38.58_{1.00}$	$38.81_{1.00}$	$38.67_{1.00}$	$32.63_{0.88}$	$38.69_{1.00}$	$39.02_{1.01}$	$35.25_{0.93}$	$26.69_{0.76}$	$33.27_{0.89}$	$28.12_{0.79}$	0.94
Meditron-70B	$44.61_{1.00}$	$45.43_{1.01}$	$44.67_{1.00}$	$44.88_{1.00}$	$45.32_{1.01}$	$38.26_{0.89}$	$45.03_{1.01}$	$44.38_{1.00}$	$42.26_{0.96}$	$27.06_{0.68}$	$42.13_{0.96}$	$38.79_{0.90}$	0.95
Average	1.00	0.96	1.00	0.98	1.00	0.77	1.00	0.99	0.93	0.69	0.84	0.77	0.91

Table 3: Medical Performance Across Compression Methods and Models

performance-related tracks. This suggests that compressed models with sustained performance are likely to demonstrate greater reliability in these trustworthiness domains.

Characteristics of Privacy, Robustness, and Fairness In contrast to the aforementioned domains, the TWY scores for Privacy remain stable, which only exhibits a tiny decreasing trend across compressed models. Conversely, Robustness exhibits moderate variability, whereas Fairness shows a significant upward trend relative to origin models.

These unpredictable variations underscore the need for systematic evaluation and benchmarking of compressed models to ensure efficient, robust, and trustworthy deployment of medical LLMs.

Track 4: Compression Cross Combination

Table 3 presents the performance of various LLMs across eleven compression settings. The results reveal consistent performance trend across specific compression methods, with performance degradation patterns aligning closely with track1 and track2. But there are also some interesting trends can be discovered in the experiments.

Resilience of Larger Models to ShortGPT. Larger models exhibit greater resilience to performance degradation under the structural sparsity method ShortGPT. This trend is consistent across model families, with larger models demonstrating higher Compression Consistency Scores (CCS) under ShortGPT compression. These findings suggest that

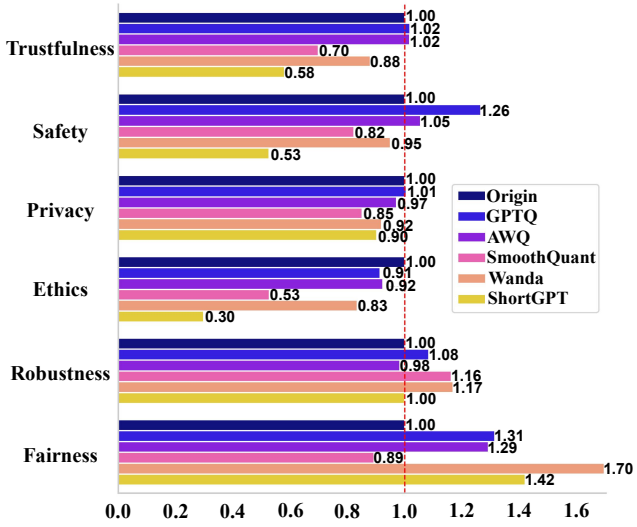

Figure 2: Normalised Trustworthy Metrics(TWY) across 6 Dimensions. Each dimension reflects the TWY score of a dataset group and metrics across 4 advanced LLMs (< 1, worse than original model, > 1, better than original model).

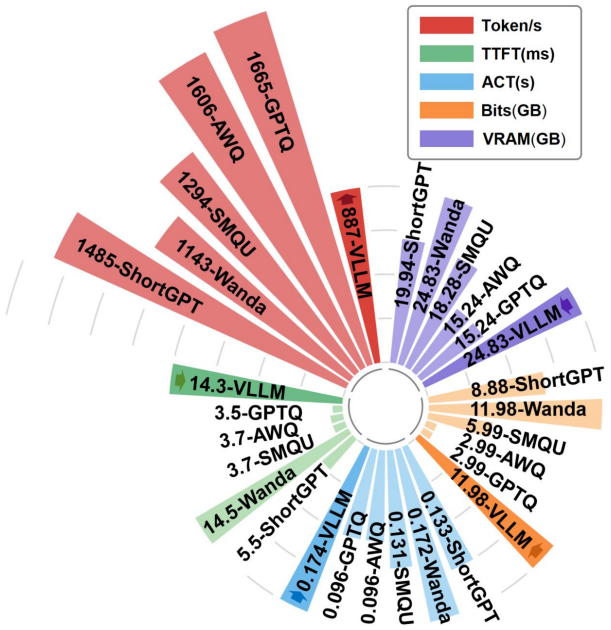

Figure 3: Efficiency of training-free compressed models across five key LLM inference performance indicators, including Resource Consumption&Inference Acceleration aspects (Arrows indicate increasing Efficiency.)

larger models possess greater layer redundancy, mitigating the impact of sparsity-induced compression. This property enhances the feasibility of deploying large-scale LLMs in resource-constrained environments, such as medical applications, where efficiency is critical.

Architectural Sensitivity shown in HuatuoGPT-o1 Series. Models sharing similar architectural foundations exhibit comparable sensitivity to compression, as observed in the HuatuoGPT-o1 series. This series is divided into two groups based on their base architectures: HuatuoGPT Huatuo-o1-70B are derived from LLaMA3.1, while HuatuoGPT-o1-7B and Huatuo-o1-72B are based on Qwen2.5. Direct comparisons within each architecture reveal consistent performance degradation patterns. For example, within LLaMA3.1 foundational group, LLaMA3.1-8B and HuatuoGPT-o1-8B shares the same performance fluctuation across compression settings, this phenomenon can also be observed on its 70B version and the Qwen2.5 based group. These results indicate that the impact of compression is strongly tied to the underlying model architecture, providing a predictive framework for assessing LLM performance in medical contexts where model reliability is paramount.

Track 5: Computational Efficiency

Quantization delivers significant computational efficiency enhancement. Fig 3 illustrates that quantization methods achieve lower resource consumption and faster inference across multiple dimensions. This phenomenon is especially evident in reducing memory consumption and improving overall token generation rate. These findings underscore quantization as a highly effective approach for deploying LLMs in resource-constrained medical settings.

Constraints of Unstructural Sparsity As depicted in Fig 3, unstructured sparsity method Wanda exhibits efficiency similar to dense models under vLLM acceleration. This is primarily due to inadequate hardware backend support for sparsity operators, as most frameworks prioritize quantization and structural sparsity. Developing robust support for sparsity operators is critical to advancing practical and efficient compression algorithms.

Discussion

Evaluation of CMedBench reveals that weight-only quantized LLMs deliver robust medical performance, high trustworthiness, and significant acceleration. making them well-suited for medical applications. Given this trend, we recommend adopting lower-bit settings, such as 4-bit quantization, to achieve acceleration while maintaining acceptable performance to acquire better efficiency in real-time diagnosis assistance. In contrast, weight-activation quantization struggles at low-bit setting, indicating challenges in lossless activation compression to further boost LLM efficiency. While unstructured sparsity maintains performance but requires specialized operators for acceleration, which highlights the limitation of unstructured sparsification's employment in medical LLMs. Structured sparsity achieves acceleration but often sacrifices medical performance, while this limitation is significantly mitigated as the scale of LLMs increases, indicating its potential for boosting large-scale LLM applications with in medical context. Addressing these challenges is crucial for advancing compression techniques to support healthcare applications effectively.

Conclusion

In this work, we introduced the Compressed Medical LLM Benchmark (CMedBench), a systematic framework designed to evaluate the impact of compression on the performance and efficiency of LLMs across a range of medical-context tasks. Through extensive experimentation, we evaluated healthcare-related capabilities across a diverse set of training free compressing algorithms, analyzing the impact of compression techniques on performance. Our findings offer valuable insights into the trade-offs between model efficiency and performance, establishing CMedBench serving as a critical resource to guide the development and deployment of compressed LLMs in medical domains. A limitation of CMedBench is its focus on several representative training free LLM compression techniques, which may not fully capture the diversity of approaches in this rapidly evolving field.

By providing a robust and transparent evaluation platform, we seek to support the responsible development of efficient, high-performing LLMs in medical applications.

Acknowledgments

This work was supported by National Key Research and Development Program of China (2023YFC2506800), the National Natural Science Foundation of China (Nos. 62476018, 62306025), the Fundamental Research Funds for the Central Universities, the Beijing Municipal Science and Technology Project (No. Z231100010323002).

References

Achiam, O. J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.; Balaji, S.; Balcom, V.; Baltescu, P.; Bao, H.-i.; Bavarian, M.; Belgum, J.; Bello, I.; Berdine, J.; Bernadett-Shapiro, G.; Berner, C.; Bogdonoff, L.; Boiko, O.; et al. 2023. GPT-4 Technical Report. *arXiv*.

Arias-Duart, A.; Martin-Torres, P. A.; Hinjos, D.; Perez, P. B.; Ganzabal, L. U.; Mallo, M. G.; Gururajan, A. K.; Lopez-Cuena, E.; Álvarez-Napagao, S.; and Garcia-Gasulla, D. 2025. Automatic Evaluation of Healthcare LLMs Beyond Question-Answering. *arXiv.org*, abs/2502.06666.

Chen, H.; Fang, Z.; Singla, Y.; and Dredze, M. 2024a. Benchmarking Large Language Models on Answering and Explaining Challenging Medical Questions. *arXiv.org*, abs/2402.18060.

Chen, J.; Cai, Z.; Ji, K.; Wang, X.; Liu, W.; Wang, R.; Hou, J.; and Wang, B. 2024b. HuatuoGPT-o1, Towards Medical Complex Reasoning with LLMs. *arXiv.org*, abs/2412.18925.

Chen, Z.; Romanou, A.; Bonnet, A.; Hernández-Cano, A.; Alkhamissi, B.; Matoba, K.; Salvi, F.; Pagliardini, M.; Fan, S.; Köpf, A.; Mohtashami, A.; Sallinen, A.; Swamy, V.; Sakhaeirad, A.; Krawczuk, I.; Bayazit, D.; Marmet, A.; Mi, L.; Boillat-Blanco, N.; Keitel, K.; Elkin, J.; Robert, B.; Montariol, S.; Bressan, S.; Chen, D.; Demers, V.; Emery, N.; Glasson, N.; Mensah, P.; Miauton, A.; Roemer, S.; Siebert, J.; Starvaggi, C.; Suttels, V.; Tan, R.; Taylor, R.; Toit, J. d.; Hartley, M.-A.; Jaggi, M.; and Bosselut, A. 2024c. MEDITRON: Open Medical Foundation Models Adapted for Clinical Practice.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The Llama 3 Herd of Models. *arXiv.org*, abs/2407.21783.

Frantar, E.; Ashkboos, S.; Hoefler, T.; and Alistarh, D. 2023. GPTQ: Accurate Quantization for Generative Pre-trained Transformers. In *The Eleventh International Conference on Learning Representations*, volume abs/2210.17323.

Gilson, A.; Safranek, C. W.; Huang, T.; Socrates, V.; Chi, L.; Taylor, R. A.; and Chartash, D. 2023. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Medical Education*, 9: e45312.

Gong, R.; Yong, Y.; Gu, S.; Huang, Y.; Zhang, Y.; Liu, X.; and Tao, D. 2024. LLMC: Benchmarking Large Language Model Quantization with a Versatile Compression Toolkit. In *Conference on Empirical Methods in Natural Language Processing*.

Gostin, L. O.; Levit, L. A.; and Nass, S. J. 2009. Beyond the HIPAA privacy rule: enhancing privacy, improving health through research.

Guo, J.; Liu, J.; and Xu, D. 2022. 3D-Pruning: A model compression framework for efficient 3D action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12): 8717–8729.

Guo, J.; Ouyang, W.; and Xu, D. 2020. Multi-Dimensional Pruning: A Unified Framework for Model Compression. In *CVPR*.

Guo, J.; Wu, J.; Wang, Z.; Liu, J.; Yang, G.; Ding, Y.; Gong, R.; Qin, H.; and Liu, X. 2024. Compressing large language models by joint sparsification and quantization. In *Forty-first International Conference on Machine Learning*.

Guo, J.; Xu, D.; and Lu, G. 2023. Cbanet: Towards complexity and bitrate adaptive deep image compression using a single network. *IEEE Transactions on Image Processing*.

Guo, J.; Xu, D.; and Ouyang, W. 2023. Multidimensional Pruning and Its Extension: A Unified Framework for Model Compression. *IEEE Transactions on Neural Networks and Learning Systems*.

Guo, J.; Zhang, W.; Ouyang, W.; and Xu, D. 2020. Model compression using progressive channel pruning. *IEEE Transactions on Circuits and Systems for Video Technology*.

He, C.; Ding, Y.; Guo, J.; Gong, R.; Qin, H.; and Liu, X. 2025. DA-KD: Difficulty-Aware Knowledge Distillation for Efficient Large Language Models. In *Forty-first International Conference on Machine Learning*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations (ICLR)*.

J. Guo, W. Ouyang, and D. Xu. 2020. Channel pruning guided by classification loss and feature importance. In *AAAI*.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2021. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences*, 11(14): 6421.

Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W. W.; and Lu, X. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2567–2577.

Kim, Y.; Wu, J.; Abdulle, Y.; and Wu, H. 2024. MedExQA: Medical Question Answering Benchmark with Multiple Explanations. In *Workshop on Biomedical Natural Language Processing (BioNLP)*, 167–181.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J.; Zhang, H.; and Stoica, I. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, 611–626. ACM.

Lin, J.; Tang, J.; Tang, H.; Yang, S.; Dang, X.; Gan, C.; and Han, S. 2023. AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration. arXiv:2306.00978.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 3214–3252.

Liu, F.; Tang, Y.; Liu, Z.; Ni, Y.; Tang, D.; Han, K.; and Wang, Y. 2024a. Kangaroo: Lossless self-speculative decoding for accelerating llms via double early exiting. *Advances in Neural Information Processing Systems*, 37: 11946–11965.

Liu, J.; Li, J.; Wang, K.; Guo, H.; Yang, J.; Peng, J.; Xu, K.; Liu, X.; and Guo, J. 2024b. LTA-PCS: Learnable Task-Agnostic Point Cloud Sampling. In *CVPR*.

Lv, C.; Chen, H.; Guo, J.; Ding, Y.; and Liu, X. 2024. PTQ4SAM: Post-Training Quantization for Segment Anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15941–15951.

Men, X.; Xu, M.; Zhang, Q.; Wang, B.; Lin, H.; Lu, Y.; Han, X.; and Chen, W. 2024. ShortGPT: Layers in Large Language Models are More Redundant Than You Expect. *arXiv.org*, abs/2403.03853.

Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In *Conference on Health, Inference, and Learning (CHIL)*, 248–260.

Rang, M.; Bi, Z.; Zhou, H.; Chen, H.; Xiao, A.; Guo, T.; Han, K.; Chen, X.; and Wang, Y. 2025. Revealing the Power of Post-Training for Small Language Models via Knowledge Distillation. *arXiv preprint arXiv:2509.26497*.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; Payne, P.; Seneviratne, M.; Gamble, P.; Kelly, C.; Babiker, A.; Schärli, N.; Chowdhery, A.; Mansfield, P.; Demner-Fushman, D.; Arcas, B. A. y.; Webster, D.; Corrado, G. S.; Matias, Y.; Chou, K.; Gottweis, J.; Tomasev, N.; Liu, Y.; Rajkomar, A.; Barral, J.; Semturs, C.; Karthikesalingam, A.; and Natarajan, V. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Amin, M.; Hou, L.; Clark, K.; Pfohl, S. R.; Cole-Lewis, H.; Neal, D.; Rashid, Q. M.; Schaekermann, M.; Wang, A.; Dash, D.; Chen, J. H.; Shah, N. H.; Lachgar, S.; Mansfield, P. A.; Prakash, S.; Green, B.; Dominowska, E.; Arcas, B. A. y.; Tomašev, N.; Liu, Y.; Wong, R.; Semturs, C.; Mahdavi, S. S.; Barral, J. K.; Webster, D. R.; Corrado, G. S.; Matias, Y.; Azizi, S.; Karthikesalingam, A.; and Natarajan, V. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3): 943–950.

Sun, L.; Huang, Y.; Wang, H.; Wu, S.; Zhang, Q.; Gao, C.; Huang, Y.; Lyu, W.; Zhang, Y.; Li, X.; et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*.

Sun, M.; Liu, Z.; Bair, A.; and Kolter, J. Z. 2023. A Simple and Effective Pruning Approach for Large Language Models. *arXiv preprint arXiv:2306.11695*.

Wang, B.; Xu, C.; Wang, S.; Gan, Z.; Cheng, Y.; Gao, J.; Awadallah, A. H.; and Li, B. 2021. Adversarial GLUE: A Multi-Task Benchmark for Robustness Evaluation of Language Models. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Wang, G.; Gao, M.; Yang, S.; Zhang, Y.; He, L.; Huang, L.; Xiao, H.; Zhang, Y.; Li, W.; Chen, L.; Fei, J.; and Li, X. 2025a. Citrus: Leveraging Expert Cognitive Pathways in a Medical Language Model for Advanced Medical Decision Support. *arXiv.org*, abs/2502.18274.

Wang, J.; Zeng, Y.; Guo, J.; Ma, Y.; Liu, A.; and Liu, X. 2025b. SLMQuant: Benchmarking Small Language Model Quantization for Practical Deployment. In *3rd International Workshop on Rich Media With Generative AI*.

Wang, Z.; Guo, J.; Gong, R.; Yong, Y.; Liu, A.; Huang, Y.; Liu, J.; and Liu, X. 2024. PTSBench: A comprehensive post-training sparsity benchmark towards algorithms and models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5742–5751.

Xiao, G.; Lin, J.; Seznec, M.; Wu, H.; Demouth, J.; and Han, S. 2023. SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models. arXiv:2211.10438.

Yang, G.; He, C.; Guo, J.; Wu, J.; Ding, Y.; Liu, A.; Qin, H.; Ji, P.; and Liu, X. 2024a. LLMCBench: Benchmarking Large Language Model Compression for Efficient Deployment. *NeurIPS*.

Yang, Q. A.; Yang, B.; Zhang, B.; and etc. 2024b. Qwen2.5 Technical Report. *arXiv.org*, abs/2412.15115.

Zuo, Y.; Qu, S.; Li, Y.; Chen, Z.; Zhu, X.; Hua, E.; Zhang, K.; Ding, N.; and Zhou, B. 2025. MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding. *arXiv.org*, abs/2501.18362.