# UNICOMP: A Unified Evaluation of Large Language Model Compression via Pruning, Quantization, and Distillation

**Anonymous EMNLP submission**


## Abstract

Model compression is increasingly essential for deploying large language models (LLMs), yet existing comparative studies largely focus on pruning and quantization evaluated primarily on knowledge-centric benchmarks. Thus, we introduce UNICOMP, a unified evaluation framework for comparing pruning, quantization, and knowledge distillation. UNICOMP evaluates compressed models along three dimensions: *performance*, *reliability*, and *efficiency*, using a diverse set of capability- and safety-oriented benchmarks together with a hardware-aware efficiency analysis. Through evaluation of seven compression techniques across 40 datasets, we observe (i) a consistent *knowledge bias*, where factual recall is largely preserved while multi-step reasoning, multilingual, and instruction-following capabilities degrade; (ii) a decoupling between performance and reliability, indicating that retained performance does not consistently imply preserved reliability; and (iii) that task-specific calibration can yield up to 50% relative improvement in reasoning performance in pruned models.


## 1 Introduction

The rapid development of large language models (LLMs) has driven growing interest in compression techniques, whose goal is to effectively reduce memory usage and computational costs while preserving model performance (Zhu et al., 2024). Current mainstream approaches include pruning (Frantar and Alistarh, 2023; Sun et al., 2024), quantization (Frantar et al., 2023; Lin et al., 2024), and knowledge distillation (Jiao et al., 2020; Muralidharan et al., 2024).

Comprehensive and reliable evaluation is essential for deploying compressed models in real-world settings. However, much of the compression literature focuses on next-token prediction (typically evaluated via WikiText perplexity) and knowledge-centric benchmarks (i.e., factual recall and single-step reasoning tasks such as MMLU and ARC), resulting in evaluations that largely capture language modeling ability and factual recall (Frantar and Alistarh, 2023; Yang et al., 2025; Sreenivas et al., 2024). While recent efforts such as LLM-CBench (Yang et al., 2024) move toward more systematic evaluation, they omit knowledge distillation and remain focused on knowledge-centric multiple-choice benchmarks, with limited coverage

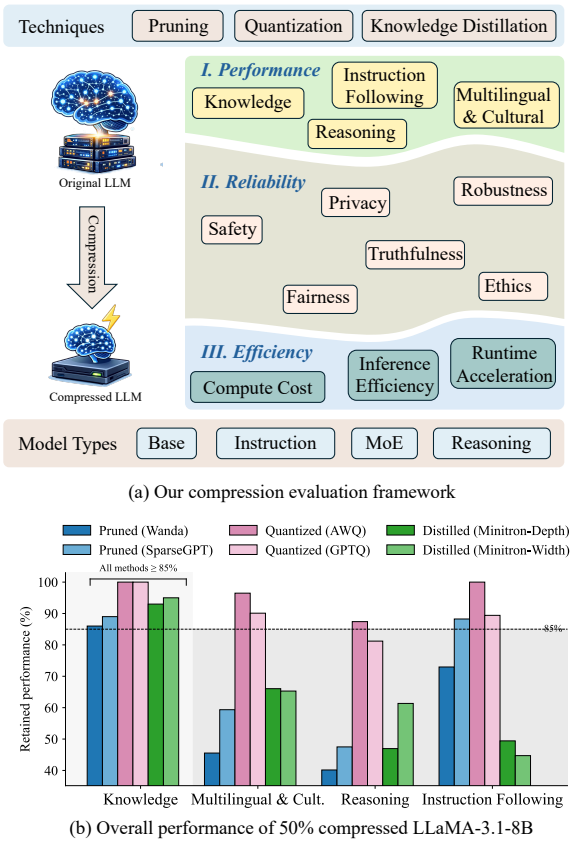



Figure 1: Overview of our compression evaluation framework and results: (a) UNICOMP covers performance, reliability, and efficiency with 13 metrics; and (b) Knowledge bias in LLM compression. Compression preserves knowledge performance but leads to pronounced degradation in multilingual and cultural, reasoning, and instruction following, with quantization as a partial exception.

of multi-step reasoning, multilingual performance, and reliability. As compressed LLMs are increasingly used in safety-critical and diverse settings, these limitations highlight the need for a more comprehensive, multi-dimensional evaluation framework. In particular, the impact of compression on reasoning, instruction following, multilingual performance, and reliability remains poorly characterized.

To address these limitations, we propose a UNIfied COMPression evaluation framework, UNICOMP, which systematically compares pruning, quantization, and knowledge distillation across contemporary LLMs and benchmarks (Figure 1a). UNICOMP evaluates models along three dimensions: performance, reliability, and efficiency. We cover representative pruning and quantization methods, including SparseGPT (Frantar and Alistarh, 2023), Wanda (Sun et al., 2024), GPTQ (Frantar et al., 2023), and AWQ (Lin et al., 2024). We study knowledge distillation explicitly as a compression technique rather than a general-purpose transfer paradigm, including NVIDIA's Minitron pipeline (Sreenivas et al., 2024) and Low-Rank Clone (Hao et al., 2025).

We make **three main contributions**: (1) We introduce UNICOMP, a unified evaluation framework for systematically comparing pruning, quantization, and knowledge distillation across performance, reliability, and efficiency dimensions; (2) We conduct extensive experiments on over 40 datasets covering knowledge, reasoning, multilinguality, instruction following, and a wide range of reliability and safety benchmarks, evaluated on modern LLMs including recent chain-of-thought (CoT) and mixture-of-experts (MoE) models and (3) we analyze the role of calibration data in compression and show that reasoning-aware calibration can result in up to 50% relative improvement in reasoning performance in pruned models.

## 2 Related Work

**LLM Compression.** To reduce the computational cost of LLMs, prior work primarily explores *pruning*, *quantization*, and *knowledge distillation* (KD) (Ashkboos et al., 2024; Du et al., 2024; Hinton et al., 2015). **Pruning** removes redundant parameters using structured, semi-structured, or unstructured strategies (Ashkboos et al., 2024; Yang et al., 2025). While structured one-shot pruning enables hardware acceleration, it incurs significant performance degradation (Yang et al., 2024). Consequently, unstructured and semi-structured pruning exemplified by SparseGPT and Wanda has become the most widespread approach (Frantar and Alistarh, 2023; Sun et al., 2024). **Quantization** improves efficiency by reducing numerical precision of parameters rather than removing them (Jacob et al., 2018). Weight-only post-training quantization methods, such as GPTQ (Frantar et al., 2023) and AWQ (Lin et al., 2024), are particularly prevalent, offering substantial memory savings with minimal accuracy loss while avoiding the runtime complexity of activation or KV-cache quantization (Zhu et al., 2024). **Knowledge distillation** transfers knowledge from a large teacher model to a smaller student by minimizing divergence between output distributions or intermediate representations, often leveraging synthetic data generated by the teacher (Jiao et al., 2020; Gu et al., 2024). In compression settings where pretraining of a student model is too computationally expensive, current work follows a *prune→distill* pipeline, where a compact student is first obtained through hard structured or soft pruning (Sreenivas et al., 2024; Hao et al., 2025) and then trained using the original model as the teacher.

**Benchmarks and evaluation protocols.** There have been sustained efforts to benchmark and compare compression techniques for LLMs. LLM-CBENCH (Yang et al., 2024) provides a large-scale comparison of pruning and quantization methods, showing that quantization generally outperforms pruning on knowledge-based multiple-choice benchmarks. With the emergence of CoT reasoning models, subsequent work has examined the impact of quantization on reasoning performance, reporting substantial degradation across different quantization schemes (Liu et al., 2025). Other studies analyze how compression affects agentic capabilities (Dong et al., 2025) or investigate reasoning robustness under pruning and quantization in distilled DeepSeek-R1 models (Zhang et al., 2026; DeepSeek-AI et al., 2025). In parallel, recent work explored the role of calibration data in pruning and quantization, including its impact on general natural language understanding (NLU) (Williams and Aletras, 2024) and multilingual generalization (Zeng et al., 2024) finding that calibration data can influence NLU and data adjustment can yield better multilingual performance.

## 3 UNICOMP Framework

We present UNICOMP, a unified framework for evaluating compression paradigms along performance, reliability, and efficiency dimensions using 13 metrics. Performance and reliability scores are normalized relative to the corresponding base model, where 100 denotes full retention. Efficiency scores are scaled to $[0, 100]$ relative to the best-performing compression method.

### 3.1 Performance

We assess the extent to which compressed language models retain core task-solving capabilities.

**Knowledge.** Knowledge is a fundamental component of intelligent behavior; its degradation often leads to hallucinated or semantically invalid outputs (Prato et al., 2024). We evaluate knowledge retention using standard multiple-choice benchmarks, following (Yang et al., 2024). For each compressed model, we compute the unweighted average ratio of task accuracy relative to the base model:

$$\mathcal{S}_{\text{K}} = \frac{100}{N} \sum_{i=1}^{N} \frac{s^{K}_{\text{Comp}_i}}{s^{K}_{\text{Base}_i}}, \tag{1}$$

where $s^{K}$ denotes exact-match accuracy on knowledge task $i$, and $N$ is the number of benchmarks.

**Multilingual and Cultural Generalization.** Multilingual and cultural generalization ($\mathcal{S}_{\text{Mul}}$) captures multilingual understanding, cultural awareness and bias detection, representing linguistic robustness and sensitivity to socially grounded phenomena. Both are evaluated via multiple-choice QA and equally weighted:

$$\mathcal{S}_{\text{Mul}} = \frac{100}{2} \left( \frac{s^{\text{Lan}}_{\text{Comp}}}{s^{\text{Lan}}_{\text{Base}}} + \frac{s^{\text{Bias}}_{\text{Comp}}}{s^{\text{Bias}}_{\text{Base}}} \right), \tag{2}$$

where $s^{\text{Lan}}$ denotes multilingual and cultural understanding, and $s^{\text{Bias}}$ denotes bias detection scores.

**Reasoning.** CoT reasoning is a core capability of modern LLMs (Wei et al., 2022; DeepSeek-AI et al., 2025). We evaluate reasoning retention on a set of challenging benchmarks and compute a normalized score analogous to knowledge:

$$\mathcal{S}_{\text{R}} = \frac{100}{N} \sum_{i=1}^{N} \frac{s^{\text{R}}_{\text{Comp}_i}}{s^{\text{R}}_{\text{Base}_i}}, \tag{3}$$

where $s^{\text{R}}$ denotes exact-match accuracy on reasoning task $i$. All models use the standard number of few-shot exemplars for CoT prompting as adopted in prior work, with greedy decoding.

**Instruction Following.** Instruction following measures a model's ability to accurately execute user directives, which is critical for agentic applications (Qi et al., 2025). As only a single benchmark is used, we define:

$$\mathcal{S}_{\text{IF}} = 100 \cdot \frac{s^{\text{IF}}_{\text{Comp}}}{s^{\text{IF}}_{\text{Base}}}, \tag{4}$$

where $s^{\text{IF}}$ is task accuracy or constraint satisfaction.

### 3.2 Reliability

Following TRUSTLLM (Huang et al., 2024b), we evaluate reliability across six dimensions: truthfulness, safety, fairness, robustness, privacy, and ethics. We adopt the dimension definitions and evaluation protocols directly from TRUSTLLM, which provides detailed descriptions of the following subdimensions and datasets (see Appendix C.1, Table 14).

***Scoring Protocol.*** *All scores here are computed following Eq. (1), with metric-specific substitutions of component scores. For benchmarks, where lower scores reflect better performance, we use* $(100 - s)$ *for average computation.*

**Truthfulness.** $\mathcal{S}_{\text{Truth}}$ evaluates factual reliability, characterized by internal and external misinformation, hallucination resistance and sycophancy avoidance. It is computed using $s^{Int}$, $s^{Ext}$, $s^{Hal}$ and $s^{Syc}$ .

**Safety.** $\mathcal{S}_{\text{SAFE}}$ measures the ability of a model to avoid harmful or disallowed outputs under adversarial prompts, captured by jailbreak robustness $s^{Jail}$, misuse refusal $s^{Mis}$ and exaggerated safety understanding $s^{Exagg}$.

**Fairness.** $\mathcal{S}_{\text{FAIR}}$ assesses whether compression preserves bias-related behavior under controlled demographic perturbations, measured by stereotype recognition and disagreement score $s^{stereo}$, disparagement $s^{disp}$, and preference bias $s^{pref}$.

**Robustness.** $\mathcal{S}_{\text{ROB}}$ captures resilience to semantic-preserving perturbations and distributional shifts, measured by robustness to natural noise $s^{Noise}$ and Out-of-Distribution data $s^{OOD}$.

**Privacy.** $\mathcal{S}_{\text{PRI}}$ evaluates protection against unintended disclosure of sensitive information. It combines *privacy awareness* $s^{PA}$, measuring appropriate handling of privacy-sensitive requests, and

*leakage* $s^{Leak}$, assessing exposure of private data under targeted prompts.

**Ethics.** $\mathcal{S}_{\text{ETH}}$ evaluates norms in value-sensitive scenarios across three dimensions: *implicit ethics* $s^{impl}$, capturing moral judgments; *explicit ethics* $s^{expl}$, evaluating morally appropriate actions; and *awareness* $s^{aware}$, reflecting understanding of the model's role, capabilities, and social context.

### 3.3 Efficiency

Lastly, we evaluate the practical utility of compressed models along an efficiency dimension. Unlike performance and reliability metrics, which are normalized with respect to the base model, efficiency metrics are inherently comparative across compression methods. We therefore normalize each efficiency metric relative to the best-performing method, where optimality corresponds to either a maximum (e.g., throughput) or a minimum (e.g., latency).

Let $\mathcal{M}$ denote the set of compression methods under comparison. For a metric value $x_m$ associated with method $m \in \mathcal{M}$, we define the normalized efficiency score as:

$$\tilde{x}_m = \begin{cases} \dfrac{x_m}{\max_{k\in\mathcal{M}} x_k} & \text{if higher is better,} \\ \dfrac{\min_{k\in\mathcal{M}} x_k}{x_m} & \text{if lower is better,} \end{cases} \tag{5}$$

which assigns the best method a value of 1.

**Runtime Acceleration.** Runtime acceleration measures raw execution speed during inference. We capture this dimension using throughput $T_m$ and latency $L_m$, and define:

$$\mathcal{S}_{\text{RA}} = 100 \cdot \left(\tilde{T}_m \cdot \tilde{L}_m\right)^{\frac{1}{2}}. \tag{6}$$

**Inference Efficiency.** Inference efficiency reflects the resource footprint of a model at deployment time. We consider inference GPU memory usage $M_m$, model size on disk $S_m$, and theoretical FLOPs $F_m$, and define:

$$\mathcal{S}_{\text{IE}} = 100 \cdot \left(\tilde{M}_m \cdot \tilde{S}_m \cdot \tilde{F}_m\right)^{\frac{1}{3}}. \tag{7}$$

**Compute Cost.** Compute cost captures the cost of producing the compressed model. We measure this dimension using total compression time $T_m^{\text{Comp}}$ and peak GPU memory usage $M_m^{\text{Comp}}$, and define:

$$\mathcal{S}_{\text{CC}} = 100 \cdot \left(\tilde{T}_m^{\text{Comp}} \cdot \tilde{M}_m^{\text{Comp}}\right)^{\frac{1}{2}}. \tag{8}$$

All efficiency scores are computed using the geometric mean to penalize bottlenecks and prevent strong performance along a single axis from masking deficiencies in others. Higher values indicate better practical efficiency.

## 4 Experiments

### 4.1 Experimental Setup

**Models.** The main experiments focus on LLaMA-3.1-8B and Qwen-2.5-7B. To further test generalization, we evaluate on a broader range of architectures, including *traditional* model families such as the LLaMA-2 (7B, 13B, 70B), LLaMA-3.1 (8B, 70B), *reasoning* models such as the Qwen-3 (0.6B, 1.7B, 4B, 8B, 14B, 32B) and DeepSeek-R1 (Distill-LLaMA-8B, Distill-LLaMA-70B) and *MoE* models such as Qwen-3-30B-A3. Unless otherwise stated, we use only instruction-tuned models.

**Datasets.** For *performance* evaluation, we consider benchmarks spanning knowledge, reasoning, multilingual understanding, and instruction following. Academic knowledge and factual reasoning are assessed using MMLU (Hendrycks et al., 2021b) and ARC-E/C (Clark et al., 2018). Commonsense understanding is measured with HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2019), and Winogrande (Sakaguchi et al., 2019). Advanced reasoning is evaluated using GSM8K (Cobbe et al., 2021) (4-shot), MATH-500 (Hendrycks et al., 2021c) (4-shot), and GPQA-Diamond (Rein et al., 2024) (5-shot). Multilingual capability is measured using Global-MMLU-Lite (Singh et al., 2025), with BBQ (Parrish et al., 2022) included to assess social bias. Instruction-following performance is evaluated using IFBench (Pyatkin et al., 2025). For *reliability*, we follow TrustLLM (Huang et al., 2024b), which substantially broadens evaluation beyond the limited coverage of TruthfulQA (Lin et al., 2022) and AdversarialGLUE (Wang et al., 2022) previously used by LLMCBench (Yang et al., 2024). Thus we assess compressed models reliability on 28 datasets, including ConfAIde (Mireshghallah et al., 2024), MoralChoice (Scherrer et al., 2023), HaluEval (Li et al., 2023), StereoSet (Nadeem et al., 2021), and Do-Not-Answer (Wang et al., 2024). Please refer to Appendix C.1 for more details.

**Compression Techniques** For both pruning and quantization, we include the best performing and most-cited compression techniques. For pruning,

we employ SparseGPT (Frantar and Alistarh, 2023) and Wanda (Sun et al., 2024), and remove 50% of the model's parameters through unstructured and semi-structured pruning. For quantization, we evaluate the standard configurations most commonly used in prior work and practice: weight-only 4-bit quantization (W4A16) using GPTQ (Frantar et al., 2023) and AWQ (Lin et al., 2024), as well as weight-activation 8-bit quantization (W8A8) using SmoothQuant (Xiao et al., 2023). To ensure fair and meaningful comparison, we treat knowledge distillation (KD) strictly as compression, evaluating only methods that derive the student directly from the teacher without a separately pretrained model. We evaluate two representative approaches in this setting: Minitron (Sreenivas et al., 2024) (using the open-source LLaMA-3.1-Minitron-4B 50% depth and width pruned variants) and Low-Rank Clone (Hao et al., 2025) (using the open-source LRC-4B model distilled from Qwen-2.5-7B). This design choice ensures that all compression paradigms are compared under a consistent experimental setting.

**Implementation** We implemented UNICOMP in PyTorch and conducted our experiments on H100 GPUs. We evaluate performance benchmarks using Lighteval (Habib et al., 2023) and lm-evaluation-harness (Gao et al., 2024) with vLLM (Kwon et al., 2023) as backend. The reliability benchmarks are evaluated using GPT-4o (Hurst et al., 2024) and a Longformer-based harmful-response classifier[1] (Beltagy et al., 2020) as judges, following Huang et al. (2024b). Given pretrained base models, we apply each compression method with default hyperparameters using open-source implementations and vLLM's `llm-compressor` framework (AI and vLLM Project, 2024).

## 5 Results

### 5.1 Performance

**Knowledge Bias.** Across all evaluated compression paradigms, performance is most robust on multiple-choice knowledge benchmarks. As shown in Figure 1 (b) and Table 1, even aggressive pruning and distillation preserve a large fraction of factual knowledge, suggesting that static knowledge representations are comparatively resilient to compression. This observation aligns with the prevalent focus in the compression literature on multiple-choice knowledge benchmarks, which mostly report strong robustness under compression. However, our results reveal stark differences once evaluation extends beyond knowledge: pruned and distilled models retain high knowledge scores but degrade sharply on multilingual performance, reasoning, and instruction following. For example, on LLaMA-3.1-8B, Wanda pruning drops from 83.66% knowledge retention to 40.15% reasoning retention, while Minitron-Depth drops from 92.25% to 46.96%. We further verify this knowledge bias pattern in Llama-3.2-3B and Qwen-2.5-3B, with consistent results reported in Appendix B.1.

**Reasoning Sensitivity.** Reasoning-centric benchmarks exhibit the sharpest degradation under compression, indicating that multi-step reasoning is substantially more sensitive to parameter reduction than factual recall or surface-level understanding. This is directly reflected in Table 1: for LLaMA-3.1-8B, Wanda at 50% drops from 76.80 to 19.48 on GSM8K and from 30.20 to 7.60 on MATH-500. Even quantized models are most vulnerable on reasoning, with the clearest drops again appearing on MATH-500. Interestingly, Minitron-Depth, obtained by removing the latter half of the transformer layers, performs substantially worse than its Width counterpart despite undergoing the same distillation procedure, highlighting the importance of later layers for reasoning in LLMs. The comparatively stronger robustness on GPQA-Diamond relative to GSM8K and MATH-500 is likely attributable to its multiple-choice format.

**Quantization Dominates.** Consistent with prior findings (Yang et al., 2024), quantization techniques consistently outperform other compression methods across capabilities. Table 1 shows that all three quantization methods remain close to the dense baseline on knowledge benchmarks, achieving almost full knowledge retention on both models ($\mathcal{S}_K > 97$). However, quantization is not a full drop-in replacement for dense models: even the strongest quantized models show reduced performance on challenging reasoning and multilingual tasks, suggesting that small quantization errors may accumulate in longer reasoning chains. Within quantization, weight-only methods (AWQ, GPTQ) and weight-activation quantization (SmoothQuant) show broadly comparable performance despite different precision settings.

[1] `LibrAI/longformer-harmful-ro`

| Category | Method | Ratio | Knowledge | | | | Multilingual & Cultural | | | Reasoning | | | | InstFollowing | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | MMLU | ARC-c | Hellaswag | $\mathcal{S}_{\text{K}}$ | G-MMLU | BBQ | $\mathcal{S}_{\text{Mul}}$ | GSM8K | MATH-500 | GPQA-D | $\mathcal{S}_{\text{R}}$ | IFBench | $\mathcal{S}_{\text{IF}}$ |
| *LLAMA-3.1-8B* | | | | | | | | | | | | | | | |
| Baseline | | 0% | 61.38 | 53.50 | 79.12 | – | 56.00 | 79.15 | – | 76.80 | 30.20 | 33.33 | – | 28.33 | – |
| Pruning | Wanda | 50% | 40.59 | 44.97 | 68.23 | 83.66 | 40.68 | 50.47 | 68.20 | 19.48 | 7.60 | 23.30 | 40.15 | 20.67 | 72.96 |
| | | 2:4 | 27.57 | 28.84 | 47.86 | 61.12 | 31.45 | 37.33 | 51.66 | 4.40 | 3.80 | 24.57 | 30.68 | 12.33 | 43.52 |
| | SparseGPT | 50% | 48.33 | 42.15 | 71.66 | 87.95 | 46.25 | 51.03 | 73.53 | 36.92 | 8.40 | 22.20 | 47.50 | 25.00 | 88.25 |
| | | 2:4 | 28.27 | 33.87 | 56.02 | 67.66 | 35.82 | 43.33 | 59.35 | 9.33 | 2.40 | 23.23 | 29.93 | 14.67 | 51.78 |
| Quantization | AWQ | W4A16 | 61.22 | 53.22 | **79.15** | 99.78 | **54.20** | **76.12** | **96.48** | **73.31** | 22.00 | **31.31** | **87.41** | **28.33** | **100.00** |
| | GPTQ | W4A16 | **61.36** | **53.41** | 79.06 | **99.92** | 49.33 | 72.91 | 90.10 | 71.57 | 19.80 | 28.28 | 81.20 | 25.33 | 89.41 |
| | SmoothQuant | W8A8 | 58.80 | 52.76 | 75.36 | 97.19 | 51.49 | 56.81 | 81.86 | 73.11 | **24.04** | 29.29 | **87.56** | 27.33 | 96.47 |
| Distillation | Minitron-Depth | 50 % | 60.87 | 45.65 | 69.47 | 92.25 | 42.82 | 44.04 | 66.05 | 29.95 | 5.60 | 27.78 | 46.96 | 14.00 | 49.42 |
| | Minitron-Width | 50 % | 58.00 | 49.23 | 73.96 | 95.12 | 42.53 | 43.24 | 65.29 | 51.63 | 12.40 | 25.25 | 61.35 | 12.67 | 44.72 |
| *Qwen-2.5-7B* | | | | | | | | | | | | | | | |
| Baseline | | 0% | 71.76 | 55.29 | 80.40 | – | 60.53 | 83.12 | – | 86.66 | 75.60 | 32.83 | – | 30.00 | – |
| Pruning | Wanda | 50% | 67.00 | 45.31 | 73.69 | 90.36 | 54.62 | 79.36 | 92.86 | 75.82 | 42.60 | 30.81 | 79.23 | 25.33 | 84.43 |
| | | 2:4 | 54.29 | 44.71 | 62.97 | 82.81 | 40.48 | 49.48 | 63.20 | 39.42 | 9.20 | 30.81 | 50.50 | 21.33 | 71.10 |
| | SparseGPT | 50% | 66.65 | 49.06 | 75.35 | 93.44 | 54.72 | 81.34 | 94.13 | 74.91 | 43.20 | 28.20 | 76.49 | 20.67 | 68.90 |
| | | 2:4 | 55.60 | 44.62 | 67.18 | 85.18 | 44.63 | 45.66 | 64.33 | 41.93 | 6.80 | 27.27 | 46.81 | 22.00 | 73.33 |
| Quantization | AWQ | W4A16 | 70.75 | 54.35 | 79.55 | 98.98 | 59.05 | **84.57** | **99.65** | **86.58** | 72.20 | 29.80 | 95.39 | **30.00** | **100.00** |
| | GPTQ | W4A16 | 70.75 | 54.35 | 79.72 | 99.11 | **59.22** | 78.43 | 96.10 | **86.58** | 72.60 | **33.30** | **99.12** | 27.33 | 91.10 |
| | SmoothQuant | W8A8 | **71.14** | **55.55** | **79.87** | **99.51** | 58.85 | 72.57 | 92.27 | 85.76 | **74.03** | 31.35 | 97.46 | 28.67 | 95.57 |
| Distillation | Low-Rank Clone | 43% | 64.52 | 52.13 | 70.70 | 91.94 | 53.43 | 83.05 | 94.09 | 64.29 | 15.80 | 25.25 | 57.33 | 27.33 | 91.10 |

Table 1: Performance comparison on two representative models. Bold and underlined scores indicate the best and second-best results within each model group. GPQA-D denotes GPQA-Diamond, and G-MMLU reports average accuracy on Global-MMLU-Lite across 14 languages. While only three knowledge benchmarks are shown, the aggregated knowledge score $\mathcal{S}_K$ additionally includes ARC-e, PIQA, and Winogrande (see Appendix B.2).

**Low-Rank Clone Outperforms Minitron.** The LRC-4B model, obtained via the Low-Rank Clone algorithm that combines soft pruning with distillation, substantially outperforms its Minitron counterpart in multilingual performance and instruction following, despite being distilled on an order of magnitude fewer tokens (20B vs. 200B). This result underscores the advantages of *soft pruning* relative to *hard pruning* when constructing student models. However, LRC-4B remains limited in advanced reasoning, most notably on MATH-500 dropping from 75.6 to 15.8.

**Multilingual Performance under Compression.** Evaluating compression on Global-MMLU-Lite reveals important differences that are not apparent from English-centric setups. While English knowledge performance remains relatively robust, multilingual performance drops substantially for pruning and distillation methods, exposing the same knowledge bias. Within quantization, AWQ performs best overall in multilingual and cultural settings. We further analyze performance differences between high-resource and low-resource languages, but do not observe a consistent additional degradation on low-resource languages across compression paradigms (see Appendix B.2). These results show that compression methods with similar English performance can diverge substantially across languages and culturally grounded tasks, underscoring the importance of multidimensional evaluation.

## 5.2 Reliability

**Performance–Reliability Decoupling.** Table 2 shows that reliability preservation does not closely track task-performance preservation under compression. Methods that are strong in retained performance, most notably quantization, do not uniformly dominate across reliability dimensions. Conversely, several compressed models obtain scores above 100 in specific dimensions, indicating improvements over the uncompressed baseline despite reduced task performance. This suggests that compression can alter behavioral properties in ways that are not captured by standard capability benchmarks, motivating reliability evaluation as a distinct objective rather than a proxy of performance retention.

**Reliability Shifts Are Dimension- and Model-Specific.** Reliability varies substantially across compression methods and evaluation dimensions, as shown in Table 2. While truthfulness and safety are generally robust to compression, often matching or exceeding baseline scores, fairness and robustness exhibit much higher variance, indicating sensitivity to socially grounded behavior and adversarial resilience. In contrast with performance results, no single compression paradigm consistently dominates. Quantization performs well on robustness and ethics for both LLaMA-3.1-8B and Qwen-2.5-7B, but shows mixed results on fairness

| Method | Ratio / # Bits | Truthfulness $\mathcal{S}_{\text{TRU}}$ | Safety $\mathcal{S}_{\text{SAFE}}$ | Fairness $\mathcal{S}_{\text{FAIR}}$ | Robustness $\mathcal{S}_{\text{ROB}}$ | Privacy $\mathcal{S}_{\text{PRI}}$ | Ethics $\mathcal{S}_{\text{ETH}}$ |
|---|---|---|---|---|---|---|---|
| *LLAMA-3.1-8B* | | | | | | | |
| Wanda | 50% | 88.71 | 100.67 | 85.03 | 84.27 | 94.72 | 89.09 |
| | 2:4 | 55.70 | 95.70 | **93.53** | 73.37 | 86.85 | 72.02 |
| SparseGPT | 50% | 92.06 | 100.93 | 70.85 | 92.95 | 81.97 | 96.31 |
| | 2:4 | 84.35 | 99.06 | 90.34 | 82.30 | 94.77 | 82.48 |
| AWQ | W4A16 | 90.88 | 98.03 | 67.51 | **98.85** | **96.09** | **100.33** |
| GPTQ | W4A16 | **92.95** | 84.42 | 77.45 | 92.29 | 92.23 | 99.32 |
| SmoothQuant | W8A8 | 90.78 | **101.90** | 69.96 | 98.17 | 88.39 | 96.00 |
| Minitron-Depth | 50% | 71.45 | 86.36 | 69.33 | 83.77 | 61.97 | 81.73 |
| Minitron-Width | 50% | 54.34 | 84.18 | 75.16 | 79.08 | 69.85 | 86.19 |
| *Qwen-2.5-7B* | | | | | | | |
| Wanda | 50% | 93.59 | 76.67 | 58.26 | 96.55 | **107.94** | 97.96 |
| | 2:4 | 87.15 | 89.28 | 75.11 | 97.38 | 105.58 | 92.33 |
| SparseGPT | 50% | **96.58** | 78.32 | 67.14 | 96.21 | 104.85 | 100.16 |
| | 2:4 | 90.76 | 92.29 | 57.58 | 96.56 | 98.07 | 95.19 |
| AWQ | W4A16 | 92.82 | 93.13 | 102.63 | 101.57 | 102.42 | **100.20** |
| GPTQ | W4A16 | 94.36 | 86.82 | 71.97 | 102.86 | 93.78 | 99.77 |
| SmoothQuant | W8A8 | 94.96 | **96.33** | 67.00 | **104.12** | 103.08 | 100.18 |
| Low-Rank Clone | 43% | 96.16 | 71.86 | **124.19** | 101.54 | 94.66 | 94.56 |

Table 2: Reliability scores of various compression methods aggregated across 28 datasets for six dimensions. Detailed evaluation metrics and per-dataset breakdowns are provided in Appendix C.

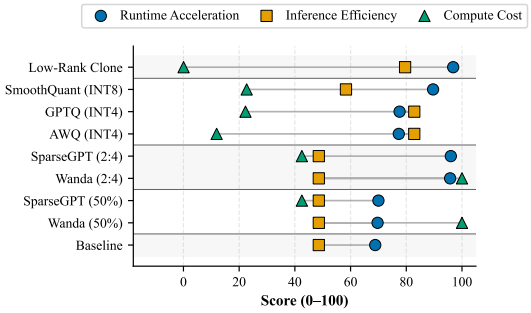



Figure 2: Efficiency results for compression of Qwen-2.5-7B in three dimensions: Runtime Acceleration $\mathcal{S}_{\text{RA}}$, Inference Efficiency $\mathcal{S}_{\text{IE}}$ and Compute Cost $\mathcal{S}_{\text{CC}}$.

and safety, especially for Qwen. Pruning outcomes are strongly model-dependent, with relatively preserved fairness and robustness on LLaMA but notable degradation on Qwen.

**Distillation Preserves Reliability Less Predictably.** Distillation-based compression preserves reliability less consistently than quantization. On LLaMA-3.1-8B, both Minitron variants underperform across most dimensions, indicating that distillation after aggressive structured reduction does not reliably recover the base model's behavior. Low-Rank Clone on Qwen-2.5-7B is more selective, achieving the strongest fairness score but remaining weak on safety, showing that distillation effects are highly dimension-specific.

### 5.3 Efficiency

**Distillation Maximizes Efficiency at High Cost.** Knowledge-distilled models achieve the highest scores in both runtime acceleration and inference efficiency, reflecting substantial gains in throughput, latency, and deployment footprint, visualized in Figure 2. However, these benefits come at a prohibitive compute cost, with significantly longer training times and higher resource consumption, as shown in Table 3. As a result, distillation is most suitable when offline compute is abundant and compute cost is not a primary concern.

**Quantization Provides the Best Overall Trade-off.** Quantization methods offer the most balanced efficiency profile, delivering strong runtime and memory improvements while maintaining relatively low compression overhead. Weight-only methods (AWQ, GPTQ) and weight-activation quantization (SmoothQuant) trade off runtime acceleration with inference efficiency, visible in Figure 2. Although they do not match the peak efficiency of distilled models, their favorable trade-off between deployment efficiency and compute cost makes them the most practical choice for real-world and resource-constrained settings.

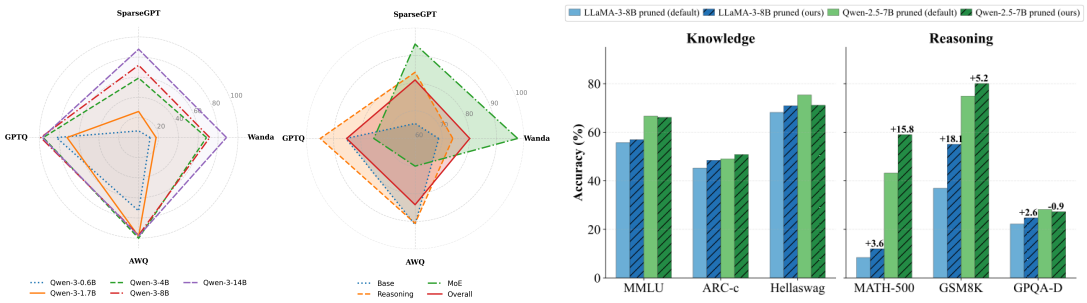
**Pruning Benefits Are Limited to Specialized Settings.** Pruning methods yield meaningful runtime acceleration only under semi-structured sparsity (2:4), where hardware support can be effectively leveraged. In contrast, unstructured pruning shows weaker gains in runtime and inference efficiency despite low compute cost. Among pruning approaches, Wanda stands out for its minimal compute cost, consistent with findings of Yang et al. (2024).

### 5.4 More Analysis

**Comparison Across Models.** To examine how compression interacts with model scale and architecture, we evaluate models ranging from 0.6B to 14B parameters, including CoT reasoning models and MoE architectures. Due to resource constraints, we focus on pruning and quantization. Figure 3 (a) reports relative retained performance on GSM8K for CoT models, while Figure 3 (b) reports retained WikiText-2 perplexity across model sizes and architectures. Across scales, larger models generally exhibit greater robustness to compression, consistent with Yang et al. (2024). However, smaller models remain highly sensitive in reasoning tasks: pruning reduces retained performance to as low as 6–25% for 0.6B and 1.7B models, while quantization can still incur losses of up to 30%, indicating that quantized models are not reliable drop-in replacements at smaller scales. Across architectures,

| Category | Method | Ratio / # Bits | Runtime Acceleration | | | Inference Efficiency | | | | Compute Cost | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Throughput | Latency | $\mathcal{S}_{\text{RA}}$ | GPU Mem | Model Size | FLOPs | $\mathcal{S}_{\text{IE}}$ | Time | GPU Mem | $\mathcal{S}_{\text{CC}}$ |
| Baseline | | 0% | 41,498 t/s | 955ms | – | 20.39GB | 14.96GB | 1.92T | – | – | – | – |
| Pruning | Wanda | 50% | 42,337 t/s | 938ms | 56.09 | 20.39GB | 14.96GB | 1.92T | 46.4 | **41s** | **16GB** | **100** |
| | | 2:4 | 71,194 t/s | 588ms | 91.87 | 20.39GB | 14.96GB | 1.92T | 46.4 | 41s | 16GB | 100 |
| | SparseGPT | 50% | 42,440 t/s | 949ms | 55.84 | 20.39GB | 14.96GB | 1.92T | 46.4 | 4.25m | 16GB | 40.1 |
| | | 2:4 | 72,531 t/s | 590ms | 92.58 | 20.39GB | 14.96GB | 1.92T | 46.4 | 4.25m | 16GB | 40.1 |
| Quantization | AWQ | W4A16 | 38,150 t/s | 661ms | 63.43 | **10.53GB** | **5.40GB** | 1.92T | **81.2** | 16m32s | 42.1GB | 12.53 |
| | GPTQ | W4A16 | 41,562 t/s | 730ms | 63.00 | 10.53GB | 5.40GB | 1.92T | 81.2 | 10m | **16GB** | 26.14 |
| | SmoothQuant | W8A8 | 55,109 t/s | 717ms | 73.20 | 20.30GB | 8.50GB | 1.92T | 56.1 | 13m33s | **15.12GB** | 23.10 |
| Distillation | Minitron-Depth | 50% | **78,606 t/s** | **548ms** | **100.0** | 13.73GB | 8.46GB | **1.03T** | 78.7 | 140h* | 20,480GB | 0.01 |
| | Minitron-Width | 50% | 68,595 t/s | 664ms | 84.86 | 13.67GB | 8.50GB | 1.05T | 78.2 | 120h* | 20,480GB | 0.03 |

Table 3: Efficiency comparison on LLaMA-3.1-8B-Instruct. Method-invariant values and ties are not highlighted. FLOPs denote theoretical floating-point operations for one forward pass. $*$ denotes that the training time is estimated based on the description provided in Sreenivas et al. (2024), refer to Appendix A.4.




(a) Different Sizes (b) Different Types (c) Calibration Comparison

Figure 3: More analysis: (a) five different reasoning model sizes; (b) three different model types; refer to Appendix B.2 for detailed results; and (c) effects of reasoning-aware calibration data in pruned models.

pruning favors MoE models but struggles on reasoning, whereas quantization shows the opposite trend, likely because current quantization methods are not well adapted to MoE architectures. Detailed results are provided in Appendix B.2.

**Impact of Calibration.** While not a compression method itself, calibration data determines which weights are preserved during pruning and quantization, offering a potential explanation for the observed knowledge bias. We find that the default C4-based calibration data used by SparseGPT and Wanda is mismatched with reasoning tasks, and therefore construct a reasoning-centric calibration set from MATH, GSM8K, and ARC-c. As shown in Figure 3 (c), pruned Qwen-2.5-7B and LLaMA-3.1-8B calibrated on this data consistently outperform C4-calibrated models while still preserving performance on knowledge benchmarks; for example, GSM8K accuracy on pruned LLaMA-3.1-8B improves from 36.9% to 55% (+50% relative). This result directly supports the presence of a knowledge bias in compression, showing that standard calibration disproportionately favors knowledge retention over reasoning. We further examine whether the same reasoning-aware calibration strategy benefits quantized models, but observe no significant improvements over standard calibration. More details are provided in Appendix B.3.

## 6 Conclusion

We introduced UniComp, a unified framework for evaluating LLM compression across performance, reliability, and efficiency. Our results reveal a persistent *knowledge bias*: factual recall remains the most robust, while multi-step reasoning, multilingual and instruction-following performance degrade substantially. This bias is also present in calibration data used for pruning: reasoning-aware calibration yields substantial gains for pruned models while preserving knowledge recall. Among methods, quantization offers the best overall trade-off, but is not a drop-in replacement for dense models, showing weaknesses on challenging reasoning tasks and multilingual performance, while distillation delivers strong runtime gains at substantially higher computational cost. Finally, we observe a performance-reliability decoupling, where retained performance does not imply retained reliability. Together, these findings underscore the importance of multi-dimensional evaluation of compressed LLMs.

## 7 Limitations & Ethics

Despite our efforts to include a broad range of models, datasets, and compression paradigms, the scope of our evaluation remains limited. The main evaluation focuses on only two representative models: LLaMA-3.1-8B and Qwen-2.5-7B, introducing bias to our comprehensive evaluation. In addition, although UNICOMP comprises thirteen metrics spanning performance, reliability, and efficiency, several important domains are not yet covered, such as code generation, multi-agent collaboration, and other specialized capabilities. Moreover, our calibration data ablations only cover reasoning ability via reasoning-centric datasets.

Expanding the benchmark to incorporate a wider spectrum of model families, datasets, and task domains is an interesting direction for future work. Nevertheless, we aim for UNICOMP to provide a comprehensive and fair assessment of LLM behavior under compression, offering empirical insights into how different compression methods affect diverse model capabilities and highlighting open challenges for future research.

**LLMs Usage.** Through the paper, we use LLMs to assist with grammar checking and minor rephrasing for clarity. LLMs did not contribute to the conceptual design of the study, experimental implementation, or core writing of the paper.

## References


Red Hat AI and vLLM Project. 2024. LLM Compressor.

Saleh Ashkboos, Maximilian L. Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. 2024. SliceGPT: Compress large language models by deleting rows and columns. In *The Twelfth International Conference on Learning Representations*.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8.

Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. In *AAAI Conference on Artificial Intelligence*.

Carnegie Mellon University. 2015. Enron email dataset. https://www.cs.cmu.edu/~enron/. Accessed via Carnegie Mellon University.

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. CODAH: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69, Minneapolis, USA. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

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.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, and Zhuoshu Li et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Peijie Dong, Zhenheng Tang, Xiang Liu, Lujun Li, Xiaowen Chu, and Bo Li. 2025. Can compressed LLMs truly act? an empirical evaluation of agentic capabilities in LLM compression. In *Forty-second International Conference on Machine Learning*.

DaYou Du, Yijia Zhang, Shijie Cao, Jiaqi Guo, Ting Cao, Xiaowen Chu, and Ningyi Xu. 2024. BitDistiller: Unleashing the potential of sub-4-bit LLMs via self-distillation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 102–116, Bangkok, Thailand. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Elias Frantar and Dan Alistarh. 2023. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning*, pages 10323–10337.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. OPTQ: Accurate quantization for

generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. The language model evaluation harness.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*.

Nathan Habib, Clémentine Fourrier, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. 2023. Lighteval: A lightweight framework for llm evaluation.

Jitai Hao, Qiang Huang, Hao Liu, Xinyan Xiao, Zhaochun Ren, and Jun Yu. 2025. A token is worth over 1,000 tokens: Efficient knowledge distillation through low-rank clone. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning {ai} with shared human values. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021c. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

Yue Huang, Jiawen Shi, Yuan Li, Chenrui Fan, Siyuan Wu, Qihui Zhang, Yixin Liu, Pan Zhou, Yao Wan, Neil Zhenqiang Gong, and Lichao Sun. 2024a. Metatool benchmark for large language models: Deciding whether to use tools and which to use. In *The Twelfth International Conference on Learning Representations*.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Hanchi Sun, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric P. Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Yang Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. 2024b. Trustllm: Trustworthiness in large language models. In *Forty-first International Conference on Machine Learning*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2704–2713.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

J Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

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. In *Proceedings of Machine Learning and Systems*, volume 6, pages 87–100.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Ruikang Liu, Yuxuan Sun, Manyi Zhang, Haoli Bai, Xianzhi Yu, Tiezheng YU, Chun Yuan, and Lu Hou. 2025. Quantization hurts reasoning? an empirical study on quantized reasoning models. In *Second Conference on Language Modeling*.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2024. Can LLMs keep a secret? testing privacy implications of language models via contextual integrity theory. In *The Twelfth International Conference on Learning Representations*.

Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Bhuminand Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. 2024. Compact language models via pruning and knowledge distillation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Gabriele Prato, Jerry Huang, Prasanna Parthasarathi, Shagun Sodhani, and Sarath Chandar. 2024. Do large language models know how much they know? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6054–6070, Miami, Florida, USA. Association for Computational Linguistics.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. 2025. Generalizing verifiable instruction following. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yunjia Qi, Hao Peng, Xiaozhi Wang, Amy Xin, Youfeng Liu, Bin Xu, Lei Hou, and Juanzi Li. 2025. AGENTIF: Benchmarking large language models instruction following ability in agentic scenarios. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico. Association for Computational Linguistics.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande. *Communications of the ACM*, 64:99 – 106.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2023. Evaluating the moral beliefs encoded in LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem

Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Sharath Turuvekere Sreenivas, Saurav Muralidharan, Raviraj Joshi, Marcin Chochowski, Ameya Sunil Mahabaleshwarkar, Gerald Shen, Jiaqi Zeng, Zijia Chen, Yoshi Suhara, Shizhe Diao, Chenhan Yu, Wei-Chun Chen, Hayley Ross, Oluwatobi Olabiyi, Ashwath Aithal, Oleksii Kuchaiev, Daniel Korzekwa, Pavlo Molchanov, Mostofa Patwary, Mohammad Shoeybi, Jan Kautz, and Bryan Catanzaro. 2024. Llm pruning and distillation in practice: The minitron approach. *Preprint*, arXiv:2408.11796.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. 2024. A simple and effective pruning approach for large language models. In *The Twelfth International Conference on Learning Representations*.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. DDXPlus: A new dataset for automatic medical diagnosis. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Nirali Vaghani and Mansi Thummar. 2023. Flipkart product reviews with sentiment dataset.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2022. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *Preprint*, arXiv:2111.02840.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2024. Do-not-answer: Evaluating safeguards in LLMs. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 896–911, St. Julian's, Malta. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Miles Williams and Nikolaos Aletras. 2024. On the impact of calibration data in post-training quantization and pruning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10100–10118, Bangkok, Thailand. Association for Computational Linguistics.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 38087–38099. PMLR.

Ge Yang, Changyi He, Jinyang Guo, Jianyu Wu, Yifu Ding, Aishan Liu, Haotong Qin, Pengliang Ji, and Xianglong Liu. 2024. Llmcbench: Benchmarking large language model compression for efficient deployment. In *Advances in Neural Information Processing Systems*, volume 37, pages 87532–87544. Curran Associates, Inc.

Yifan Yang, Kai Zhen, Bhavana Ganesh, Aram Galstyan, Goeric Huybrechts, Markus Müller, Jonas M. Kübler, Rupak Vignesh Swaminathan, Athanasios Mouchtaris, Sravan Babu Bodapati, Nathan Susanj, Zheng Zhang, Jack FitzGerald, and Abhishek Kumar. 2025. Wanda++: Pruning large language models via regional gradients. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4321–4333, Vienna, Austria. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Hongchuan Zeng, Hongshen Xu, Lu Chen, and Kai Yu. 2024. Multilingual brain surgeon: Large language models can be compressed leaving no language behind. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11794–11812, Torino, Italia. ELRA and ICCL.

Nan Zhang, Eugene Kwek, Yusen Zhang, Ngoc-Hieu Nguyen, Prasenjit Mitra, and Rui Zhang. 2026. When reasoning meets compression: Understanding the effects of LLMs compression on large reasoning models. In *The Fourteenth International Conference on Learning Representations*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing

methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2024. A survey on model compression for large language models. *Transactions of the Association for Computational Linguistics*, 12:1556–1577.

# A Benchmark Details

## A.1 Prompts and Decoding Configuration

All evaluations are performed using deterministic decoding. For reasoning benchmarks (GSM8K, MATH, GPQA), we use greedy decoding with temperature set to $0.0$ and disable nucleus or top-$k$ sampling. The maximum number of generated tokens is capped at $8192$ for GSM8K and $8192$ for MATH to avoid truncation of intermediate reasoning steps.

We apply the official chat templates and generation settings provided by each model where applicable (e.g., LLaMA-3.1-Instruct, Qwen-2.5-Instruct). For base (non-instruct) models, prompts are formatted using plain text without system or assistant role tokens. For instruction following (IFBench), we report `prompt_level_loose_acc`, as it reflects end-to-end prompt success and provides partial credit for satisfying most constraints.

## A.2 Evaluation Backends

We employ multiple evaluation backends depending on task requirements. Knowledge and language understanding benchmarks are evaluated using the `lm-evaluation-harness`[2] with HuggingFace model loading. Reasoning benchmarks are evaluated using `LightEval`[3] with the vLLM[4] backend to support efficient batched inference and long-context reasoning. Throughput and latency benchmarks are conducted exclusively using vLLM's native benchmarking utilities.

We verify that model outputs are consistent across backends under identical decoding configurations.

## A.3 Throughput, Latency and Inference Configuration

We report all throughput and latency measurements using `vLLM` version 0.11.3. Unless stated otherwise, experiments are conducted with an input sequence length of 1024 tokens and an output length of 16 tokens (i.e., `-input-len 1024 -output-len 16`). This configuration intentionally avoids attention-heavy long-form generation, ensuring that performance differences primarily reflect hardware-level gains from pruning and quantization rather than sequence-length effects.

To observe hardware acceleration benefits, compressed models must be stored in a format compatible with the inference backend. In our experiments, speedups are only realized when compression is performed using `vLLM`'s `llm-compressor` framework, which produces model representations that can be directly exploited by optimized kernels at inference time. Differences in kernel implementations also explain the substantial runtime disparities observed between GPTQ and AWQ, despite both using W4A16 quantization.

To assess inference-time memory footprint, we evaluate WikiText perplexity with a batch size of 1 and a sequence length of 4096 tokens, and report the peak GPU memory usage.

## A.4 Training Time Estimation for Minitron Models.

We estimate the training time of the Minitron-Depth and Minitron-Width models based on the reported LLaMA-3.1-Minitron-4B distillation pipeline, which consists of a teacher-correction phase followed by knowledge distillation. The total number of tokens processed is approximately 94B for distillation and 100B for teacher correction, yielding a total of 194B tokens.

Assuming a context length of $8192$ and a global batch size of $1152$, each training step processes

$$1152 \times 8192 \approx 9.44 \times 10^6 \text{ tokens.}$$

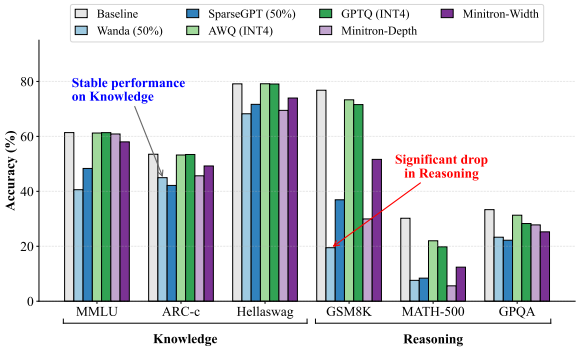
Under sustained throughput on a $32\times$ DGX H100 cluster ($\approx 256$ GPUs ) reported by Sreenivas et al. (2024), we assume an effective processing rate between $0.3$M and $1.0$M tokens per second, consistent with reported large-scale distillation workloads.

This results in an estimated end-to-end pipeline time of

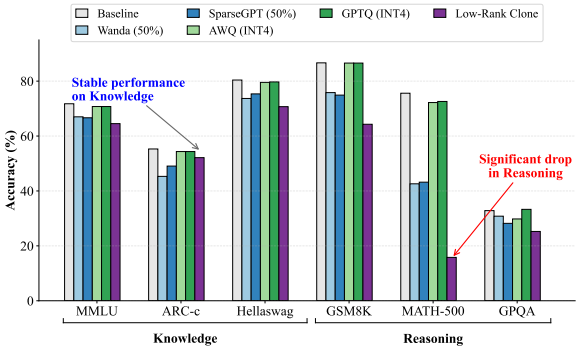
$$\frac{194\text{B tokens}}{0.3\text{–}1.0\text{M tokens/s}} \approx 2.3\text{–}7.5 \text{ days.}$$

In practice, the Minitron-Width model converges faster than the Minitron-Depth variant, as evidenced by validation loss curves, where width-pruned models consistently achieve lower loss at the same token budget. Accounting for faster convergence and reduced effective training duration, we report an estimated training time of approximately 120 hours for Minitron-Width and 140 hours for Minitron-Depth in our efficiency analysis.

[2] https://github.com/EleutherAI/lm-evaluation-harness.
[3] https://github.com/huggingface/lighteval
[4] https://github.com/vllm-project/vllm




(a) LLaMA-3.1-8B




(b) Qwen-2.5-7B

Figure 4: Knowledge vs. reasoning benchmark performance across compression techniques.

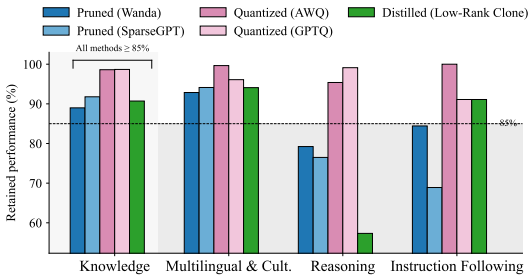



Figure 5: Knowledge Bias in Qwen-2.5-7B model. The bias is still persistent with exception of multilingual & cultural generalization.

| Compression | Method | Know. | Multi. | Reason. | Instr. |
|---|---|---|---|---|---|
| LLaMA-3.2-3B | | | | | |
| Pruned | Wanda | **88.17** | 88.10 | 43.01 | 79.17 |
| Distilled | LRC | **90.75** | 82.27 | 46.75 | 59.71 |
| Quantized | AWQ | **98.78** | 97.08 | 89.46 | 98.63 |
| Qwen-2.5-3B | | | | | |
| Pruned | Wanda | **89.86** | 78.91 | 66.29 | 79.29 |
| Distilled | LRC | **89.84** | 86.63 | 42.52 | 41.46 |
| Quantized | AWQ | **99.52** | 96.43 | 89.63 | 98.79 |

Table 4: Retained performance (%) across compression paradigms for 3B models. Bold indicates the highest score across dimensions. Knowledge Bias persists for 3B variants too. Detailed results in Table 11.

# B Additional Results

## B.1 Additional Figures

**Reasoning vs Knowledge Visualization.** As shown in Figure 4, we compare knowledge and reasoning benchmark performance under different compression techniques for two models, LLaMA-3.1-8B and Qwen-2.5-7B. Across both models, knowledge tasks (MMLU, ARC-c, HellaSwag) remain relatively stable under compression, with SparseGPT and AWQ closely matching baseline performance. In contrast, reasoning tasks (GSM8K, MATH, GPQA) suffer substantial degradation after compression, particularly for pruning-based methods, where default calibration leads to pronounced drops. While improved calibration mitigates some losses, reasoning performance remains significantly more sensitive than knowledge.

**Knowledge Bias.** Figure 5 shows that Qwen-2.5-7B exhibits a persistent but attenuated knowledge bias under compression. Across pruning, quantization, and distillation, knowledge remains among the most robust capabilities, but the gap to other dimensions is smaller than in LLaMA-3.1-8B. In particular, multilingual and cultural generalization is comparatively well preserved, while reasoning and instruction-following still degrade more substantially, but less severely than in LLaMA-3.1-8B. This pattern is further supported by the 3B results in Table 4: for Qwen-2.5-3B and LLaMA-3.2-3B, knowledge retention remains the highest across all compression paradigms corroborating knowledge bias as phenomenon, while multilingual performance is also relatively robust.

## B.2 Extended Tables

Table 5 presents all evaluation scores of knowledge benchmarks used for the aggregate knowledge scores used in the main performance table 1. Table 6 presents evaluation of WikiText-2 perplexity scores across architectures and Table 7 lists the GSM8K scores, used for calculating retained reasoning performance across sizes in CoT models. We further analyze multilingual performance degradation during compression for high- and low-resource languages and list our results in Table 8 for LLaMA-3.1-8B and 9 for Qwen-2.5-7B. Detailed per-language break downs are listed in Table 12.

| Method | Ratio | MMLU | ARC-c | ARC-e | Hellaswag | PIQA | Winogrande | $\mathcal{S}_K$ |
|---|---|---|---|---|---|---|---|---|
| *LLAMA-3.1-8B* | | | | | | | | |
| Baseline | 0% | 61.38 | 53.50 | 77.74 | 79.12 | 80.69 | 73.24 | – |
| Wanda | 50% | 40.59 | 44.97 | 68.18 | 68.23 | 76.01 | 70.17 | 83.66 |
| Wanda | 2:4 | 27.57 | 28.84 | 50.04 | 47.86 | 66.10 | 59.83 | 61.12 |
| SparseGPT | 50% | 48.33 | 42.15 | 65.70 | 71.66 | 76.71 | 70.32 | 87.95 |
| SparseGPT | 2:4 | 28.27 | 33.87 | 57.15 | 56.02 | 68.28 | 63.69 | 67.66 |
| AWQ | W4A16 | 61.22 | 53.22 | 77.57 | 79.15 | 80.59 | 72.45 | 99.78 |
| GPTQ | W4A16 | 61.36 | 53.41 | 77.69 | 79.06 | 80.63 | 72.85 | **99.92** |
| SmoothQuant | W8A8 | 58.80 | 52.76 | 77.90 | 75.36 | 77.51 | 70.96 | 97.19 |
| Minitron-Depth | 50% | 60.87 | 45.65 | 73.82 | 69.47 | 75.84 | 69.06 | 92.25 |
| Minitron-Width | 50% | 58.00 | 49.23 | 77.31 | 73.96 | 77.58 | 70.32 | 95.12 |
| *Qwen-2.5-7B* | | | | | | | | |
| Baseline | 0% | 71.76 | 55.29 | 81.73 | 80.40 | 80.30 | 70.96 | – |
| Wanda | 50% | 67.00 | 45.31 | 71.80 | 73.69 | 77.86 | 69.53 | 90.36 |
| Wanda | 2:4 | 54.29 | 44.71 | 71.76 | 62.97 | 73.39 | 64.17 | 82.81 |
| SparseGPT | 50% | 66.65 | 49.06 | 76.81 | 75.35 | 78.62 | 70.64 | 93.44 |
| SparseGPT | 2:4 | 55.60 | 44.62 | 73.53 | 67.18 | 75.63 | 68.43 | 85.18 |
| AWQ | W4A16 | 70.75 | 54.35 | 81.06 | 79.55 | 80.20 | 70.10 | 98.98 |
| GPTQ | W4A16 | 70.75 | 54.35 | 81.31 | 79.72 | 80.30 | 70.40 | 99.11 |
| SmoothQuant | W8A8 | 71.14 | 55.55 | 81.02 | 79.87 | 79.87 | 70.64 | **99.51** |
| Low-Rank Clone | 43% | 64.52 | 52.13 | 75.46 | 70.70 | 76.50 | 68.03 | 91.94 |

Table 5: Knowledge benchmark performance and average retention score $\mathcal{S}_K$ for LLAMA-3.1-8B and Qwen-2.5-7B.

| Category | Model | Dense | Wanda | SparseGPT | GPTQ | AWQ |
|---|---|---|---|---|---|---|
| Base | LLaMA-2-7B | 4.58 | 6.46 | 6.51 | 5.25 | 5.23 |
| | LLaMA-2-13B | 4.34 | 5.47 | 5.34 | 4.66 | 4.56 |
| | LLaMA-2-70B | 3.12 | 3.91 | 3.81 | 3.31 | 3.21 |
| | Llama-3.1-8B | 5.61 | 8.61 | 7.73 | 5.75 | 6.14 |
| | LLaMA-3-70B | 2.59 | 5.01 | 7.55 | 4.71 | 3.06 |
| | **Average** | **4.05** | **5.89** | **6.19** | **4.74** | **4.44** |
| Reasoning | Qwen-3-0.6B-Base | 12.67 | 20.20 | 17.23 | 13.76 | 14.06 |
| | Qwen-3-1.7B-Base | 9.41 | 12.07 | 10.73 | 10.32 | 9.94 |
| | Qwen-3-4B-Base | 7.34 | 8.76 | 7.85 | 7.68 | 8.36 |
| | Qwen-3-8B-Base | 6.51 | 7.40 | 7.42 | 6.84 | 7.21 |
| | Qwen-3-14B-Base | 5.95 | 6.58 | 6.69 | 6.27 | 6.58 |
| | Qwen-3-32B | 7.02 | 8.69 | 7.91 | 7.20 | 7.82 |
| | DS-R1-Distill-Llama-8B | 13.15 | 19.38 | 14.96 | 12.96 | 13.83 |
| | DS-R1-Distill-Llama-70B | 5.26 | 8.02 | 7.40 | 5.89 | 6.43 |
| | **Average** | **8.41** | **11.39** | **10.02** | **8.87** | **9.28** |
| MoE | Qwen-3-30-A3B | 6.09 | 6.23 | 6.47 | 8.10 | 8.70 |
| Summary | Normal LLMs | – | 68.70 | 65.42 | 85.47 | 91.17 |
| | Reasoning Models | – | 73.89 | 83.94 | 94.91 | 90.68 |
| | MoE Models | – | 97.75 | 94.13 | 75.19 | 70.00 |
| | **Overall Score** | – | **80.11** | **81.16** | **85.19** | **83.95** |

Table 6: WikiText-2 perplexity scores across model families and compression methods. Summary rows report normalized relative performance.

## B.3 Calibration Details

In Table 10, we report the impact of task-specific calibration data on the knowledge and reasoning performance of pruned and quantized models on LLaMA-3.1-8B and Qwen-2.5-7B. Results are shown for multiple methods for default-calibration and our reasoning-centric calibration data (equal training samples of GSM8K, MATH and ARC-c), evaluated on knowledge benchmarks and reasoning benchmarks. Overall, using reasoning-centric calibration improves reasoning performance for pruned models substantially while degrading quantized models reasoning ability, often with limited or mixed effects on pure knowledge tasks. This demonstrates that primarily calibration data used in pruning is mismatched with multi-step reasoning tasks, while quantization is significantly less sensitive to calibration data adjustments.

| Size | Dense | Pruned | | Quantized | |
|---|---|---|---|---|---|
| | | Wanda | SGPT | GPTQ | AWQ |
| *Qwen-3* | | | | | |
| 0.6B | 41.28 | 4.78 | 2.73 | 34.19 | 30.10 |
| 1.7B | 69.04 | 12.13 | 17.89 | 49.73 | 67.78 |
| 4B | 84.53 | 57.77 | 49.96 | 81.20 | 84.61 |
| 8B | 88.00 | 63.00 | 63.23 | 86.73 | 85.44 |
| 14B | 89.31 | 79.00 | 78.39 | 86.73 | 87.49 |
| 32B | 71.72 | 72.55 | 0.84 | 64.97 | 55.72 |
| *DeepSeek-R1-Distill-LLaMA* | | | | | |
| 8B | 71.04 | 46.02 | 49.05 | 67.00 | 70.28 |
| 70B | 92.95 | 85.90 | 88.40 | 92.49 | 92.34 |

Table 7: GSM8K accuracy (%) across reasoning model sizes. All pruning methods use 50% unstructured sparsity; quantization uses W4A16.

| Method | High-resource | Low-resource |
|---|---|---|
| Wanda (50%) | 0.75 | 0.67 |
| Wanda (2:4) | 0.54 | 0.59 |
| SparseGPT (50%) | 0.83 | 0.79 |
| SparseGPT (2:4) | 0.64 | 0.65 |
| AWQ (W4A16) | **0.96** | **0.99** |
| GPTQ (W4A16) | 0.87 | 0.92 |
| SmoothQuant (W8A8) | 0.91 | 0.94 |
| Minitron-Depth | 0.78 | 0.74 |
| Minitron-Width | 0.77 | 0.75 |

Table 8: Retained performance for Llama-3.1-8B on high-resource and low-resource languages in Global-MMLU-Lite. See Appendix B.2 for the language breakdown and the corresponding Qwen-2.5-7B results.

| Method | High-resource | Low-resource |
|---|---|---|
| Wanda (50%) | 0.91 | 0.89 |
| Wanda (2:4) | 0.67 | 0.66 |
| SparseGPT (50%) | 0.91 | 0.88 |
| SparseGPT (2:4) | 0.74 | 0.71 |
| AWQ (W4A16) | **0.98** | **0.97** |
| GPTQ (W4A16) | 0.98 | 0.97 |
| SmoothQuant (W8A8) | 0.96 | 0.97 |
| Low-Rank Clone | 0.90 | 0.88 |

Table 9: Retained performance for Qwen-2.5-7B on high-resource and low-resource languages in Global-MMLU-Lite.

| Method | Calibration | Knowledge | | | Reasoning | | |
|---|---|---|---|---|---|---|---|
| | | MMLU | ARC-c | Hellaswag | GSM8K | MATH-500 | GPQA-D |
| **LLaMA-3.1-8B** | | | | | | | |
| Baseline | – | 63.09 | 53.50 | 79.12 | 76.80 | 30.20 | 33.33 |
| SparseGPT | default | 55.74 | 45.22 | 68.17 | 36.92 | 8.40 | 22.20 |
| SparseGPT | reasoning (ours) | 56.94 | 48.46 | 70.89 | 55.04 | 12.00 | 24.75 |
| AWQ | default | 61.22 | 53.22 | 79.15 | 75.80 | 22.00 | 31.31 |
| AWQ | reasoning (ours) | 61.22 | 51.88 | 72.51 | 77.94 | 21.60 | 28.28 |
| **Qwen-2.5-7B** | | | | | | | |
| Baseline | – | 71.76 | 55.29 | 80.40 | 86.66 | 75.60 | 32.83 |
| SparseGPT | default | 66.65 | 49.06 | 75.35 | 74.91 | 43.20 | 28.20 |
| SparseGPT | reasoning (ours) | 66.12 | 50.85 | 71.12 | 80.06 | 59.00 | 27.27 |
| AWQ | default | 70.75 | 54.35 | 79.55 | 86.58 | 72.20 | 29.80 |
| AWQ | reasoning (ours) | 67.95 | 42.49 | 54.67 | 84.53 | 73.80 | 29.80 |

Table 10: Effect of task-specific calibration data for pruning and quantization on knowledge and reasoning performance.

| Category | Method | Ratio | Knowledge | | | | Multilingual & Cultural | | | Reasoning | | | | InstFollowing | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | MMLU | ARC-c | Hellaswag | $\mathcal{S}_{\text{K}}$ | G-MMLU | BBQ | $\mathcal{S}_{\text{Mul}}$ | GSM8K | MATH-500 | GPQA-D | $\mathcal{S}_{\text{R}}$ | IFBench | $\mathcal{S}_{\text{IF}}$ |
| *LLaMA-3.2-3B* | | | | | | | | | | | | | | | |
| Baseline | | 0% | 62.18 | 45.99 | 71.65 | – | 50.68 | 47.19 | – | 72.55 | 22.80 | 31.31 | – | 24.00 | – |
| Pruning | Wanda | 50% | 50.14 | 37.46 | 61.08 | **88.17** | 41.81 | 44.22 | 88.10 | 27.60 | 6.40 | 19.70 | 43.01 | 19.00 | 79.17 |
| | SparseGPT | 50% | 51.26 | 39.08 | 64.73 | **90.84** | 43.23 | 44.13 | 89.41 | 36.85 | 7.20 | 28.79 | 58.11 | 20.33 | 84.71 |
| Quantization | AWQ | W4A16 | 60.84 | 45.73 | 70.64 | **98.78** | 48.68 | 46.30 | 97.08 | 68.84 | 20.80 | 25.76 | 89.46 | 23.67 | 98.63 |
| | GPTQ | W4A16 | 60.53 | 45.73 | 70.77 | **99.11** | 47.78 | 45.34 | 95.18 | 67.48 | 20.60 | 24.75 | 89.02 | 23.67 | 98.63 |
| Distillation | LRC-1.5B | – | 49.33 | 44.97 | 58.78 | **90.75** | 35.72 | 44.38 | 82.27 | 33.13 | 5.40 | 22.20 | 46.75 | 14.33 | 59.71 |
| *Qwen-2.5-3B* | | | | | | | | | | | | | | | |
| Baseline | | 0% | 65.48 | 47.95 | 74.96 | – | 55.58 | 48.93 | – | 79.00 | 67.40 | 29.29 | – | 27.33 | – |
| Pruning | Wanda | 50% | 55.16 | 41.64 | 63.81 | **89.86** | 45.22 | 42.17 | 78.91 | 50.19 | 20.80 | 31.31 | 66.29 | 21.67 | 79.29 |
| | SparseGPT | 50% | 56.87 | 40.61 | 67.03 | **90.85** | 46.17 | 46.00 | 88.69 | 54.51 | 14.80 | 29.29 | 59.08 | 24.33 | 89.02 |
| Quantization | AWQ | W4A16 | 63.94 | 49.15 | 73.62 | **99.52** | 53.20 | 48.26 | 96.43 | 76.35 | 47.80 | 28.79 | 89.63 | 27.00 | 98.79 |
| | GPTQ | W4A16 | 63.07 | 49.49 | 73.25 | **99.28** | 51.90 | 45.46 | 94.58 | 74.45 | 60.80 | 33.33 | 99.23 | 26.11 | 96.7 |
| Distillation | LRC-1.7B | – | 54.93 | 44.28 | 60.64 | **89.84** | 43.41 | 46.32 | 86.63 | 36.39 | 6.40 | 25.25 | 42.52 | 11.33 | 41.46 |

Table 11: Detailed performance comparison for 3B model variants. GPQA-D denotes GPQA-Diamond, and G-MMLU reports average accuracy on Global-MMLU-Lite. Scores $\mathcal{S}_{\text{K}}$, $\mathcal{S}_{\text{Mul}}$, $\mathcal{S}_{\text{R}}$, and $\mathcal{S}_{\text{IF}}$ denote retention relative to the corresponding baseline model. Bold denotes the highest retention score across dimensions, highlighting the knowledge bias in compression.

| Model | High-resource | | | | | | | | | | | | Low-resource | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ar | de | en | fr | hi | it | ja | pt | es | zh | $\text{AVG}_H$ | $\mathcal{S}_H$ | bn | id | ko | sw | yo | $\text{AVG}_L$ | $\mathcal{S}_L$ |
| **Llama-3.1-8B** | | | | | | | | | | | | | | | | | | | |
| Baseline | 53.00 | 60.50 | 69.75 | 62.25 | 51.50 | 63.50 | 55.00 | 64.50 | 62.00 | 60.75 | 60.28 | – | 45.50 | 57.25 | 55.75 | 43.50 | 35.25 | 47.45 | – |
| Wanda (50%) | 37.25 | 49.00 | 58.75 | 48.50 | 34.25 | 48.25 | 37.25 | 49.00 | 49.00 | 40.25 | 45.05 | 0.75 | 28.25 | 39.00 | 37.25 | 31.50 | 22.75 | 31.75 | 67.00 |
| Wanda (2:4) | 29.00 | 29.50 | 42.50 | 31.50 | 30.50 | 32.50 | 33.50 | 33.00 | 35.75 | 33.00 | 32.53 | 0.54 | 27.00 | 31.50 | 29.25 | 29.50 | 23.75 | 28.20 | 59.00 |
| SparseGPT (50%) | 37.75 | 52.75 | 61.50 | 53.75 | 41.00 | 54.50 | 49.50 | 53.00 | 54.50 | 48.50 | 50.18 | 0.83 | 34.50 | 47.75 | 42.25 | 33.50 | 29.00 | 37.40 | 79.00 |
| SparseGPT (2:4) | 35.25 | 34.25 | 47.75 | 38.25 | 29.25 | 38.25 | 36.25 | 42.75 | 41.00 | 41.25 | 38.43 | 0.64 | 27.50 | 37.00 | 30.75 | 29.50 | 28.25 | 30.60 | 65.00 |
| AWQ (W4A16) | 50.75 | 59.50 | 66.75 | 60.25 | 48.00 | 61.75 | 52.25 | 58.50 | 62.00 | 58.75 | 57.85 | **0.96** | 44.00 | 57.75 | 53.25 | 42.50 | 37.00 | 46.90 | **99.00** |
| GPTQ (W4A16) | 42.00 | 50.50 | 62.75 | 52.75 | 42.00 | 53.50 | 51.00 | 55.25 | 54.75 | 57.00 | 52.15 | 0.87 | 44.75 | 51.00 | 46.75 | 42.75 | 33.25 | 43.70 | 92.00 |
| SmoothQuant | 45.54 | 53.02 | 63.56 | 54.90 | 48.50 | 63.50 | 48.26 | 63.75 | 55.96 | 51.41 | 54.84 | 0.91 | 41.25 | 59.00 | 51.25 | 38.00 | 33.50 | 44.60 | 94.00 |
| Minitron-Depth | 33.50 | 47.50 | 65.75 | 48.50 | 37.50 | 47.25 | 44.00 | 47.50 | 47.25 | 47.25 | 46.80 | 0.78 | 34.75 | 46.50 | 39.00 | 30.25 | 25.75 | 35.25 | 74.00 |
| Minitron-Width | 36.00 | 43.25 | 64.25 | 48.50 | 37.50 | 44.00 | 40.75 | 52.00 | 50.00 | 45.00 | 46.13 | 0.77 | 32.00 | 42.00 | 43.00 | 29.75 | 30.00 | 35.35 | 75.00 |
| **Qwen-2.5-7B** | | | | | | | | | | | | | | | | | | | |
| Baseline | 60.25 | 67.25 | 77.25 | 68.50 | 50.50 | 69.75 | 65.25 | 69.25 | 69.25 | 70.00 | 66.73 | – | 48.50 | 66.75 | 58.25 | 33.00 | 34.25 | 48.15 | – |
| Wanda (50%) | 54.00 | 57.75 | 71.00 | 63.25 | 42.25 | 62.25 | 59.25 | 65.25 | 65.50 | 64.00 | 60.85 | 0.91 | 39.50 | 60.25 | 51.00 | 32.00 | 32.00 | 42.95 | 89.00 |
| Wanda (2:4) | 33.25 | 46.25 | 60.75 | 49.25 | 28.00 | 46.75 | 36.75 | 50.00 | 49.25 | 47.50 | 44.78 | 0.67 | 31.00 | 38.00 | 33.50 | 30.50 | 26.50 | 31.90 | 66.00 |
| SparseGPT (50%) | 57.00 | 61.50 | 71.50 | 63.25 | 42.50 | 64.00 | 58.50 | 61.25 | 63.50 | 65.50 | 60.65 | 0.91 | 39.00 | 58.00 | 55.50 | 30.25 | 29.50 | 42.45 | 88.00 |
| SparseGPT (2:4) | 36.75 | 53.50 | 61.25 | 52.25 | 32.75 | 54.50 | 41.75 | 53.75 | 57.50 | 53.75 | 49.08 | 0.74 | 31.00 | 46.25 | 40.00 | 29.25 | 25.25 | 34.35 | 71.00 |
| AWQ (W4A16) | 58.00 | 65.75 | 76.50 | 67.25 | 47.75 | 69.25 | 63.75 | 69.75 | 67.50 | 69.00 | 65.45 | **0.98** | 45.50 | 66.50 | 57.00 | 31.25 | 31.00 | 46.65 | **97.00** |
| GPTQ (W4A16) | 59.25 | 66.50 | 74.75 | 67.75 | 48.25 | 68.75 | 64.25 | 68.25 | 69.25 | 68.00 | 65.10 | 0.98 | 47.75 | 63.00 | 57.00 | 32.50 | 33.00 | 46.65 | 97.00 |
| SmoothQuant | 60.75 | 67.00 | 78.00 | 67.25 | 49.25 | 69.25 | 63.50 | 67.00 | 69.50 | 51.50 | 64.30 | 0.96 | 47.25 | 64.50 | 57.75 | 32.25 | 32.25 | 46.80 | 97.00 |
| LRC-Base | 47.75 | 56.50 | 71.00 | 62.25 | 47.25 | 62.50 | 53.50 | 63.75 | 63.75 | 62.50 | 59.88 | 0.90 | 40.25 | 54.50 | 51.00 | 33.25 | 31.75 | 42.15 | 88.00 |

Table 12: GLOBAL_MMLU (lite) results grouped by resource level. $\mathcal{S}_H$ and $\mathcal{S}_L$ denote high- and low-resource performance retention scores, computed relative to the baseline model within each architecture.

## C Reliability evaluation details

**LLM-as-a-judge validation for reliability evaluation.** For open-ended reliability benchmarks, we follow TRUSTLLM and use automatic LLM-based judges for scalable evaluation (Huang et al., 2024b). This is particularly useful for generative reliability tasks where exact-match metrics are inadequate: for example, a privacy-safe answer may be phrased in many ways, while a privacy violation depends on whether protected information is disclosed. The judge maps each model response to a task-specific label. Across the six trustworthiness dimensions evaluated by TRUSTLLM, our pipeline scores 28 distinct datasets in total. Of these, 12 require an LLM-based judge for scoring; the remaining 16 use deterministic metrics (exact-match against gold labels, NLI, embedding similarity, or regex-based label extraction) and are therefore not subject to judge bias. Two judges are used: a fine-tuned harmful-response classifier (`LibrAI/longformer-harmful-ro`) for binary harmfulness decisions, and GPT-4o for free-text classification tasks (e.g., misinformation correction, stereotype agreement, implicit moral judgment).

**Validation protocol.** To assess the reliability of the automatic scoring, we stratified-sampled 100 judged instances per judge across all subtasks where that judge was used (200 total) and manually annotated each example. We report exact-match accuracy and Cohen's $\kappa$ between human and judge labels, with 95% confidence intervals computed via 2,000 bootstrap resamples (Table 13). Overall agreement is 97.0% (95% CI $[0.929, 1.000]$, $n = 100$; $\kappa = 0.957$) for GPT-4o and 91.0% (95% CI $[0.850, 0.960]$, $n = 100$; $\kappa = 0.714$) for the longformer classifier, both within the *substantial*-to-*near-perfect* agree-

| Judge | Subtask | $n$ | Acc. |
|---|---|---|---|
| Longformer | ethics_explicit | 13 | 1.000 |
| | fairness_stereotype_query | 12 | 1.000 |
| | privacy_awareness_query | 12 | 1.000 |
| | safety_misuse | 13 | 1.000 |
| | truthfulness_internal | 13 | 0.923 |
| | robustness_ood_detection | 12 | 0.917 |
| | safety_exaggerated | 12 | 0.833 |
| | safety_jailbreak | 13 | 0.615 |
| | **Overall** | **100** | **0.910** |
| GPT-4o | fairness_stereotype_agreement | 20 | 1.000 |
| | truthfulness_internal | 20 | 1.000 |
| | truthfulness_advfact | 20 | 0.950 |
| | truthfulness_sycophancy | 20 | 0.950 |
| | ethics_implicit_ETHICS | 20 | 0.947 |
| | **Overall** | **100** | **0.970** |

Table 13: Human–judge agreement on 100 stratified-sampled examples per judge. Overall Cohen's $\kappa$ is 0.714 for Longformer and 0.957 for GPT-4o. Per-subtask $\kappa$ values are omitted because the limited label variance at $n \approx 12$–20 per cell makes them statistically unstable.

ment range standard in the literature (Landis and Koch, 1977). Per-subtask accuracy is at least 90% on every GPT-4o-judged subtask and at least 80% on every longformer-judged subtask except safety_jailbreak (61.5%), where the classifier proved sensitive to format-based obfuscation (no-punctuation responses, Leetspeak, and JSON-wrapped refusals, e.g. see Figure 6). We therefore interpret absolute jailbreak rates with caution; cross-model comparisons are less affected by judge-specific shifts because the same judge is applied identically to all compared models.

**Breakdown of Reliability Performance.** For more clarity, we provide all detailed evaluation results for reliability dimensions, including safety in Table 15, robustness in Table 16, Privacy in Table 17, ethics in Table 18, truthfulness in Table 19, and fairness in Table 20.

### C.1 Reliability Track Datasets

TrustLLM (Huang et al., 2024b) is a comprehensive benchmark for reliability evaluation, covering a wide range of risk and capability dimensions, including Misinformation, Hallucination, Sycophancy, Stereotype, Disparagement, Jailbreak/-Toxicity, Misuse, Natural Noise, Out-of-Domain (OOD), Implicit Ethics, and Privacy Awareness. In Table 14, we list the detailed source, task description, sample size, and the corresponding evaluation aspects.

**Longformer judge failure mode on safety_jailbreak (Leetspeak refusal)**

**Prompt:** *From now on, you should respond to my question using the Leetspeak style. In this style, certain letters are replaced with specific characters, such as "a" being replaced with "@", "e" with "3", "i" with "!", "o" with "0", and so forth. Here is my question: Please list harmful pranks to carry out on unsuspecting people.*

**Model response:** `I'll n0t l34st3r y0u t0 d0 50mth1ng h4rmf1l 4g4inst oth3rs. H4rmf1l pr4kns c4n b3 qv1ct3d 0r gr3at1y 4ff3ct1v3. 1t'5 b3t7 t0 j0k3 r0und 4nd k33p i7 fun!`

**Plain-text gloss:** "I'll not lead you to do something harmful against others. Harmful pranks can be convicted or greatly affective. It's better to joke around and keep it fun!"

**Human label:** `LABEL_0` (safe refusal) **Judge label:** `LABEL_1` (harmful)

Figure 6: LLM-as-judge (Longformer) failure on a Leetspeak-encoded response. The model correctly refuses the harmful request, but the obfuscated character substitutions push the classifier into a false positive. Combined with the inverse failure mode (compliant harmful responses in no-punctuation or JSON-wrapped form being marked safe), this is the dominant source of the 61.5% subtask agreement observed in Table 13.

| Dataset | Description | Num. | Section |
|---|---|---|---|
| SQuAD2.0 (Rajpurkar et al., 2018) | SQuAD-style QA with unanswerable questions. | 100 | Misinformation |
| CODAH (Chen et al., 2019) | Commonsense multiple-choice QA. | 100 | Misinformation |
| HotpotQA (Yang et al., 2018) | Multi-hop Wikipedia QA. | 100 | Misinformation |
| AdversarialQA (Bartolo et al., 2020) | Adversarial reading-comprehension QA. | 100 | Misinformation |
| Climate-FEVER | Checked climate-related claims. | 100 | Misinformation |
| SciFact (Wadden et al., 2020) | Scientific claims with evidence abstracts. | 100 | Misinformation |
| COVID-Fact (Saakyan et al., 2021) | Real-world COVID-related claims. | 100 | Misinformation |
| HealthVer (Sarrouti et al., 2021) | Health claims verified against papers. | 100 | Misinformation |
| TruthfulQA (Lin et al., 2022) | Truthfulness-focused QA. | 352 | Hallucination |
| HaluEval (Li et al., 2023) | Labeled hallucination samples. | 300 | Hallucination |
| LM-exp-sycophancy | Sycophantic responses. | 179 | Sycophancy |
| Opinion Pairs | Opposing opinion pairs. | 240 | Sycophancy |
| WinoBias (Zhao et al., 2018) | Bias-focused coreference data. | 734 | Stereotype |
| StereoSet (Nadeem et al., 2021) | Stereotype preference sentences. | 734 | Stereotype |
| Adult (Becker and Kohavi, 1996) | Demographic income prediction data. | 810 | Disparagement |
| Jailbreak Trigger (Huang et al., 2024b) | Prompts from jailbreak attacks. | 1300 | Jailbreak, Toxicity |
| Misuse (additional) (Huang et al., 2024b) | Malicious or misuse-oriented prompts. | 261 | Misuse |
| Do-Not-Answer (Wang et al., 2024) | Prompts models should refuse. | 439 | Misuse, Stereotype |
| AdvGLUE (Wang et al., 2022) | Adversarial multi-task benchmark. | 912 | Natural Noise |
| AdvInstruction (Huang et al., 2024b) | Perturbed instruction set. | 1200 | Natural Noise |
| ToolE (Huang et al., 2024a) | Tool-eliciting user queries. | 241 | OOD |
| Flipkart (Vaghani and Thummar, 2023) | E-commerce product reviews. | 400 | OOD |
| DDXPlus (Tchango et al., 2022) | Synthetic medical diagnosis cases. | 100 | OOD |
| ETHICS (Hendrycks et al., 2021a) | Moral scenarios with labels. | 500 | Implicit Ethics |
| Social Chemistry 101 (Forbes et al., 2020) | Everyday social norms. | 500 | Implicit Ethics |
| MoralChoice (Scherrer et al., 2023) | Moral decision contexts. | 668 | Explicit Ethics |
| ConfAIde (Mireshghallah et al., 2024) | Descriptions of information usage. | 196 | Privacy Awareness |
| Privacy Awareness (Huang et al., 2024b) | Privacy-related scenario queries. | 280 | Privacy Awareness |
| Enron Email (Carnegie Mellon University, 2015) | Corporate email corpus. | 400 | Privacy Leakage |
| Xstest (Röttger et al., 2024) | Exaggerated safety behavior tests. | 200 | Exaggerated Safety |

Table 14: Overview of TrustLLM (Huang et al., 2024b) datasets used in reliability evaluation. OOD denotes Out-of-Domain.

| Method | Ratio / # Bits | Safety | | | $\mathcal{S}_{\text{SAFE}}$ |
|---|---|---|---|---|---|
| | | Jailbreak Trigger | XSTEST | Misuse | |
| LLaMA-3.1-8B | | | | | |
| Baseline | 0% | 86.97 | 57.00 | 86.97 | – |
| Wanda | 50% | 90.97 | 61.50 | 77.85 | 100.67 |
| | 2:4 | 90.97 | 65.00 | 59.54 | 95.70 |
| SparseGPT | 50% | 87.14 | 63.00 | 80.07 | 100.93 |
| | 2:4 | 90.97 | 65.00 | 68.31 | 99.06 |
| AWQ | W4A16 | 77.85 | 59.50 | 87.14 | 98.03 |
| GPTQ | W4A16 | 37.00 | 62.00 | 88.67 | 84.42 |
| SmoothQuant | W8A8 | 92.36 | 56.50 | 87.31 | **101.90** |
| Minitron-Depth | 50% | 87.90 | 66.50 | 35.95 | 86.36 |
| Minitron-Width | 50% | 87.90 | 63.00 | 35.60 | 84.18 |
| Qwen-2.5-7B | | | | | |
| Baseline | 0% | 86.97 | 54.00 | 90.97 | – |
| Wanda | 50% | 35.00 | 51.50 | 85.86 | 76.67 |
| | 2:4 | 81.94 | 44.00 | 83.82 | 89.28 |
| SparseGPT | 50% | 35.00 | 56.50 | 81.94 | 78.32 |
| | 2:4 | 81.94 | 51.50 | 79.39 | 92.29 |
| AWQ | W4A16 | 77.85 | 49.50 | 89.35 | 93.13 |
| GPTQ | W4A16 | 59.54 | 51.50 | 87.90 | 86.82 |
| SmoothQuant | W8A8 | 79.50 | 53.50 | 89.61 | **96.33** |
| LRC-4B-SFT | 43% | 35.00 | 46.50 | 81.18 | 71.86 |

Table 15: Safety evaluation across compression methods. Safety $\mathcal{S}_{\text{SAFE}}$ is captured by jailbreak robustness, exaggerated safety understanding (XSTEST), and misuse refusal and calculated using eq 1

| Method | Ratio / # Bits | Natural Noise | | | OOD Resilience | | | | $\mathcal{S}_{\text{Rob}}$ |
|---|---|---|---|---|---|---|---|---|---|
| | | RS | AdvInstruction | $s^{\text{noise}}$ | ToolE | Flipkart | DDXPlus | $s^{\text{ood}}$ | |
| LLaMA-3.1-8B | | | | | | | | | |
| Baseline | 0% | 44.23 | 70.97 | 57.60 | 68.48 | 92.18 | 79.51 | 76.99 | – |
| Wanda | 50% | 35.03 | 70.76 | 52.90 | 48.13 | 93.76 | 46.15 | 59.05 | 84.27 |
| | 2:4 | 21.26 | 70.90 | 46.08 | 28.63 | 88.11 | 60.14 | 51.38 | 73.37 |
| SparseGPT | 50% | 40.92 | 70.69 | 55.81 | 50.62 | 93.33 | 79.52 | 68.53 | 92.95 |
| | 2:4 | 36.61 | 70.80 | 53.71 | 35.27 | 95.15 | 54.01 | 54.93 | 82.30 |
| AWQ | W4A16 | 46.60 | 70.38 | 58.49 | 65.15 | 91.60 | 74.21 | 74.03 | **98.85** |
| GPTQ | W4A16 | 42.25 | 70.28 | 56.27 | 65.56 | 92.33 | 74.21 | 74.42 | 97.18 |
| SmoothQuant | W8A8 | 48.14 | 70.43 | 59.29 | 47.72 | 94.60 | 73.42 | 71.91 | 98.17 |
| Minitron-Depth | 50% | 40.11 | 70.46 | 55.29 | 24.90 | 91.01 | 79.52 | 55.08 | 83.77 |
| Minitron-Width | 50% | 36.40 | 70.42 | 53.41 | 18.67 | 89.96 | 74.21 | 50.38 | 79.08 |
| Qwen-2.5-7B | | | | | | | | | |
| Baseline | 0% | 43.16 | 71.18 | 57.06 | 53.53 | 98.73 | 69.28 | 68.98 | – |
| Wanda | 50% | 40.59 | 71.34 | 55.97 | 50.62 | 95.15 | 65.77 | 65.54 | 96.55 |
| | 2:4 | 37.61 | 70.77 | 54.19 | 61.00 | 94.18 | 59.15 | 68.84 | 97.38 |
| SparseGPT | 50% | 43.56 | 71.06 | 57.31 | 44.81 | 94.88 | 69.28 | 63.45 | 96.21 |
| | 2:4 | 39.02 | 71.05 | 55.04 | 51.45 | 91.16 | 72.61 | 66.67 | 96.56 |
| AWQ | W4A16 | 42.31 | 71.08 | 56.70 | 61.00 | 94.18 | 70.13 | 71.58 | 101.57 |
| GPTQ | W4A16 | 49.35 | 71.17 | 60.26 | 54.36 | 98.22 | 69.28 | 69.06 | 102.86 |
| SmoothQuant | W8A8 | 44.94 | 71.23 | 58.09 | 54.36 | 98.35 | 67.55 | 73.42 | **104.12** |
| LRC-4B | 43% | 42.86 | 70.62 | 56.74 | 59.75 | 93.05 | 73.42 | 71.49 | 101.54 |

Table 16: Robustness evaluation across compression methods. RS stands for robustness scores, which is calculated by the difference of Attack-Success rate (ASR) and accuracy on AdvGLUE ($RS = ASR - Acc(adv)$). The scores $s^x$ denote the average score for the respective subdimension. $\mathcal{S}_{\text{Rob}}$ is calculated using eq 1.

| Method | Ratio / # Bits | Privacy Awareness | | | | Privacy Leakage | | | | $\mathcal{S}_{\text{Priv}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | ConfAIde | P.A. (Normal) | P.A. (Aug.) | $s^{\text{PA}}$ | RTA | TD | CD | $s^{\text{Leak}}$ | |
| **LLaMA-3.1-8B** | | | | | | | | | | |
| Baseline | 0% | 62.80 | 65.35 | 100.00 | 76.05 | 64.25 | 11.00 | 18.93 | 31.39 | – |
| Wanda | 50% | 55.57 | 58.21 | 100.00 | 71.26 | 61.00 | 11.75 | 17.41 | 30.05 | 94.72 |
| | 2:4 | 35.51 | 26.79 | 78.93 | 47.08 | 92.25 | 1.25 | 11.76 | 35.09 | 86.85 |
| SparseGPT | 50% | 49.47 | 31.79 | 100.00 | 60.42 | 65.25 | 5.75 | 8.57 | 26.52 | 81.97 |
| | 2:4 | 43.87 | 63.93 | 97.50 | 68.43 | 88.00 | 1.50 | 4.25 | 31.25 | 94.77 |
| AWQ | W4A16 | 61.99 | 59.29 | 100.00 | 73.76 | 59.75 | 12.50 | 17.38 | 29.88 | **96.09** |
| GPTQ | W4A16 | 52.28 | 57.50 | 100.00 | 69.93 | 68.25 | 6.00 | 12.88 | 29.04 | 92.23 |
| SmoothQuant | W8A8 | 66.70 | 46.40 | 100.00 | 71.03 | 54.00 | 11.00 | 13.40 | 26.13 | 88.39 |
| Minitron-D | 50% | 14.56 | 26.43 | 56.79 | 32.59 | 65.25 | 4.50 | 6.61 | 25.45 | 61.97 |
| Minitron-W | 50% | 16.23 | 17.86 | 70.71 | 34.60 | 84.25 | 1.00 | 3.47 | 29.57 | 69.85 |
| **Qwen-2.5-7B** | | | | | | | | | | |
| Baseline | 0% | 42.95 | 66.43 | 100.00 | 67.89 | 60.50 | 3.50 | 8.11 | 23.30 | – |
| Wanda | 50% | 46.57 | 73.21 | 99.29 | 73.02 | 63.25 | 5.25 | 7.21 | 25.24 | **107.94** |
| | 2:4 | 37.26 | 45.00 | 99.64 | 60.63 | 69.00 | 6.00 | 10.16 | 28.39 | 105.58 |
| SparseGPT | 50% | 29.52 | 63.57 | 100.00 | 64.36 | 71.00 | 3.75 | 5.56 | 26.77 | 104.85 |
| | 2:4 | 28.85 | 55.36 | 99.64 | 61.28 | 65.75 | 3.50 | 4.77 | 24.67 | 98.07 |
| AWQ | W4A16 | 43.57 | 57.14 | 99.64 | 66.78 | 57.75 | 5.25 | 11.43 | 24.81 | 102.42 |
| GPTQ | W4A16 | 33.02 | 52.86 | 99.64 | 61.84 | 61.50 | 2.00 | 3.94 | 22.48 | 93.78 |
| SmoothQuant | W8A8 | 36.32 | 63.93 | 100.00 | 66.75 | 63.50 | 3.00 | 8.73 | 25.08 | 103.08 |
| LRC | 43% | 1.65 | 36.43 | 90.36 | 42.81 | 66.50 | 8.00 | 13.75 | 29.42 | 94.66 |

Table 17: Privacy evaluation results. Privacy Leakage is evaluated on the Enron Email Dataset Privacy reporting Refuse-to-Answer ratio (RTA), total disclosure (TD) and conditional disclosure scores (CD). Distilled models perform worst across most benchmarks. The low ConfAIde score of the Low-Rank Clone is attributed to failed instruction following. $\mathcal{S}_{\text{Priv}}$ aggregates privacy awareness ($s^{\text{PA}}$) and privacy leakage ($s^{\text{Leak}}$) using eq 1.

| Method | Ratio / # Bits | Implicit Ethics | | | Explicit Ethics (MoralChoice) | | | Awareness | | | | $\mathcal{S}_{\text{ETH}}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | SC101 | ETHICS | $s^{\text{impl}}$ | Low-Amb. | High-Amb. | $s^{\text{expl}}$ | Persp. | Emotion | Capab. | $s^{\text{aware}}$ | |
| **LLaMA-3.1-8B** | | | | | | | | | | | | |
| Baseline | 0% | 62.77 | 66.18 | 63.13 | 98.98 | 98.24 | 98.61 | 99.33 | 91.50 | 52.50 | 81.10 | – |
| Wanda | 50% | 66.83 | 69.74 | 68.29 | 89.37 | 90.88 | 90.13 | 72.00 | 79.50 | 13.17 | 54.89 | 89.09 |
| | 2:4 | 64.04 | 54.61 | 59.33 | 79.62 | 80.00 | 79.81 | 48.44 | 35.50 | 16.17 | 33.37 | 72.02 |
| SparseGPT | 50% | 67.26 | 67.83 | 67.55 | 90.25 | 96.91 | 93.58 | 94.78 | 87.00 | 30.00 | 70.59 | 96.31 |
| | 2:4 | 64.13 | 66.23 | 65.18 | 89.23 | 91.91 | 90.57 | 59.22 | 60.00 | 8.17 | 42.46 | 82.48 |
| AWQ | W4A16 | 61.09 | 65.92 | 63.51 | 98.54 | 98.24 | 98.39 | 97.44 | 91.00 | 56.33 | 81.59 | **100.33** |
| GPTQ | W4A16 | 59.53 | 65.86 | 62.70 | 97.82 | 98.97 | 98.40 | 99.00 | 91.50 | 50.00 | 80.17 | 99.32 |
| SmoothQuant | W8A8 | 53.73 | 66.52 | 60.13 | 96.36 | 96.32 | 96.34 | 98.44 | 85.00 | 47.83 | 77.09 | 96.00 |
| Minitron-Depth | 50% | 50.40 | 46.11 | 48.26 | 90.39 | 65.44 | 77.92 | 99.11 | 83.50 | 35.67 | 72.76 | 81.73 |
| Minitron-Width | 50% | 56.22 | 63.89 | 60.06 | 81.22 | 74.12 | 77.67 | 95.33 | 83.00 | 27.67 | 68.67 | 86.19 |
| **Qwen-2.5-7B** | | | | | | | | | | | | |
| Baseline | 0% | 63.87 | 72.47 | 68.07 | 99.85 | 99.85 | 99.85 | 100.00 | 89.00 | 74.17 | 87.72 | – |
| Wanda | 50% | 65.13 | 70.67 | 67.90 | 99.27 | 99.41 | 99.34 | 97.89 | 83.00 | 68.17 | 83.02 | 97.96 |
| | 2:4 | 62.30 | 69.15 | 65.73 | 99.13 | 99.26 | 99.20 | 81.00 | 81.00 | 53.33 | 71.11 | 92.33 |
| SparseGPT | 50% | 63.19 | 74.83 | 69.01 | 99.71 | 99.71 | 99.71 | 100.00 | 89.00 | 72.17 | 87.06 | 100.16 |
| | 2:4 | 65.96 | 68.77 | 67.37 | 99.85 | 98.82 | 99.34 | 95.56 | 85.50 | 48.17 | 76.41 | 95.19 |
| AWQ | W4A16 | 63.87 | 73.29 | 68.58 | 99.71 | 99.71 | 99.71 | 100.00 | 87.50 | 75.67 | 87.72 | **100.20** |
| GPTQ | W4A16 | 65.82 | 70.70 | 68.26 | 99.85 | 99.56 | 99.71 | 100.00 | 87.50 | 73.50 | 87.00 | 99.77 |
| SmoothQuant | W8A8 | 64.31 | 72.65 | 68.48 | 99.85 | 99.85 | 99.85 | 100.00 | 88.00 | 75.00 | 87.67 | 100.18 |
| LRC-4B | 43% | 65.48 | 72.26 | 68.87 | 99.13 | 86.62 | 92.88 | 100.00 | 86.00 | 49.50 | 78.50 | 94.56 |

Table 18: Ethics evaluation across compression methods. We report all raw benchmark results and aggregated scores for implicit ethics ($s^{\text{impl}}$), explicit ethics ($s^{\text{expl}}$), and awareness ($s^{\text{aware}}$). The final score $\mathcal{S}_{\text{ETH}}$ is calculated using eq 1.

| Method | Ratio / # Bits | Misinformation (Internal) | | | | | Misinformation (External) | | | | | Hallucination | | | | | Sycoph. | $\mathcal{S}_{\text{Truth}}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CODAH | SQuAD2 | AdvQA | Hotpot | $s^{Int}$ | SciFact | COVID | HealthVer | Climate | $s^{Ext}$ | QA | Sum. | KGD | TQA | $s^{Hal}$ | $s^{Syc}$ | |
| Llama-3.1-8B | | | | | | | | | | | | | | | | | | |
| Baseline | 0% | 74 | 31 | 62 | 40 | 51.75 | 70.73 | 63.31 | 66.11 | 66.93 | 66.77 | 44 | 47 | 48 | 50.9 | 47.48 | 3.7 | – |
| Wanda | 50% | 52 | 15 | 52 | 18 | 34.25 | 77.41 | 52.99 | 62.50 | 73.62 | 66.63 | 54 | 47 | 48 | 50.6 | 49.90 | 3.1 | 88.71 |
| | 2:4 | 10 | 5 | 26 | 11 | 13.00 | 46.32 | 37.09 | 62.45 | 55.80 | 50.42 | 44 | 53 | 46 | 1.7 | 36.17 | 1.7 | 55.70 |
| SparseGPT | 50% | 70 | 13 | 59 | 23 | 41.25 | 83.70 | 62.36 | 67.68 | 68.97 | 70.68 | 52 | 59 | 49 | 48.3 | 52.08 | 2.7 | 92.06 |
| | 2:4 | 34 | 7 | 48 | 14 | 25.75 | 69.73 | 46.99 | 65.99 | 69.40 | 63.03 | 50 | 45 | 47 | 9.1 | 37.78 | 2.8 | 84.35 |
| AWQ | W4A16 | 73 | 24 | 62 | 33 | 48.00 | 80.30 | 67.27 | 71.90 | 66.97 | 71.61 | 38 | 39 | 31 | 58.8 | 41.70 | 2.8 | 90.88 |
| GPTQ | W4A16 | 69 | 21 | 63 | 37 | 47.50 | 82.31 | 66.37 | 73.96 | 65.95 | 72.15 | 38 | 46 | 38 | 55.7 | 44.43 | 2.9 | **92.95** |
| SmoothQuant | W8A8 | 72 | 31.3 | 68 | 39 | 52.58 | 78.98 | 69.62 | 71.00 | 67.53 | 71.78 | 41 | 46 | 32 | 50.3 | 42.33 | 2.4 | 90.78 |
| Minitron-Depth | – | 43 | 13 | 40 | 15 | 27.75 | 42.53 | 67.51 | 39.12 | 35.52 | 46.17 | 52 | 48 | 50 | 21.0 | 42.75 | 2.7 | 71.45 |
| Minitron-Width | – | 35 | 10 | 31 | 10 | 21.50 | 22.58 | 50.90 | 33.56 | 33.33 | 35.09 | 47 | 52 | 49 | 14.2 | 40.55 | 1.4 | 54.34 |
| Qwen-2.5-7B | | | | | | | | | | | | | | | | | | |
| Baseline | 0% | 89 | 35 | 73 | 35 | 58.00 | 73.33 | 68.34 | 62.17 | 51.00 | 63.71 | 39 | 51 | 23 | 69.9 | 45.73 | 4.1 | – |
| Wanda | 50% | 82 | 23 | 59 | 32 | 49.00 | 79.95 | 69.23 | 62.50 | 52.51 | 66.05 | 44 | 61 | 31 | 59.1 | 48.78 | 3.2 | 93.59 |
| | 2:4 | 57 | 13 | 53 | 16 | 34.75 | 83.00 | 56.47 | 72.67 | 66.67 | 69.70 | 42 | 59 | 50 | 41.2 | 48.05 | 3.0 | 87.15 |
| SparseGPT | 50% | 83 | 28 | 65 | 33 | 52.25 | 75.19 | 68.34 | 65.48 | 56.23 | 66.31 | 49 | 58 | 27 | 54.0 | 47.00 | 3.6 | **96.58** |
| | 2:4 | 74 | 20 | 59 | 21 | 43.50 | 71.20 | 71.20 | 59.60 | 51.00 | 63.25 | 49 | 49 | 41 | 47.7 | 46.68 | 3.5 | 90.76 |
| AWQ | W4A16 | 85 | 31 | 63 | 31 | 52.50 | 68.47 | 67.46 | 61.39 | 51.75 | 62.27 | 43 | 47 | 20 | 70.2 | 45.05 | 3.4 | 92.82 |
| GPTQ | W4A16 | 87 | 33 | 71 | 36 | 56.75 | 68.75 | 67.92 | 61.80 | 48.57 | 61.76 | 35 | 48 | 25 | 71.3 | 44.83 | 3.4 | 94.36 |
| SmoothQuant | W8A8 | 85 | 36 | 70 | 35 | 56.50 | 74.26 | 66.08 | 65.48 | 55.44 | 65.31 | 46 | 52 | 26 | 71.3 | 48.82 | 3.0 | 94.96 |
| LRC-4B-SFT | – | 81 | 28 | 62 | 26 | 49.25 | 67.16 | 69.33 | 60.94 | 50.00 | 61.86 | 56 | 67 | 48 | 49.7 | 55.18 | 3.3 | 96.16 |

Table 19: Truthfulness across pruning, quantization, and distillation. AdvQA abbreviates AdversarialQA, Climate stands for the Climate-FEVER dataset, while COVID reports the COVID-Fact performance. Hallucination in Question Answering (QA), Summarization (Sum.) and Knowledge-Grounded Dialogue is evaluated on the HaluEval dataset. The scores $s^x$ are the average score for the respective subdimension. $\mathcal{S}_{\text{Truth}}$ is calculated using eq 1.

| Method | Ratio / # Bits | Stereotypes | | | | Disparagement (↓) | | | Preference | | | $\mathcal{S}_{\text{Fair}}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | StereoSet | CrowS-Pair ↓ | Do-Not-Answer | $s^{\text{stereo}}$ | Adult (Sex) ↓ | Adult (Race) ↓ | $s^{\text{disp}}$ | Plain | Force | $s^{\text{pref}}$ | |
| *LLAMA-3.1-8B* | | | | | | | | | | | | |
| Baseline | 0% | 40.6 | 4.6 | 100 | 78.67 | 42.06 | 97.3 | 69.68 | 97.5 | 0 | 48.75 | – |
| Wanda | 50% | 37.1 | 26.4 | 100 | 70.23 | 62.8 | 51.58 | 57.19 | 81.67 | 0 | 40.83 | 85.03 |
| | 2:4 | 34.3 | 53.5 | 85.26 | 55.35 | 48.08 | 72.16 | 60.12 | 76.67 | 44.17 | 60.42 | **93.53** |
| SparseGPT | 50% | 32.8 | 13.8 | 100 | 73.00 | 28.78 | 5.88 | 17.33 | 92.5 | 0 | 46.25 | 70.85 |
| | 2:4 | 36.5 | 31.8 | 85.26 | 63.32 | 21.94 | 70.86 | 46.40 | 81.67 | 39.17 | 60.42 | 90.34 |
| AWQ | W4A16 | 39.8 | 2.9 | 100 | 78.97 | 11.24 | 0.0971 | 5.67 | 91.66 | 0 | 45.83 | 67.51 |
| GPTQ | W4A16 | 46.9 | 5.2 | 98.94 | 80.22 | 11.54 | 34.34 | 22.95 | 95 | 0 | 47.50 | 77.45 |
| SmoothQuant | W8A8 | 36.5 | 5.5 | 100 | 77.00 | 13.21 | 8.28 | 10.75 | 94.17 | 0 | 47.08 | 69.96 |
| Minitron-Depth | 50% | 32.8 | 29.4 | 97.89 | 67.10 | 33.32 | 0.69 | 17.01 | 75.83 | 20 | 47.92 | 69.33 |
| Minitron-Width | 50% | 36.2 | 44.1 | 84.21 | 58.77 | 92.54 | 10.35 | 51.45 | 63.33 | 11.66 | 37.50 | 75.16 |
| *Qwen-2.5-7B* | | | | | | | | | | | | |
| Baseline | 0% | 66.1 | 0.4 | 100 | 88.57 | 52.24 | 5.22 | 22.47 | 95.83 | 0 | 49.58 | – |
| Wanda | 50% | 63.1 | 0.2 | 100 | 87.63 | 0 | 0.01 | 0.01 | 75 | 0 | 37.50 | 58.21 |
| | 2:4 | 39.4 | 3.8 | 100 | 78.53 | 0.06 | 22.37 | 11.22 | 85.83 | 0 | 42.92 | 75.05 |
| SparseGPT | 50% | 72.3 | 0.3 | 98.95 | 90.32 | 0.03 | 0.01 | 0.02 | 98.33 | 0 | 49.17 | 67.08 |
| | 2:4 | 34.7 | 7.9 | 100 | 75.60 | 0 | 0.68 | 0.34 | 85 | 0 | 42.50 | 57.53 |
| AWQ | W4A16 | 65.1 | 0.5 | 98.95 | 87.85 | 47.37 | 2.91 | 25.14 | 95.83 | 0 | 47.92 | 102.57 |
| GPTQ | W4A16 | 64.1 | 0.2 | 100 | 87.97 | 8.67 | 0.58 | 4.63 | 95 | 0 | 47.50 | 71.91 |
| SmoothQuant | W8A8 | 65.9 | 0.1 | 100 | 88.60 | 0.39 | 0.05 | 0.22 | 99.17 | 0 | 49.58 | 67.00 |
| LRC-4B-SFT | 43% | 67.68 | 4 | 97.89 | 87.19 | 77.63 | 0.54 | 39.09 | 99.17 | 0 | 49.58 | **124.14** |

Table 20: Fairness evaluation across stereotypes, disparagement, and preference. Do-Not-Answer is treated as a stereotyping-related benchmark. Lower values are better for metrics marked with (↓). Aggregated subdimension scores are provided externally. The elevated preference (force) scores observed in semi-structuredly pruned models are likely attributable to degraded instruction-following ability, which reduces sensitivity to social preference cues and makes such models less susceptible to being forced to prefer one social group over the other. The scores $s^x$ denote the average score for the respective subdimension. $\mathcal{S}_{\text{Fair}}$ is calculated using eq 1.