Task-Stratified Knowledge Scaling Laws for Post-Training Quantized Large Language Models

Chenxi Zhou[1,2], Pengfei Cao[2,3]*, Jiang Li[4], Bohan Yu[1,2], Jinyu Ye[2], Jun Zhao[2,3], Kang Liu[2,3]*

[1]School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences
[2]The Key Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences
[4]College of Computer Science, Inner Mongolia University
zhouchenxi2025@ia.ac.cn, {pengfei.cao, jzhao, kliu}@nlpr.ia.ac.cn

Abstract

Post-Training Quantization (PTQ) is a critical strategy for efficient Large Language Models (LLMs) deployment. However, existing scaling laws primarily focus on general performance, overlooking crucial fine-grained factors and how quantization differentially impacts diverse knowledge capabilities. To address this, we establish Task-Stratified Knowledge Scaling Laws. By stratifying capabilities into memorization, application, and reasoning, we develop a framework that unifies model size, bit-width, and fine-grained factors: group size and calibration set size. Validated on 293 diverse PTQ configurations, our framework demonstrates strong fit and cross-architecture consistency. It reveals distinct sensitivities across knowledge capabilities: reasoning is precision-critical, application is scale-responsive, and memorization is calibration-sensitive. We highlight that in low-bit scenarios, optimizing these fine-grained factors is essential for preventing performance collapse. These findings provide an empirically-backed foundation for designing knowledge-aware quantization strategies.

1 Introduction

Large language models (LLMs) have achieved impressive performance across diverse tasks (Guo et al., 2023), but their growing scale poses deployment challenges due to high memory and computational costs (Zhu et al., 2024; Lang et al., 2024). Post-training quantization (PTQ) emerges as a practical solution by compressing LLMs without expensive retraining (Yao et al., 2023). A recent study shows that nearly 70% of quantization-related research since 2022 has focused on PTQ for LLMs (Zhao et al., 2025).

Despite the widespread use of PTQ, a comprehensive understanding of how LLM performance is precisely impacted under quantization remains elusive. Current evaluations offer general insights, such as performance cliffs below 4-bit precision (Li et al., 2024) and task-specific sensitivities (Marchisio et al., 2024; Liu et al., 2025). However, these studies typically lack a systematic and predictive framework. This deficiency makes it difficult for practitioners to make informed decisions when configuring PTQ strategies. To this end, some researchers have initiated the exploration of scaling laws for quantized models, aiming to establish relationships between model performance and factors, such as model size or bit-width (Ouyang et al., 2024; Kumar et al., 2025; Xu et al., 2024). Such scaling laws enable the prediction of post-quantization performance. However, they still have two notable limitations:

1) The role of fine-grained PTQ factors is overlooked. Current studies predominantly focus on factors like model size, bit-width, and pre-training data volume (Ouyang et al., 2024; Kumar et al., 2025). In contrast, tunable parameters inherent in widely adopted algorithms (e.g., GPTQ (Frantar et al., 2023)), such as group size (Elangovan et al., 2025) and calibration set size (Zhang et al., 2025), are often treated as constants. However, our empirical observations reveal that these fine-grained parameters are decisive factors for maintaining model capabilities, especially under low-bit quantization.

2) The impact of quantization on diverse knowledge capabilities remains underexplored. Existing scaling laws mainly focus on the overall performance of quantized LLMs, often overlooking the fact that LLMs possess diverse knowledge capabilities. This is critical as they rely on core capabilities, ranging from memorization to application and reasoning, to support diverse downstream tasks (Wang et al., 2024; Yu et al., 2024). Crucially, these capabilities are hypothesized to exhibit divergent sensitivities to quantization due to their distinct underlying mechanisms, which gen-

*Corresponding authors

Findings of the Association for Computational Linguistics: ACL 2026, pages 23268–23285

eral scaling laws fail to capture.

To address these limitations, we conduct an extensive empirical investigation to establish **Task-Stratified Knowledge Scaling Laws** for post-training quantized LLMs. Specifically, this involves: 1) *systematically incorporating model size, bit-width, calibration set size, and group size into a unified power-law framework*; and 2) *comprehensively investigating the impact of quantization configurations on the diverse knowledge capabilities of LLMs*. Validated on 293 diverse PTQ configurations spanning the Qwen3 and Llama-3 families, our framework demonstrates a strong fit and cross-architecture universality. We reveal that different knowledge capabilities exhibit distinct sensitivities to quantization variables. Specifically, while reasoning is bottlenecked by precision (bit-width and group size), knowledge application scales significantly with model size, and memorization is particularly sensitive to calibration set size. Furthermore, we highlight that under low-bit quantization, smaller group sizes and sufficient calibration data are no longer optional but essential to prevent performance collapse.

In summary, our contributions are twofold:

- We establish the first task-stratified knowledge scaling laws for PTQ. Our unified framework incorporates model size and bit-width alongside crucial fine-grained factors (group size and calibration set size), and models diverse knowledge capabilities separately.
- We empirically reveal divergent sensitivities across knowledge capabilities (memorization, application, and reasoning) to quantization, and highlight that optimizing fine-grained factors is essential for preventing performance collapse under low-bit scenarios.

2 Related Work

2.1 Post-Training Quantization of LLMs

Post-Training Quantization (PTQ) has emerged as a dominant strategy for LLM compression, offering superior efficiency over Quantization-Aware Training (QAT) by eliminating retraining (Lang et al., 2024; Hasan, 2024). While PTQ methods vary widely, they generally balance compression and performance via sophisticated calibration techniques (Williams and Aletras, 2024; Ji et al., 2024).

Among these, optimization-based approaches like GPTQ (Frantar et al., 2023) have become industry standards. GPTQ leverages second-order information (Hessian matrix) and calibration data to minimize quantization error layer-by-layer. Crucially, the performance of such methods is intricately tied to hyperparameters like calibration set size and group granularity (Zhang et al., 2025; Elangovan et al., 2025). However, prior works typically treat these as static settings rather than dynamic scaling variables, leaving their systematic impact on model capabilities underexplored.

2.2 Scaling Laws for Quantized LLMs

Neural scaling laws provide a predictive framework linking model performance to resources. Pioneering works by Kaplan et al. (2020) and Hoffmann et al. (2022) establish that uncompressed LLM performance follows power laws with model size, training tokens, and training compute.

Recently, this framework has been extended to the quantization domain. For instance, Ouyang et al. (2024) investigate scaling laws for quantization-induced degradation (QiD), linking QiD to training data volume, model size, and bit-width. Kumar et al. (2025) explore the interplay between training precision and PTQ precision. Sun et al. (2025) explore the scaling behavior of floating-point representation structures during the training phase. Furthermore, Xu et al. (2024) attempt to build predictive models for post-PTQ quality considering various factors.

Despite these advancements, prior works primarily focus on generic performance metrics, overlooking how varying quantization configurations differentially impact distinct knowledge capabilities. The lack of a unified framework incorporating fine-grained factors leaves the scaling dynamics of diverse capabilities largely unquantified.

3 Task-Stratified Knowledge Scaling Laws for PTQ LLMs

3.1 Task Capability Definitions for Quantization Analysis

To systematically investigate the impact of PTQ on LLMs, we refine the knowledge capability taxonomy into three hierarchical levels of increasing cognitive complexity, as illustrated in Figure 1: *knowledge memorization*, *knowledge application*, and *knowledge reasoning*.

This stratification draws from Bloom's Taxonomy (Krathwohl, 2002; Huber and Niklaus, 2025), its adaptation for LLM benchmarks (e.g., KoLA (Yu et al., 2024)), and recent studies on

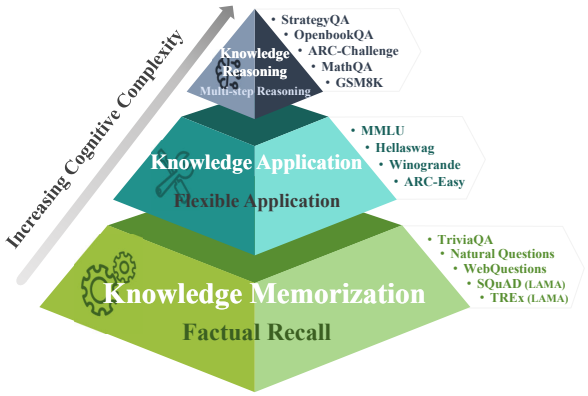


Figure 1: Overview of the task-stratified knowledge taxonomy defined in this study.

knowledge mechanisms in LLMs (Wang et al., 2024). We posit that these knowledge capabilities exhibit divergent sensitivities to quantization, necessitating a task-stratified scaling analysis.

Level 1: Knowledge Memorization (KM). Aligning with Bloom's *Remembering* level, this capability refers to an LLM's ability to accurately store and recall specific factual knowledge learned during pre-training. Tasks at this level are characterized by an "exact lookup" nature, where the model must recall precise facts (e.g., names, dates) from the internal knowledge base without complex contextual transformation.

Level 2: Knowledge Application (KA). Combining Bloom's *Understanding* and *Applying* levels, KA transcends static storage, focusing on comprehending inquiries and leveraging internalized knowledge to formulate appropriate answers. Unlike simple recall, this level requires the model to understand the context and apply generalized knowledge to specific scenarios, emphasizing flexible application rather than strict factual knowledge lookup.

Level 3: Knowledge Reasoning (KR). Aligning with Bloom's deep thinking skills (primarily *Analyzing* (Huber and Niklaus, 2025)), KR involves complex cognitive processes including multi-step logic, mathematical problem-solving, and chain-of-thought deduction (Wei et al., 2022). Unlike application, complex reasoning requires the model to construct multi-step logical chains to handle novel scenarios beyond simple pattern matching.

Based on this stratification, we aim to construct distinct scaling laws for each level, predicting how PTQ configurations impact diverse knowledge capabilities.

3.2 Factors under Investigation

To establish task-stratified scaling laws, we focus on four key factors governing the quantization process. Fundamentally, PTQ compresses a model of size N by mapping high-precision weights $\mathbf{W}$ to B-bit representations $\hat{\mathbf{W}}$. This process typically aims to minimize the reconstruction error $\|\mathbf{WX} - \hat{\mathbf{W}}\mathbf{X}\|_F^2$ on calibration inputs $\mathbf{X}$ (with set size C_b). Furthermore, the quantization granularity is determined by the group size G, which defines the block size of weights sharing the same quantization scale (and zero-point). We examine the scaling behaviors of these factors below:

(1) Model Size (N): Defined as the total number of non-embedding parameters (Ouyang et al., 2024), model size determines representational capacity and robustness to quantization noise. Figure 2 (left) confirms that accuracy consistently increases with model size across most bit-widths, following a power-law trend as in full-precision models (Kaplan et al., 2020; Hoffmann et al., 2022). However, the 2-bit models remain near the random baseline and improve only slightly at large scales, deviating markedly from higher-precision trends.

(2) Bit-width (B): As shown in Figure 2 (right), we observe a sharp recovery: performance rises steeply from the random baseline at 2-bit to a usable level at 3-bit, before saturating near BF16 performance at higher bit-widths. This observation highlights the non-linear impact of bit-width on model capabilities.

(3) Calibration Set Size (C_b): While the importance of calibration data is acknowledged (Zhang et al., 2025; Williams and Aletras, 2024; Ji et al., 2024), its systematic scaling behavior remains under-explored. As shown in Figure 3 (left), increasing C_b improves accuracy, but the benefits saturate at larger sizes. This non-linear saturation motivates its inclusion as a key factor to quantify its impact on knowledge preservation.

(4) Group Size (G): Group size serves as a trade-off between compression ratio and error compensation. Figure 3 (right) demonstrates a pronounced inverse relationship: smaller group sizes (e.g., 32, 64) mitigate accuracy loss via finer-grained quantization, whereas larger groups (e.g., 1024) cause obvious degradation. This confirms that G acts as a critical granularity regulator in PTQ.

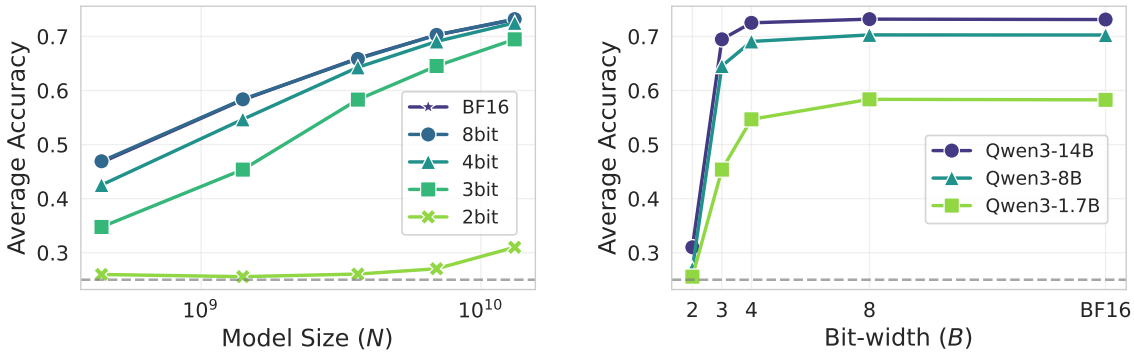


Figure 2: Scaling trends of Model Size (N) and Bit-width (B) for Qwen3 models ($C_b = 128, G = 128$). Accuracy is averaged across five representative 4-choice tasks: Hellaswag, ARC-e/c, MMLU, and OpenbookQA. The dashed grey line represents the random baseline (0.25). (BF16 and 8-bit curves visually overlap).

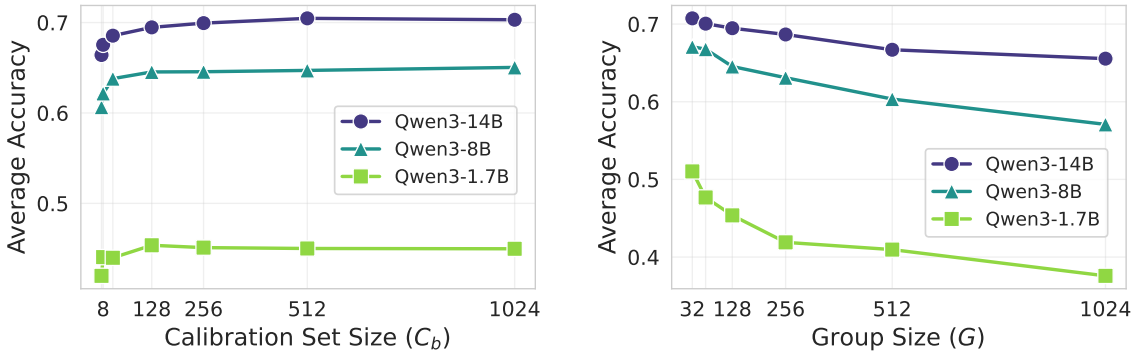


Figure 3: Scaling trends of Calibration Set Size (C_b) and Group Size (G) under 3-bit quantization. Benchmarks are the same as in Figure 2. (Left) Impact of C_b with fixed $G = 128$. (Right) Impact of G with fixed $C_b = 128$.

3.3 Scaling Law Formulation and Fitting Method

3.3.1 Task-Stratified Scaling Law

To quantitatively model the impact of quantization configurations on knowledge capabilities, we propose a unified multiplicative power-law function. The performance metric, denoted as the negative log-adjusted accuracy, is modeled as follows:

$$-\ln(\text{Acc}_{\text{adj}}) = A_{\text{task}} \cdot N^{\alpha_{\text{task}}} (\log_2 B)^{\beta_{\text{task}}} (\log_2 C_b)^{\gamma_{\text{task}}} G^{\delta_{\text{task}}}, \quad (1)$$

where A_{task} is a task-specific constant scaling coefficient. The exponents $\alpha_{\text{task}}, \beta_{\text{task}}, \gamma_{\text{task}}$, and δ_{task} are task-specific scaling parameters, quantifying the sensitivity of performance on that task type to each respective factor.

Note that since higher performance corresponds to a lower value of $-\ln(\text{Acc}_{\text{adj}})$, we expect negative exponents for resource-related factors (N, B, C_b), as scaling them up reduces this "loss" metric. Conversely, we anticipate a positive exponent for group size (G), since a larger group size implies coarser quantization granularity, which typically degrades performance (increases the "loss").

Theoretical Support. The adoption of this functional form is based on two key foundations. First, the multiplicative power-law structure successfully describes how neural networks scale, capturing the relationship between influential factors and model performance (Kaplan et al., 2020; Hoffmann et al., 2022). Second, we fit the negative natural logarithm of the adjusted accuracy instead of raw accuracy. As highlighted by Schaeffer et al. (2025), downstream metrics like accuracy are bounded in $[0, 1]$ and exhibit complex non-linear behaviors that are difficult to fit directly. Transforming accuracy into an unbounded "loss-like" space ($-\ln(\text{Acc})$) restores the monotonic, convex properties required for robust modeling (Krajewski et al., 2025). This form also allows the exponents to be understood as elasticities, quantifying the sensitivity of performance to relative changes in each factor.

Adjustment for Diverse Task Baselines. Our evaluation spans a diverse three-layer knowl-

edge taxonomy where random guessing baselines ($\text{Acc}_{\text{random}}$) vary significantly. For instance, generative tasks in knowledge memorization have a baseline approaching zero, whereas multiple-choice tasks in knowledge application have a baseline of 0.25 or 0.5. To eliminate this bias and ensure a unified scaling metric across different task types, we use the baseline-adjusted accuracy instead of raw accuracy:

$$\text{Acc}_{\text{adj}} = \frac{\text{Acc} - \text{Acc}_{\text{random}}}{1 - \text{Acc}_{\text{random}}}. \tag{2}$$

This adjustment ensures that Acc_{adj} reflects knowledge gain over random guessing, enabling consistent comparison in our task-stratified analysis.

3.3.2 Illustration for Logarithmic Transformation of C_b and B

As introduced in Eq. 1, we apply a logarithmic transformation ($\log_2$) to both calibration set size (C_b) and bit-width (B) to explicitly model their non-linear "diminishing returns" on model accuracy. Specifically, as observed in our preliminary experiments (Figure 2 and 3), initial increases in C_b or B yield substantial performance gains, but these benefits progressively diminish as the values become larger. The logarithmic transformation linearizes this saturation behavior, ensuring robust fitting across the effective range. This modeling choice aligns with prior work suggesting that the utility of additional calibration data (Williams and Aletras, 2024) and increased bit-width (Li et al., 2024) often follows such a non-linear pattern.

3.3.3 Fitting Method

To robustly estimate the coefficients ($A_{\text{task}}, \alpha_{\text{task}}, \beta_{\text{task}}, \gamma_{\text{task}}, \delta_{\text{task}}$), we transform the multiplicative scaling law into a linear form by taking the natural logarithm of both sides of Eq. 1:

$$\begin{aligned}\ln(-\ln(\text{Acc}_{\text{adj}})) = \ln A_{\text{task}} + \alpha_{\text{task}} \ln N \\ + \beta_{\text{task}} \ln(\log_2 B) \\ + \gamma_{\text{task}} \ln(\log_2 C_b) + \delta_{\text{task}} \ln G.\end{aligned} \tag{3}$$

We employ Ordinary Least Squares (OLS) linear regression (Zdaniuk, 2014) on this log-log data, filtering out collapsed configurations ($\text{Acc}_{\text{adj}} \leq 0.01$) to ensure numerical stability (Appendix A.4). Compared to direct Non-linear Least Squares (NLS) optimization, this linearized approach offers a closed-form solution and ensures convexity, avoiding local optima (Sengupta et al., 2025).

To rigorously evaluate the model's explanatory power, we employ the Adjusted R^2 statistic (Appendix B). We report this metric in two spaces: (1) the log-space ($\ln(-\ln(\text{Acc}_{\text{adj}}))$) to assess regression quality, and (2) the original space (Acc_{adj}) to validate practical predictive capability. Furthermore, we utilize Mean Absolute Error (MAE) to verify absolute accuracy and extrapolation robustness (Appendix D.2).

4 Experiments

4.1 Experimental Setup

We design a comprehensive setup to evaluate how PTQ parameters affect distinct knowledge capabilities. The implementation details, along with the rationale for benchmark stratification, are provided in Appendix A.

Models. We primarily study the Qwen3 family (Yang et al., 2025), chosen for its recency and the broad coverage of available model sizes, which facilitates robust scaling analysis. We use five sizes for scaling law fitting: 0.6B, 1.7B, 4B, 8B, and 14B. Additionally, Qwen3-32B is reserved to validate the extrapolation of our proposed laws.

Benchmarks. We evaluate diverse knowledge capabilities using 14 representative benchmarks aligned with the taxonomy defined in Section 3.1.

- **L1 (KM).** Assessed via benchmarks requiring exact facts recall, including TriviaQA (Joshi et al., 2017), Natural Questions (Kwiatkowski et al., 2019), WebQuestions (Berant et al., 2013), and the TREx and SQuAD subsets from LAMA (Petroni et al., 2019).
- **L2 (KA).** Evaluated on tasks focusing on flexible knowledge application, specifically Hellaswag (Zellers et al., 2019), Winogrande (Sakaguchi et al., 2021), MMLU (Hendrycks et al., 2021), and ARC-Easy (Clark et al., 2018).
- **L3 (KR).** Tested using multi-step reasoning datasets, namely StrategyQA (Geva et al., 2021a), OpenbookQA (Mihaylov et al., 2018), ARC-Challenge (Clark et al., 2018), GSM8K (Cobbe et al., 2021), and MathQA (Amini et al., 2019).

Quantization Strategy. Establishing robust scaling laws requires systematic sweeps over multiple quantization variables. We employ GPTQ (Frantar et al., 2023) because it is the most widely adopted weight-only PTQ method, and its mature libraries readily support the flexible configurations essential

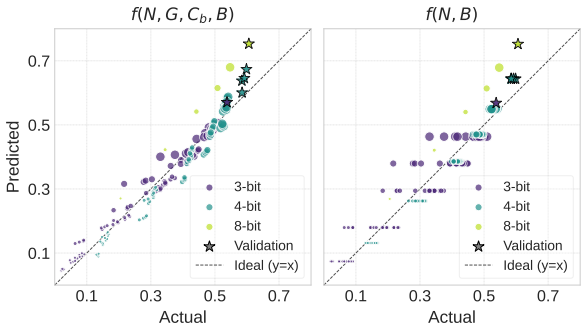


Figure 4: Goodness-of-fit: Predicted vs. actual adjusted accuracy for (Left) our proposed four-factor law (N, B, C_b, G) and (Right) the baseline (N, B). Points are colored by bit-width (B) and sized by model size (N). Stars $(\star)$ denote the validation data (Qwen3-32B). Dashed line represents ideal prediction.

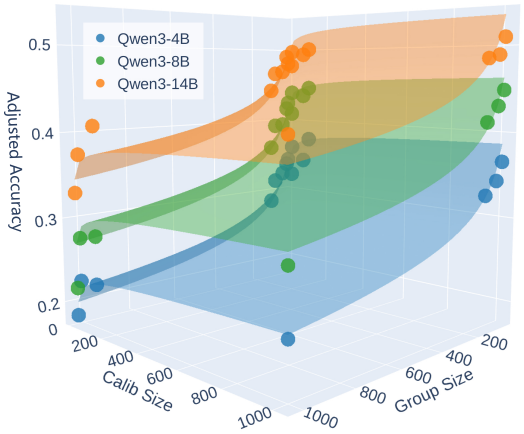


Figure 5: Performance surface of the General Scaling Law in the 3-bit region ($\text{Acc}_{\text{adj}} = \exp\left[-966.56 \cdot N^{-0.322}(\log_2 C_b)^{-0.103} G^{0.117}\right]$, Adj.$R^2_{\mathcal{O}} = 0.97$). Points represent empirical data.

for our analysis. In contrast, implementations of alternative methods (e.g., AWQ (Lin et al., 2024), QuIP (Chee et al., 2023)) often restrict accessible bit-widths or architectures. We apply a targeted sampling strategy to different compression zones. In the effective compression zone (3/4-bit), we execute a full grid search ($C_b \in \{8, 32, 128, 1024\}$, $G \in \{32, 64, 128, 1024\}$) to capture fine-grained sensitivities. Conversely, 8-bit configurations are fixed ($C_b = 128, G = 128$) due to marginal variance, and 2-bit is excluded from overall fitting to strictly preserve power-law assumptions.

4.2 Validation of the Unified Scaling Law

We first validate our unified scaling law on aggregated performance across all knowledge levels, offering an overall view of how PTQ factors influence general model performance.

4.2.1 Goodness-of-Fit and Ablation Analysis

We perform an ablation study to quantify the contribution of each factor. The results, summarized in Table 1 and visualized in Figure 4, reveal several key insights regarding factor importance:

(1) The comprehensive model achieves superior fit. The full four-factor model yields the highest Adj.$R^2_{\mathcal{O}}$ of 0.9475, indicating robust predictive capability. As shown in Figure 4 (Left), empirical data points tightly cluster around the ideal diagonal, while the held-out large-scale models (stars) validate extrapolation potential.

(2) Foundational role of N and B. The baseline model considering only model size (N) and bit-width (B) achieves a respectable foundation (Adj.$R^2_{\mathcal{O}} = 0.9125$). The large negative exponents for N (-0.359) and $\log_2 B$ (-1.067) confirm them as primary drivers for reducing the "loss" metric $(-\ln(\text{Acc}))$. However, the visible scatter in Figure 4 (Right) and the explanatory gap compared to the full formulation (0.91 vs. 0.95) indicate that neglecting granular parameters fails to capture critical performance variations.

(3) Significance of fine-grained factors (G and C_b). Combining group size (G) and calibration set size (C_b) bridges the performance gap. Notably, adding G alone boosts the $Adj.R^2_{\mathcal{O}}$ significantly to 0.9466, identifying it as a critical regulator. While adding C_b yields a marginal statistical gain overall (consistent with saturation effects), it remains indispensable for stability in low-bit scenarios, as discussed below.

4.2.2 Parameter Sensitivity in Low-Bit Scenarios

While the general model captures global trends, it obscures the nuanced behaviors in the critical 3-bit region. As illustrated in Figure 5, the "Effective Compression Zone" exhibits a dramatic sensitivity amplification to fine-grained parameters.

Specifically, when fitting solely to 3-bit data, the elasticity of calibration data (C_b) triples $(|-0.032| \rightarrow |-0.103|)$, confirming its shift from a diminishing factor to a critical constraint. Simultaneously, the group size (G) coefficient surges $(0.073 \rightarrow 0.117)$, indicating that coarse grouping becomes penalizing at lower precisions. These trends further intensify in the 2-bit region, as we will discuss in Section 4.3.2.

Formulation	Fitted Function	Adj. $R^2_{\mathcal{L}}$	Adj. $R^2_{\mathcal{O}}$
$f(N, B, C_b, G)$	$3.98 \times 10^3 \, N^{-0.359} \, (\log_2 B)^{-1.067} \, (\log_2 C_b)^{-0.032} \, G^{0.073}$	**0.9425**	**0.9475**
$f(N, B)$	$5.39 \times 10^3 \, N^{-0.359} \, (\log_2 B)^{-1.071}$	0.9038	0.9125
$f(N, B, G)$	$3.77 \times 10^3 \, N^{-0.359} \, (\log_2 B)^{-1.071} \, G^{0.073}$	0.9420	0.9466
$f(N, B, C_b)$	$5.69 \times 10^3 \, N^{-0.359} \, (\log_2 B)^{-1.067} \, (\log_2 C_b)^{-0.032}$	0.9041	0.9131

Table 1: Ablation analysis of the scaling law formulation modeling $-\ln(\text{Acc}_{\text{adj}})$. Adj. $R^2_{\mathcal{L}}$ and Adj. $R^2_{\mathcal{O}}$ denote the adjusted R^2 in the log-transformed and original accuracy spaces, respectively. The full formulation achieves the highest explanatory power, accurately capturing the variance across 165 fitted configurations.

Task Level	Const (A)	Scaling Exponents (Sensitivity)				Goodness-of-Fit	
		$\alpha(N)$	$\beta(B)$	$\gamma(C_b)$	$\delta(G)$	Adj. $R^2_{\mathcal{L}}$	Adj. $R^2_{\mathcal{O}}$
General	3.98×10^3	-0.359	-1.067	-0.032	0.073	0.9425	0.9475
L1: Memorization (KM)	2.08×10^3	-0.315	-0.964	**-0.040**	0.064	0.9341	0.9350
L2: Application (KA)	7.37×10^3	**-0.409**	-0.982	-0.023	0.069	0.9550	0.9626
L3: Reasoning (KR)	1.27×10^4	-0.405	**-1.356**	-0.034	**0.087**	0.9156	0.9218

Table 2: Fitted scaling parameters for task-stratified scaling laws. The model form is $-\ln(\text{Acc}_{\text{adj}}) = A \cdot N^{\alpha}(\log_2 B)^{\beta}(\log_2 C_b)^{\gamma} G^{\delta}$.

4.3 Task-Stratified Scaling Laws

While the general scaling law provides a macroscopic view, it inevitably masks the distinct scaling behaviors of different knowledge capabilities. To dissect these nuances, we derive separate scaling laws for the three knowledge levels: knowledge memorization, application, and reasoning. We fit the full four-variable formulation to each task level independently. Detailed ablation studies for each level are provided in Appendix C.1.

4.3.1 Heterogeneous Sensitivity Analysis

Table 2 details the fitted parameters for each knowledge level (standard errors and 95% confidence intervals are provided in Appendix C.2 to confirm statistical significance). As shown, all stratified formulations achieve high goodness-of-fit, confirming the universality of the proposed power-law formulation. However, a cross-comparison of the exponents reveals divergent sensitivities to quantization.

(1) Reasoning (KR) is Precision-Critical. L3 tasks exhibit the highest sensitivity to bit-width ($\beta = -1.356$) and group granularity ($\delta = 0.087$). Notably, the bit-width sensitivity exceeds that of KM and KA by nearly 40%. This supports the hypothesis that reasoning relies on long-chain logical deductions, where quantization noise accumulates at each step ("error propagation"), rendering the process highly fragile to precision loss.

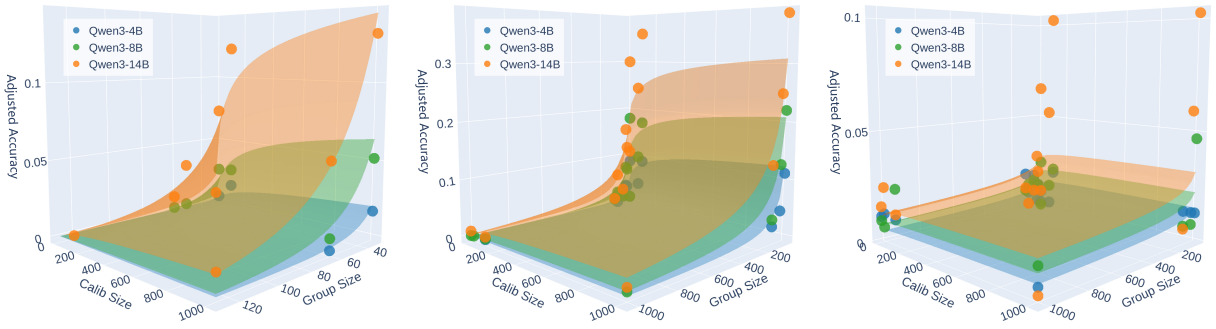
(2) Application (KA) is Scale-Responsive. In terms of model size, KA exhibits a high scaling exponent ($\alpha = -0.409$), contrasting with the notably lower exponent of KM ($\alpha = -0.315$). This implies that while memorization capacity saturates faster, application benefits significantly from scaling up, consistent with the "emergence" properties often observed in high-level cognitive tasks.

(3) Memorization (KM) is Calibration-Sensitive. L1 tasks show a pronounced sensitivity to calibration data ($\gamma = -0.040$), nearly double that of the more robust KA tasks. We attribute this to KM's reliance on precise activation alignment to trigger Key-Value pairs in FFN layers (Geva et al., 2021b). Unlike KA tasks, which rely on generalized patterns robust to numerical shifts, KM's "exact lookup" mechanism is susceptible to distribution shifts, necessitating richer calibration data.

4.3.2 The "Phase Transition" at 2-bit

We characterize the entry into the 2-bit region as a critical "Phase Transition," where the scaling behavior diverges sharply depending on model size and task type.

(1) Systemic Collapse in Small-Scale Models. For models with $N < 2$B, we observe a universal performance collapse across all tasks. Scaling laws fail to converge (Adj.$R^2_{\mathcal{O}} < 0$). Consequently, PTQ tuning becomes ineffective, as the model lacks the fundamental capacity to retain utility.



(a) Memorization (KM) (b) Application (KA) (c) Reasoning (KR)

Figure 6: Fitted performance surfaces under 2-bit quantization ($N \geq 4$B). (a) KM and (b) KA retain robust scaling behaviors with high goodness-of-fit (Adj.$R^2_{\mathcal{O}} \approx 0.91$ and 0.87, respectively), exhibiting pronounced sensitivity to G and C_b. In contrast, (c) KR exhibits a flat surface with poor fit (Adj.$R^2_{\mathcal{O}} \approx 0.22$), indicating a structural collapse of reasoning capabilities regardless of configuration adjustments.

(2) Capability Recovery in Large-Scale Models. In contrast, larger models ($N \geq 4$B) can maintain capabilities, but with certain conditions. As shown in Figure 6, while reasoning (KR) fails completely, memorization (KM) and application (KA) are effectively recovered if fine-grained parameters are optimized. Specifically, the scaling exponents for G surges from ~ 0.07 (General) to ~ 0.60 (KM) and ~ 0.33 (KA), and calibration dependence intensifies ($\gamma \approx -0.58$). This implies that using smaller group sizes and sufficient calibration data is no longer optional, but essential for preventing failure in the 2-bit region.

4.4 Cross-Architecture Validation on Llama-3

To verify the universality of our framework beyond Qwen, we extend the evaluation to the Llama-3 family (1B, 3B, 8B) (Grattafiori et al., 2024) using consistent quantization strategy and benchmarks. We assess a representative subset of 42 configurations within the effective compression zone.

Universality of the Scaling Framework. As shown in Table 3, fitting the four-factor formulation yields exceptional goodness-of-fit, with Adj.$R^2_{\mathcal{O}}$ exceeding 0.92 across all knowledge levels. This confirms that our multiplicative power-law formulation captures fundamental quantization dynamics independent of architecture. Appendix D.1 provides further visualizations and statistical validation for these results.

Consistency of Knowledge Sensitivities. Crucially, the fitted coefficients reinforce the distinct sensitivities observed in Qwen3:

- *Precision Critical*: KR remains the most fragile, showing the highest sensitivity to both bit-width (β) and group size (δ).

Task	Const(A)	$\alpha(N)$	$\beta(B)$	$\gamma(C_b)$	$\delta(G)$	Adj. $R^2_{\mathcal{O}}$
General	2.91e3	-0.333	-1.501	-0.056	0.072	0.9595
L1: KM	5.32e2	-0.249	-1.596	-0.060	0.074	0.9622
L2: KA	2.19e4	**-0.447**	-1.462	-0.045	0.073	0.9709
L3: KR	9.66e3	-0.373	**-1.645**	**-0.071**	**0.080**	0.9277

Table 3: Fitted scaling parameters for Llama-3 family.

- *Scale Responsive*: KA exhibits the highest scaling exponent (α) while maintaining the lowest sensitivity to quantization coefficients. This confirms it benefits most from model scaling and is relatively robust to quantization.
- *Calibration Sensitive*: Both KM and KR exhibit heightened sensitivity to calibration data compared to the robust KA. This reinforces our finding that while KA is largely scale-driven, retaining memorization and reasoning capabilities necessitates high-quality quantization parameters.

5 Conclusion

In this work, we formulate Task-Stratified Knowledge Scaling Laws, integrating model size, bit-width, and crucial fine-grained factors (group size and calibration set size) into a unified framework. Validated on 293 diverse configurations, our framework demonstrates strong fit and cross-architecture consistency. We identify distinct sensitivities across knowledge capabilities: reasoning is precision-critical, application is scale-responsive, and memorization is calibration-sensitive. Furthermore, we emphasize that under low-bit quantization, optimizing fine-grained factors is essential to prevent performance collapse.

Limitations

Our study primarily establishes task-stratified PTQ scaling laws for representative dense Transformer architectures under weight-only quantization. While the proposed framework covers diverse knowledge capabilities, future research could extend these laws to other quantization paradigms (e.g., activation quantization) and alternative architectures, such as Mixture-of-Experts (MoE).

Acknowledgments

This work was supported by Beijing Natural Science Foundation (L243006), the National Natural Science Foundation of China (No.62406321), the independent research project of the Key Laboratory of Cognition and Decision Intelligence for Complex Systems and CIPS-SMP-Zhipu Large Model Fund.

References

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards Interpretable Math Word Problem Solving with Operation-Based Formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic Parsing on Freebase from Question-Answer Pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher M. De Sa. 2023. [QuIP: 2-Bit Quantization of Large Language Models With Guarantees. In *Advances in Neural Information Processing Systems*, volume 36, pages 4396–4429.

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 preprint*. ArXiv: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. *arXiv preprint*. ArXiv:2110.14168 [cs].

Reena Elangovan, Charbel Sakr, Anand Raghunathan, and Brucek Khailany. 2025. BCQ: Block Clustered Quantization for 4-bit (W4A4) LLM Inference. *arXiv preprint*. ArXiv:2502.05376.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and Improving Consistency in Pretrained Language Models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. arXiv. ArXiv:2210.17323.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021a. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021b. Transformer Feed-Forward Layers Are Key-Value Memories. arXiv.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 181 others. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ADS Bibcode: 2024arXiv240721783G.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating Large Language Models: A Comprehensive Survey. *arXiv preprint*. ArXiv:2310.19736.

Jahid Hasan. 2024. Optimizing Large Language Models through Quantization: A Comparative Analysis of PTQ and QAT Techniques. *arXiv preprint*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *arXiv preprint*. ArXiv:2009.03300 [cs].

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training Compute-Optimal Large Language Models. *arXiv preprint*. ArXiv:2203.15556.

Thomas Huber and Christina Niklaus. 2025. LLMs meet Bloom's Taxonomy: A Cognitive View on

Large Language Model Evaluations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5211–5246, Abu Dhabi, UAE. Association for Computational Linguistics.

Yixin Ji, Yang Xiang, Juntao Li, Qingrong Xia, Ping Li, Xinyu Duan, Zhefeng Wang, and Min Zhang. 2024. Beware of Calibration Data for Pruning Large Language Models. *arXiv preprint*. ArXiv:2410.17711.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *arXiv preprint*. ArXiv:1705.03551 [cs].

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint*. ArXiv:2001.08361.

Jakub Krajewski, Amitis Shidani, Dan Busbridge, Sam Wiseman, and Jason Ramapuram. 2025. Revisiting the Scaling Properties of Downstream Metrics in Large Language Model Training. *arXiv preprint*.

David R. Krathwohl. 2002. A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice*, 41(4):212–218.

Tanishq Kumar, Zachary Ankner, Benjamin F. Spector, Blake Bordelon, Niklas Muennighoff, Mansheej Paul, Cengiz Pehlevan, Christopher Ré, and Aditi Raghunathan. 2025. Scaling Laws for Precision. arXiv.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Jiedong Lang, Zhehao Guo, and Shuyu Huang. 2024. A Comprehensive Study on Quantization Techniques for Large Language Models. *arXiv preprint*. ArXiv:2411.02530.

Shiyao Li, Xuefei Ning, Luning Wang, Tengxuan Liu, Xiangsheng Shi, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. 2024. Evaluating Quantized Large Language Models. arXiv.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration. *Proceedings of Machine Learning and Systems*, 6:87–100.

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. *arXiv preprint*. ArXiv:2504.04823.

Kelly Marchisio, Saurabh Dash, Hongyu Chen, Dennis Aumiller, Ahmet Üstün, Sara Hooker, and Sebastian Ruder. 2024. How Does Quantization Affect Multilingual LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15928–15947, Miami, Florida, USA. Association for Computational Linguistics.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. *arXiv preprint*. ArXiv:1809.02789 [cs].

Xu Ouyang, Tao Ge, Thomas Hartvigsen, Zhisong Zhang, Haitao Mi, and Dong Yu. 2024. Low-Bit Quantization Favors Undertrained LLMs: Scaling Laws for Quantized LLMs with 100T Training Tokens. *arXiv preprint*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. WinoGrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Rylan Schaeffer, Hailey Schoelkopf, Brando Miranda, Gabriel Mukobi, Varun Madan, Adam Ibrahim, Herbie Bradley, Stella Biderman, and Sanmi Koyejo. 2025. Why Has Predicting Downstream Capabilities of Frontier AI Models with Scale Remained Elusive? *arXiv preprint*. ArXiv:2406.04391 [cs].

Ayan Sengupta, Siddhant Chaudhary, and Tanmoy Chakraborty. 2025. Compression Laws for Large Language Models. *arXiv preprint*.

Xingwu Sun, Shuaipeng Li, Ruobing Xie, Weidong Han, Kan Wu, Zhen Yang, Yixing Li, An Wang, Shuai Li, Jinbao Xue, Yu Cheng, Yangyu Tao, Zhanhui Kang, Chengzhong Xu, Di Wang, and Jie Jiang. 2025. Scaling Laws for Floating Point Quantization Training. arXiv. ArXiv:2501.02423.

Lintang Sutawika, Hailey Schoelkopf, Leo Gao, Baber Abbasi, Stella Biderman, Jonathan Tow, ben fattori, Charles Lovering, farzanehnakhaee70, Jason Phang, Anish Thite, Fazz, Thomas Wang, Niklas Muennighoff, Aflah, sdtblck, nopperl, gakada, tttyuntian, and 11 others. 2025. EleutherAI/lm-evaluation-harness: v0.4.9.

Mengru Wang, Yunzhi Yao, Ziwen Xu, Shuofei Qiao, Shumin Deng, Peng Wang, Xiang Chen, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2024. Knowledge Mechanisms in Large Language Models: A Survey and Perspective. arXiv.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed 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*, volume 35, pages 24824–24837.

Miles Williams and Nikolaos Aletras. 2024. On the Impact of Calibration Data in Post-training Quantization and Pruning. ArXiv:2311.09755.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zifei Xu, Alexander Lan, Wanzin Yazar, Tristan Webb, Sayeh Sharify, and Xin Wang. 2024. Scaling Laws for Post Training Quantized Large Language Models. *arXiv preprint*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint*. ArXiv:2505.09388 [cs].

Zhewei Yao, Xiaoxia Wu, Cheng Li, Stephen Youn, and Yuxiong He. 2023. A Comprehensive Study on Post-Training Quantization for Large Language Models. *arXiv preprint*.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chunyang Li, Zheyuan Zhang, Yushi Bai, Yantao Liu, Amy Xin, Nianyi Lin, Kaifeng Yun, Linlu Gong, Jianhui Chen, and 16 others. 2024. KoLA: Carefully Benchmarking World Knowledge of Large Language Models. arXiv.

Bozena Zdaniuk. 2014. Ordinary Least-Squares (OLS) Model. In *Encyclopedia of Quality of Life and Well-Being Research*, pages 4515–4517. Springer, Dordrecht.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? *arXiv preprint*. ArXiv:1905.07830.

Zhao Zhang, Yangcheng Gao, Jicong Fan, Zhongqiu Zhao, Yi Yang, and Shuicheng Yan. 2025. SelectQ: Calibration Data Selection for Post-training Quantization. *Machine Intelligence Research*.

Jiaqi Zhao, Ming Wang, Miao Zhang, Yuzhang Shang, Xuebo Liu, Yaowei Wang, Min Zhang, and Liqiang Nie. 2025. Benchmarking Post-Training Quantization in LLMs: Comprehensive Taxonomy, Unified Evaluation, and Comparative Analysis. *arXiv preprint*. ArXiv:2502.13178.

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 Experimental Details

This appendix provides supplementary details to support the reproducibility of our experiments, covering implementation specifics, benchmark stratification rationale, and full model configurations.

A.1 Implementation and Evaluation Setup

Quantization Implementation. Experiments are conducted using the Hugging Face Transformers library (Wolf et al., 2020), with GPTQ implemented via the GPTQModel library. We employ default hyperparameters unless otherwise specified. To establish a domain-agnostic baseline, we chose the universal C4 dataset (Raffel et al., 2020) as our calibration corpus. Samples are randomly drawn with a fixed sequence length of 2048.

Evaluation Framework. We utilize the Language Model Evaluation Harness (`lm-eval`, v0.4.9) framework (Sutawika et al., 2025) for standardized testing. Most tasks are evaluated in a 5-shot setting. For multiple-choice tasks, we report the "acc_norm" (accuracy normalized by choice length) to mitigate length bias, while generative tasks use "exact_match". A specific exception is the TREx benchmark (part of LAMA). We strictly control the prompt variance: for each of the 39 relation types, we select the single prompt template from the Pararel dataset (Elazar et al., 2021) where the object [Y] is positioned at the end of the sentence. Performance for TREx is reported using the Precision@5 (P@5) metric.

A.2 Benchmarks Mapping and Statistics

Table 4 provides a comprehensive mapping of the 14 benchmarks to our cognitive taxonomy, along with their statistical details.

A.3 Full Experimental Configurations

To ensure reproducibility and transparency, Table 9 enumerates all 293 experimental configurations evaluated in this study, covering the Main (scaling fit), Validation, and Generalization groups.

A.4 Numerical Stabilization in Regression

To ensure numerical stability during regression, we implement filtering rules for the transformation $\ln(-\ln(\text{Acc}_{\text{adj}}))$. Because this term is undefined for $\text{Acc}_{\text{adj}} \leq 0$ and approaches a mathematical singularity as $\text{Acc}_{\text{adj}} \to 0^+$, we establish a lower-bound threshold of $\epsilon = 0.01$. Configurations yielding $\text{Acc}_{\text{adj}} \leq 0.01$ are considered "collapsed to random guessing" and are excluded. At this boundary, the transformation yields $\ln(-\ln(0.01)) \approx 1.527$, ensuring stable computation.

In our main experiments on the Qwen3 family, exactly 6 configurations trigger this filter. All of them share the most aggressive compression setting: the smallest model size ($N = 0.6$B) at 3-bit weight precision with the coarsest group size ($G = 1024$).

B Definition of Adjusted R^2

While the standard coefficient of determination (R^2) measures the proportion of variance explained by the model, it tends to increase when more variables are added, regardless of their actual predictive power. To provide a robust assessment that accounts for model complexity, we employ the Adj. R^2 (denoted as R^2_{adj}).

First, the standard R^2 is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{4}$$

where y_i is the true value, $\hat{y}_i$ is the model prediction, and $\bar{y}$ is the empirical mean of the true values.

The Adj. R^2 is then calculated as:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}, \tag{5}$$

where n is the sample size (number of observations) and p is the number of predictors (independent variables) in the fitted model. Unlike standard R^2, the Adj. R^2 penalizes the inclusion of non-informative parameters, ensuring that the reported goodness-of-fit accurately reflects the model's explanatory power relative to its complexity.

C Detailed Ablation and Statistical Significance for Qwen3

C.1 Ablation Study on Fine-Grained Factors

To validate the necessity of including Group Size (G) and Calibration Set Size (C_b) in our task-stratified scaling laws, we conduct ablation studies across the three knowledge capabilities (KM, KA, KR), as detailed in Table 5.

The results consistently highlight two patterns. First, adding G significantly enhances the goodness-of-fit across all tasks (e.g., KR improves from $0.8775 \to 0.9212$), confirming quantization granularity as a universal determinant. Second, the

Level	Benchmark	Domain	Type	Metric	Baseline	Size	Characteristics
KM	TriviaQA	Trivia & Web	Gen	EM	≈ 0	17,944	**Factual Recall**: Requires recalling precise entities (names, dates) from the internal knowledge base without complex contextual transformation.
	Natural Questions	Wikipedia	Gen	EM	≈ 0	3,610	
	WebQuestions	KB (Freebase)	Gen	EM	≈ 0	2,032	
	TREx (LAMA)	KB (Wikidata)	Gen	P@5	≈ 0	27,610	
	SQuAD (LAMA)	Wikipedia	Gen	P@5	≈ 0	212	
KA	MMLU	57 Subjects	MC (4)	Acc	0.25	14,042	**Flexible Application**: Requires understanding contexts and applying internalized knowledge for specific scenarios, emphasizing flexible utilization rather than strict facts lookup.
	Hellaswag	Commonsense	MC (4)	Acc	0.25	10,042	
	Winogrande	Commonsense	MC (2)	Acc	0.50	1,267	
	ARC-Easy	Science (Basic)	MC (4)	Acc	0.25	2,376	
KR	ARC-Challenge	Science (Hard)	MC (4)	Acc	0.25	1,172	**Multi-step Reasoning**: Requires constructing sequential logical chains (mathematical derivation or multi-hop logic).
	StrategyQA	Open-Domain	MC (2)	Acc	0.50	2,289	
	OpenbookQA	Science & Common	MC (4)	Acc	0.25	500	
	MathQA	Math	MC (5)	Acc	0.20	2,985	
	GSM8K	Math	Gen	EM	≈ 0	1,319	

Table 4: Detailed statistics and cognitive mapping of benchmarks. **Type** denotes the task format (Generative vs. Multiple-Choice). **Metric** denotes Exact Match (EM), Accuracy (Acc), or Precision@5 (P@5). **Characteristics** justifies the classification by highlighting the underlying task nature.

Formulation	Fitted Function	Adj. $R^2_{\mathcal{L}}$	Adj. $R^2_{\mathcal{O}}$
L1: Knowledge Memorization (KM)			
$f(N, B, C_b, G)$	$2.08 \times 10^3\, N^{-0.315}\, (\log_2 B)^{-0.964}\, (\log_2 C_b)^{-0.040}\, G^{0.064}$	**0.9341**	**0.9350**
$f(N, B)$	$2.51 \times 10^3\, N^{-0.313}\, (\log_2 B)^{-0.959}$	0.8946	0.8993
$f(N, B, G)$	$1.95 \times 10^3\, N^{-0.315}\, (\log_2 B)^{-0.969}\, G^{0.064}$	0.9328	0.9326
$f(N, B, C_b)$	$2.69 \times 10^3\, N^{-0.313}\, (\log_2 B)^{-0.953}\, (\log_2 C_b)^{-0.040}$	0.8957	0.9015
L2: Knowledge Application (KA)			
$f(N, B, C_b, G)$	$7.37 \times 10^3\, N^{-0.409}\, (\log_2 B)^{-0.982}\, (\log_2 C_b)^{-0.023}\, G^{0.069}$	**0.9550**	**0.9626**
$f(N, B)$	$9.89 \times 10^3\, N^{-0.409}\, (\log_2 B)^{-0.986}$	0.9276	0.9362
$f(N, B, G)$	$7.09 \times 10^3\, N^{-0.409}\, (\log_2 B)^{-0.986}\, G^{0.069}$	0.9549	0.9624
$f(N, B, C_b)$	$1.03 \times 10^4\, N^{-0.409}\, (\log_2 B)^{-0.982}\, (\log_2 C_b)^{-0.023}$	0.9275	0.9362
L3: Knowledge Reasoning (KR)			
$f(N, B, C_b, G)$	$1.27 \times 10^4\, N^{-0.405}\, (\log_2 B)^{-1.356}\, (\log_2 C_b)^{-0.034}\, G^{0.087}$	**0.9156**	**0.9218**
$f(N, B)$	$1.55 \times 10^4\, N^{-0.398}\, (\log_2 B)^{-1.330}$	0.8738	0.8775
$f(N, B, G)$	$1.20 \times 10^4\, N^{-0.405}\, (\log_2 B)^{-1.361}\, G^{0.087}$	0.9154	0.9212
$f(N, B, C_b)$	$1.64 \times 10^4\, N^{-0.398}\, (\log_2 B)^{-1.325}\, (\log_2 C_b)^{-0.034}$	0.8738	0.8779

Table 5: Ablation analysis for task-stratified scaling laws across three knowledge levels. Including fine-grained factors (G, C_b) consistently improves goodness-of-fit. The model form is $-\ln(\text{Acc}_{\text{adj}}) = A \cdot N^{\alpha}(\log_2 B)^{\beta}(\log_2 C_b)^{\gamma} G^{\delta}$.

impact of C_b varies by task nature: it yields negligible improvement for the robust Knowledge Application (KA) task, but provides detectable gains for Knowledge Memorization (KM) and Reasoning (KR). This empirical evidence reinforces the sensitivity hierarchy discussed in Section 4.3, where specific capabilities rely more heavily on precise distribution alignment.

	Qwen3		Llama-3
Task Level	**Fit**	**Validation**	**Fit**
General	0.0267	0.0630	0.0171
L1: KM	0.0219	0.0946	0.0147
L2: KA	0.0254	0.0292	0.0172
L3: KR	0.0361	0.0555	0.0227

Table 6: Mean Absolute Error (MAE) of the predicted accuracy. 'Validation' denotes the held-out Qwen3-32B.

C.2 Statistical Significance of Scaling Exponents

To rigorously validate that the observed sensitivities in the Qwen3 family are not artifacts of random variance, we compute the Standard Errors (SE) and 95% Confidence Intervals (CI) for all fitted exponents ($\alpha, \beta, \gamma, \delta$). The regressions were evaluated using the `statsmodels` library. An exponent is considered statistically significant if its 95% CI strictly excludes zero.

The results, presented in Table 7, align with our qualitative observations. Across all task levels, the primary drivers (N, B, G) are highly significant. Notably, the coefficient for calibration set size ($\gamma(C_b)$) is strictly negative for KM (e.g., $[-0.078, -0.002]$), confirming its calibration-sensitive nature, whereas it crosses zero for KA ($[-0.063, +0.016]$), indicating that application capabilities rely fundamentally on model scale rather than granular calibration alignment.

D Cross-Architecture Validation and Predictive Robustness

D.1 Scaling Law Analysis on Llama-3

Experimental Configuration Details. For the Llama-3 generalization experiments, we analyze 42 representative quantization configurations in the effective compression zone (3-bit and 4-bit). To efficiently traverse the hyperparameter space, we adopt a controlled grid search strategy: (1) Fixed Group Size ($G = 128$) with varying Calibration Set Sizes ($C_b \in \{8, 32, 128, 1024\}$); and (2) Fixed Calibration Set Size ($C_b = 128$) with varying Group Sizes ($G \in \{32, 64, 128, 1024\}$). This setup ensures coverage of key sensitivity thresholds while maintaining computational feasibility.

Visualization. Figure 7 compares the predicted and actual adjusted accuracy across all three knowledge levels for the Llama-3 family, visually confirming the high goodness-of-fit reported in the main text.

Statistical Significance. As shown in Table 8, the statistical behaviors strictly mirror those of Qwen3. The primary factors remain highly significant (95% CIs exclude zero). Crucially, the unique sensitivity of Knowledge Memorization to calibration is preserved ($\gamma(C_b)$ CI is $[-0.113, -0.007]$), confirming that this calibration dependence is an architecture-agnostic property of memorization tasks.

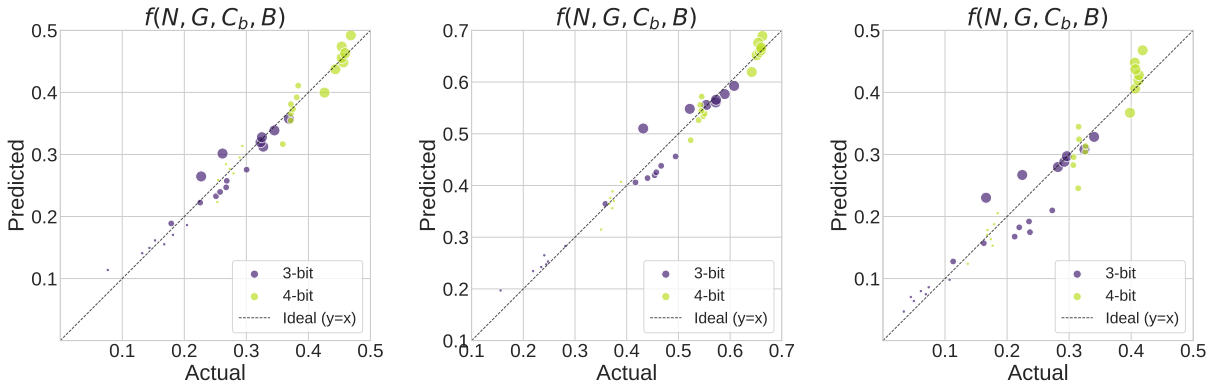
D.2 Predictive Quality and Extrapolation

While the Adjusted R^2 measures the explained variance, deployment decisions often require evaluating the absolute predictive error. To this end, we compute the Mean Absolute Error (MAE) across both the Qwen3 and Llama-3 families.

As shown in Table 6, the MAE remains remarkably low. Crucially, scale-extrapolation on the held-out Qwen3-32B validation model is highly reliable, with prediction errors for KA and KR strictly bounded to $\approx 2.9\%$ and $\approx 5.5\%$ respectively.

Task Level	Metric	$\alpha(N)$	$\beta(B)$	$\gamma(C_b)$	$\delta(G)$
General	Est. $\pm$ SE	-0.359 ± 0.008	-1.067 ± 0.062	-0.033 ± 0.021	$+0.074 \pm 0.007$
	95% CI	$[-0.374, -0.344]$	$[-1.189, -0.944]$	$[-0.073, +0.008]$	$[+0.060, +0.087]$
L1: KM	Est. $\pm$ SE	-0.315 ± 0.007	-0.964 ± 0.058	$\mathbf{-0.040 \pm 0.019}$	$+0.065 \pm 0.007$
	95% CI	$[-0.329, -0.301]$	$[-1.078, -0.849]$	$\mathbf{[-0.078, -0.002]}$	$[+0.051, +0.078]$
L2: KA	Est. $\pm$ SE	$\mathbf{-0.409 \pm 0.007}$	-0.982 ± 0.061	-0.023 ± 0.020	$+0.069 \pm 0.007$
	95% CI	$\mathbf{[-0.424, -0.395]}$	$[-1.102, -0.863]$	$[-0.063, +0.016]$	$[+0.055, +0.082]$
L3: KR	Est. $\pm$ SE	-0.405 ± 0.011	$\mathbf{-1.356 \pm 0.085}$	-0.034 ± 0.028	$\mathbf{+0.087 \pm 0.010}$
	95% CI	$[-0.425, -0.384]$	$\mathbf{[-1.523, -1.189]}$	$[-0.089, +0.022]$	$\mathbf{[+0.068, +0.107]}$

Table 7: Statistical significance of the fitted scaling exponents for the Qwen3 family. CIs that strictly exclude zero indicate statistical significance.



(a) Memorization (KM) (b) Application (KA) (c) Reasoning (KR)

Figure 7: Goodness-of-fit visualization for Llama-3 family. The scatter plots compare the predicted adjusted accuracy (y-axis) against the actual empirical values (x-axis) for (a) Memorization, (b) Application, and (c) Reasoning. The close alignment with the dashed diagonal line ($y = x$) indicates high predictive accuracy. Point size corresponds to model size (1B, 3B, 8B), and color indicates bit-width.

Task Level	Metric	$\alpha(N)$	$\beta(B)$	$\gamma(C_b)$	$\delta(G)$
General	Est. $\pm$ SE	-0.333 ± 0.013	-1.501 ± 0.091	-0.056 ± 0.030	$+0.072 \pm 0.011$
	95% CI	$[-0.360, -0.307]$	$[-1.685, -1.318]$	$[-0.117, +0.005]$	$[+0.050, +0.093]$
L1: KM	Est. $\pm$ SE	-0.249 ± 0.011	-1.596 ± 0.079	$\mathbf{-0.060 \pm 0.026}$	$+0.074 \pm 0.009$
	95% CI	$[-0.273, -0.226]$	$[-1.756, -1.437]$	$\mathbf{[-0.113, -0.007]}$	$[+0.055, +0.093]$
L2: KA	Est. $\pm$ SE	$\mathbf{-0.447 \pm 0.014}$	-1.462 ± 0.095	-0.045 ± 0.032	$+0.073 \pm 0.011$
	95% CI	$\mathbf{[-0.475, -0.419]}$	$[-1.655, -1.268]$	$[-0.109, +0.019]$	$[+0.050, +0.096]$
L3: KR	Est. $\pm$ SE	-0.373 ± 0.019	$\mathbf{-1.645 \pm 0.131}$	-0.071 ± 0.044	$\mathbf{+0.080 \pm 0.016}$
	95% CI	$[-0.411, -0.334]$	$\mathbf{[-1.911, -1.379]}$	$[-0.159, +0.018]$	$\mathbf{[+0.048, +0.111]}$

Table 8: Statistical significance of the fitted scaling exponents for the Llama-3 family, demonstrating cross-architecture consistency.

Table 9: All configurations of experiments. The Type column classifies the 293 data points into three roles: Fit (245 Qwen3 configurations for fitting scaling coefficients), Val (6 held-out Qwen3-32B configurations for extrapolation validation), and Gen (42 Llama-3 configurations for cross-architecture generalization).

No.	Model	N	B	G	C_b	Type	No.	Model	N	B	G	C_b	Type
0	Qwen3-0.6B	440,467,456	8	128	128	Fit	1	Qwen3-0.6B	440,467,456	4	32	8	Fit
2	Qwen3-0.6B	440,467,456	4	32	32	Fit	3	Qwen3-0.6B	440,467,456	4	32	128	Fit
4	Qwen3-0.6B	440,467,456	4	32	1024	Fit	5	Qwen3-0.6B	440,467,456	4	64	8	Fit
6	Qwen3-0.6B	440,467,456	4	64	32	Fit	7	Qwen3-0.6B	440,467,456	4	64	128	Fit
8	Qwen3-0.6B	440,467,456	4	64	1024	Fit	9	Qwen3-0.6B	440,467,456	4	128	8	Fit
10	Qwen3-0.6B	440,467,456	4	128	32	Fit	11	Qwen3-0.6B	440,467,456	4	128	128	Fit
12	Qwen3-0.6B	440,467,456	4	128	1024	Fit	13	Qwen3-0.6B	440,467,456	4	1024	8	Fit
14	Qwen3-0.6B	440,467,456	4	1024	32	Fit	15	Qwen3-0.6B	440,467,456	4	1024	128	Fit
16	Qwen3-0.6B	440,467,456	4	1024	1024	Fit	17	Qwen3-0.6B	440,467,456	3	32	8	Fit
18	Qwen3-0.6B	440,467,456	3	32	32	Fit	19	Qwen3-0.6B	440,467,456	3	32	128	Fit
20	Qwen3-0.6B	440,467,456	3	32	1024	Fit	21	Qwen3-0.6B	440,467,456	3	64	8	Fit
22	Qwen3-0.6B	440,467,456	3	64	32	Fit	23	Qwen3-0.6B	440,467,456	3	64	128	Fit
24	Qwen3-0.6B	440,467,456	3	64	1024	Fit	25	Qwen3-0.6B	440,467,456	3	128	8	Fit
26	Qwen3-0.6B	440,467,456	3	128	32	Fit	27	Qwen3-0.6B	440,467,456	3	128	128	Fit
28	Qwen3-0.6B	440,467,456	3	128	1024	Fit	29	Qwen3-0.6B	440,467,456	3	1024	8	Fit
30	Qwen3-0.6B	440,467,456	3	1024	32	Fit	31	Qwen3-0.6B	440,467,456	3	1024	128	Fit
32	Qwen3-0.6B	440,467,456	3	1024	1024	Fit	33	Qwen3-0.6B	440,467,456	2	32	8	Fit
34	Qwen3-0.6B	440,467,456	2	32	32	Fit	35	Qwen3-0.6B	440,467,456	2	32	128	Fit
36	Qwen3-0.6B	440,467,456	2	32	1024	Fit	37	Qwen3-0.6B	440,467,456	2	64	8	Fit
38	Qwen3-0.6B	440,467,456	2	64	32	Fit	39	Qwen3-0.6B	440,467,456	2	64	128	Fit
40	Qwen3-0.6B	440,467,456	2	64	1024	Fit	41	Qwen3-0.6B	440,467,456	2	128	8	Fit
42	Qwen3-0.6B	440,467,456	2	128	32	Fit	43	Qwen3-0.6B	440,467,456	2	128	128	Fit
44	Qwen3-0.6B	440,467,456	2	128	1024	Fit	45	Qwen3-0.6B	440,467,456	2	1024	8	Fit
46	Qwen3-0.6B	440,467,456	2	1024	32	Fit	47	Qwen3-0.6B	440,467,456	2	1024	128	Fit
48	Qwen3-0.6B	440,467,456	2	1024	1024	Fit	49	Qwen3-1.7B	1,409,410,048	8	128	128	Fit
50	Qwen3-1.7B	1,409,410,048	4	32	8	Fit	51	Qwen3-1.7B	1,409,410,048	4	32	32	Fit
52	Qwen3-1.7B	1,409,410,048	4	32	128	Fit	53	Qwen3-1.7B	1,409,410,048	4	32	1024	Fit
54	Qwen3-1.7B	1,409,410,048	4	64	8	Fit	55	Qwen3-1.7B	1,409,410,048	4	64	32	Fit
56	Qwen3-1.7B	1,409,410,048	4	64	128	Fit	57	Qwen3-1.7B	1,409,410,048	4	64	1024	Fit
58	Qwen3-1.7B	1,409,410,048	4	128	8	Fit	59	Qwen3-1.7B	1,409,410,048	4	128	32	Fit
60	Qwen3-1.7B	1,409,410,048	4	128	128	Fit	61	Qwen3-1.7B	1,409,410,048	4	128	1024	Fit
62	Qwen3-1.7B	1,409,410,048	4	1024	8	Fit	63	Qwen3-1.7B	1,409,410,048	4	1024	32	Fit
64	Qwen3-1.7B	1,409,410,048	4	1024	128	Fit	65	Qwen3-1.7B	1,409,410,048	4	1024	1024	Fit
66	Qwen3-1.7B	1,409,410,048	3	32	8	Fit	67	Qwen3-1.7B	1,409,410,048	3	32	32	Fit
68	Qwen3-1.7B	1,409,410,048	3	32	128	Fit	69	Qwen3-1.7B	1,409,410,048	3	32	1024	Fit
70	Qwen3-1.7B	1,409,410,048	3	64	8	Fit	71	Qwen3-1.7B	1,409,410,048	3	64	32	Fit
72	Qwen3-1.7B	1,409,410,048	3	64	128	Fit	73	Qwen3-1.7B	1,409,410,048	3	64	1024	Fit
74	Qwen3-1.7B	1,409,410,048	3	128	8	Fit	75	Qwen3-1.7B	1,409,410,048	3	128	32	Fit
76	Qwen3-1.7B	1,409,410,048	3	128	128	Fit	77	Qwen3-1.7B	1,409,410,048	3	128	1024	Fit
78	Qwen3-1.7B	1,409,410,048	3	1024	8	Fit	79	Qwen3-1.7B	1,409,410,048	3	1024	32	Fit
80	Qwen3-1.7B	1,409,410,048	3	1024	128	Fit	81	Qwen3-1.7B	1,409,410,048	3	1024	1024	Fit
82	Qwen3-1.7B	1,409,410,048	2	32	8	Fit	83	Qwen3-1.7B	1,409,410,048	2	32	32	Fit
84	Qwen3-1.7B	1,409,410,048	2	32	128	Fit	85	Qwen3-1.7B	1,409,410,048	2	32	1024	Fit
86	Qwen3-1.7B	1,409,410,048	2	64	8	Fit	87	Qwen3-1.7B	1,409,410,048	2	64	32	Fit
88	Qwen3-1.7B	1,409,410,048	2	64	128	Fit	89	Qwen3-1.7B	1,409,410,048	2	64	1024	Fit
90	Qwen3-1.7B	1,409,410,048	2	128	8	Fit	91	Qwen3-1.7B	1,409,410,048	2	128	32	Fit
92	Qwen3-1.7B	1,409,410,048	2	128	128	Fit	93	Qwen3-1.7B	1,409,410,048	2	128	1024	Fit
94	Qwen3-1.7B	1,409,410,048	2	1024	8	Fit	95	Qwen3-1.7B	1,409,410,048	2	1024	32	Fit
96	Qwen3-1.7B	1,409,410,048	2	1024	128	Fit	97	Qwen3-1.7B	1,409,410,048	2	1024	1024	Fit
98	Qwen3-4B	3,633,511,936	8	128	128	Fit	99	Qwen3-4B	3,633,511,936	4	32	8	Fit
100	Qwen3-4B	3,633,511,936	4	32	32	Fit	101	Qwen3-4B	3,633,511,936	4	32	128	Fit
102	Qwen3-4B	3,633,511,936	4	32	1024	Fit	103	Qwen3-4B	3,633,511,936	4	64	8	Fit
104	Qwen3-4B	3,633,511,936	4	64	32	Fit	105	Qwen3-4B	3,633,511,936	4	64	128	Fit
106	Qwen3-4B	3,633,511,936	4	64	1024	Fit	107	Qwen3-4B	3,633,511,936	4	128	8	Fit
108	Qwen3-4B	3,633,511,936	4	128	32	Fit	109	Qwen3-4B	3,633,511,936	4	128	128	Fit
110	Qwen3-4B	3,633,511,936	4	128	1024	Fit	111	Qwen3-4B	3,633,511,936	4	1024	8	Fit

Continued on next page

Table 9 – continued from previous page

No.	Model	N	B	G	C_b	Type	No.	Model	N	B	G	C_b	Type
112	Qwen3-4B	3,633,511,936	4	1024	32	Fit	113	Qwen3-4B	3,633,511,936	4	1024	128	Fit
114	Qwen3-4B	3,633,511,936	4	1024	1024	Fit	115	Qwen3-4B	3,633,511,936	3	32	8	Fit
116	Qwen3-4B	3,633,511,936	3	32	32	Fit	117	Qwen3-4B	3,633,511,936	3	32	128	Fit
118	Qwen3-4B	3,633,511,936	3	32	1024	Fit	119	Qwen3-4B	3,633,511,936	3	64	8	Fit
120	Qwen3-4B	3,633,511,936	3	64	32	Fit	121	Qwen3-4B	3,633,511,936	3	64	128	Fit
122	Qwen3-4B	3,633,511,936	3	64	1024	Fit	123	Qwen3-4B	3,633,511,936	3	128	8	Fit
124	Qwen3-4B	3,633,511,936	3	128	32	Fit	125	Qwen3-4B	3,633,511,936	3	128	128	Fit
126	Qwen3-4B	3,633,511,936	3	128	1024	Fit	127	Qwen3-4B	3,633,511,936	3	1024	8	Fit
128	Qwen3-4B	3,633,511,936	3	1024	32	Fit	129	Qwen3-4B	3,633,511,936	3	1024	128	Fit
130	Qwen3-4B	3,633,511,936	3	1024	1024	Fit	131	Qwen3-4B	3,633,511,936	2	32	8	Fit
132	Qwen3-4B	3,633,511,936	2	32	32	Fit	133	Qwen3-4B	3,633,511,936	2	32	128	Fit
134	Qwen3-4B	3,633,511,936	2	32	1024	Fit	135	Qwen3-4B	3,633,511,936	2	64	8	Fit
136	Qwen3-4B	3,633,511,936	2	64	32	Fit	137	Qwen3-4B	3,633,511,936	2	64	128	Fit
138	Qwen3-4B	3,633,511,936	2	64	1024	Fit	139	Qwen3-4B	3,633,511,936	2	128	8	Fit
140	Qwen3-4B	3,633,511,936	2	128	32	Fit	141	Qwen3-4B	3,633,511,936	2	128	128	Fit
142	Qwen3-4B	3,633,511,936	2	128	1024	Fit	143	Qwen3-4B	3,633,511,936	2	1024	8	Fit
144	Qwen3-4B	3,633,511,936	2	1024	32	Fit	145	Qwen3-4B	3,633,511,936	2	1024	128	Fit
146	Qwen3-4B	3,633,511,936	2	1024	1024	Fit	147	Qwen3-8B	6,946,075,648	8	128	128	Fit
148	Qwen3-8B	6,946,075,648	4	32	8	Fit	149	Qwen3-8B	6,946,075,648	4	32	32	Fit
150	Qwen3-8B	6,946,075,648	4	32	128	Fit	151	Qwen3-8B	6,946,075,648	4	32	1024	Fit
152	Qwen3-8B	6,946,075,648	4	64	8	Fit	153	Qwen3-8B	6,946,075,648	4	64	32	Fit
154	Qwen3-8B	6,946,075,648	4	64	128	Fit	155	Qwen3-8B	6,946,075,648	4	64	1024	Fit
156	Qwen3-8B	6,946,075,648	4	128	8	Fit	157	Qwen3-8B	6,946,075,648	4	128	32	Fit
158	Qwen3-8B	6,946,075,648	4	128	128	Fit	159	Qwen3-8B	6,946,075,648	4	128	1024	Fit
160	Qwen3-8B	6,946,075,648	4	1024	8	Fit	161	Qwen3-8B	6,946,075,648	4	1024	32	Fit
162	Qwen3-8B	6,946,075,648	4	1024	128	Fit	163	Qwen3-8B	6,946,075,648	4	1024	1024	Fit
164	Qwen3-8B	6,946,075,648	3	32	8	Fit	165	Qwen3-8B	6,946,075,648	3	32	32	Fit
166	Qwen3-8B	6,946,075,648	3	32	128	Fit	167	Qwen3-8B	6,946,075,648	3	32	1024	Fit
168	Qwen3-8B	6,946,075,648	3	64	8	Fit	169	Qwen3-8B	6,946,075,648	3	64	32	Fit
170	Qwen3-8B	6,946,075,648	3	64	128	Fit	171	Qwen3-8B	6,946,075,648	3	64	1024	Fit
172	Qwen3-8B	6,946,075,648	3	128	8	Fit	173	Qwen3-8B	6,946,075,648	3	128	32	Fit
174	Qwen3-8B	6,946,075,648	3	128	128	Fit	175	Qwen3-8B	6,946,075,648	3	128	1024	Fit
176	Qwen3-8B	6,946,075,648	3	1024	8	Fit	177	Qwen3-8B	6,946,075,648	3	1024	32	Fit
178	Qwen3-8B	6,946,075,648	3	1024	128	Fit	179	Qwen3-8B	6,946,075,648	3	1024	1024	Fit
180	Qwen3-8B	6,946,075,648	2	32	8	Fit	181	Qwen3-8B	6,946,075,648	2	32	32	Fit
182	Qwen3-8B	6,946,075,648	2	32	128	Fit	183	Qwen3-8B	6,946,075,648	2	32	1024	Fit
184	Qwen3-8B	6,946,075,648	2	64	8	Fit	185	Qwen3-8B	6,946,075,648	2	64	32	Fit
186	Qwen3-8B	6,946,075,648	2	64	128	Fit	187	Qwen3-8B	6,946,075,648	2	64	1024	Fit
188	Qwen3-8B	6,946,075,648	2	128	8	Fit	189	Qwen3-8B	6,946,075,648	2	128	32	Fit
190	Qwen3-8B	6,946,075,648	2	128	128	Fit	191	Qwen3-8B	6,946,075,648	2	128	1024	Fit
192	Qwen3-8B	6,946,075,648	2	1024	8	Fit	193	Qwen3-8B	6,946,075,648	2	1024	32	Fit
194	Qwen3-8B	6,946,075,648	2	1024	128	Fit	195	Qwen3-8B	6,946,075,648	2	1024	1024	Fit
196	Qwen3-14B	13,212,482,560	8	128	128	Fit	197	Qwen3-14B	13,212,482,560	4	32	8	Fit
198	Qwen3-14B	13,212,482,560	4	32	32	Fit	199	Qwen3-14B	13,212,482,560	4	32	128	Fit
200	Qwen3-14B	13,212,482,560	4	32	1024	Fit	201	Qwen3-14B	13,212,482,560	4	64	8	Fit
202	Qwen3-14B	13,212,482,560	4	64	32	Fit	203	Qwen3-14B	13,212,482,560	4	64	128	Fit
204	Qwen3-14B	13,212,482,560	4	64	1024	Fit	205	Qwen3-14B	13,212,482,560	4	128	8	Fit
206	Qwen3-14B	13,212,482,560	4	128	32	Fit	207	Qwen3-14B	13,212,482,560	4	128	128	Fit
208	Qwen3-14B	13,212,482,560	4	128	1024	Fit	209	Qwen3-14B	13,212,482,560	4	1024	8	Fit
210	Qwen3-14B	13,212,482,560	4	1024	32	Fit	211	Qwen3-14B	13,212,482,560	4	1024	128	Fit
212	Qwen3-14B	13,212,482,560	4	1024	1024	Fit	213	Qwen3-14B	13,212,482,560	3	32	8	Fit
214	Qwen3-14B	13,212,482,560	3	32	32	Fit	215	Qwen3-14B	13,212,482,560	3	32	128	Fit
216	Qwen3-14B	13,212,482,560	3	32	1024	Fit	217	Qwen3-14B	13,212,482,560	3	64	8	Fit
218	Qwen3-14B	13,212,482,560	3	64	32	Fit	219	Qwen3-14B	13,212,482,560	3	64	128	Fit
220	Qwen3-14B	13,212,482,560	3	64	1024	Fit	221	Qwen3-14B	13,212,482,560	3	128	8	Fit
222	Qwen3-14B	13,212,482,560	3	128	32	Fit	223	Qwen3-14B	13,212,482,560	3	128	128	Fit
224	Qwen3-14B	13,212,482,560	3	128	1024	Fit	225	Qwen3-14B	13,212,482,560	3	1024	8	Fit
226	Qwen3-14B	13,212,482,560	3	1024	32	Fit	227	Qwen3-14B	13,212,482,560	3	1024	128	Fit
228	Qwen3-14B	13,212,482,560	3	1024	1024	Fit	229	Qwen3-14B	13,212,482,560	2	32	8	Fit
230	Qwen3-14B	13,212,482,560	2	32	32	Fit	231	Qwen3-14B	13,212,482,560	2	32	128	Fit

Continued on next page

Table 9 – continued from previous page

No.	Model	N	B	G	C_b	Type	No.	Model	N	B	G	C_b	Type
232	Qwen3-14B	13,212,482,560	2	32	1024	Fit	233	Qwen3-14B	13,212,482,560	2	64	8	Fit
234	Qwen3-14B	13,212,482,560	2	64	32	Fit	235	Qwen3-14B	13,212,482,560	2	64	128	Fit
236	Qwen3-14B	13,212,482,560	2	64	1024	Fit	237	Qwen3-14B	13,212,482,560	2	128	8	Fit
238	Qwen3-14B	13,212,482,560	2	128	32	Fit	239	Qwen3-14B	13,212,482,560	2	128	128	Fit
240	Qwen3-14B	13,212,482,560	2	128	1024	Fit	241	Qwen3-14B	13,212,482,560	2	1024	8	Fit
242	Qwen3-14B	13,212,482,560	2	1024	32	Fit	243	Qwen3-14B	13,212,482,560	2	1024	128	Fit
244	Qwen3-14B	13,212,482,560	2	1024	1024	Fit	245	Qwen3-32B	31,206,298,624	8	128	128	Val
246	Qwen3-32B	31,206,298,624	4	32	128	Val	247	Qwen3-32B	31,206,298,624	4	128	8	Val
248	Qwen3-32B	31,206,298,624	4	128	128	Val	249	Qwen3-32B	31,206,298,624	4	1024	128	Val
250	Qwen3-32B	31,206,298,624	3	128	128	Val	251	Llama-3.2-1B	973,146,112	4	32	128	Gen
252	Llama-3.2-1B	973,146,112	4	64	128	Gen	253	Llama-3.2-1B	973,146,112	4	128	8	Gen
254	Llama-3.2-1B	973,146,112	4	128	32	Gen	255	Llama-3.2-1B	973,146,112	4	128	128	Gen
256	Llama-3.2-1B	973,146,112	4	128	1024	Gen	257	Llama-3.2-1B	973,146,112	4	1024	128	Gen
258	Llama-3.2-1B	973,146,112	3	32	128	Gen	259	Llama-3.2-1B	973,146,112	3	64	128	Gen
260	Llama-3.2-1B	973,146,112	3	128	8	Gen	261	Llama-3.2-1B	973,146,112	3	128	32	Gen
262	Llama-3.2-1B	973,146,112	3	128	128	Gen	263	Llama-3.2-1B	973,146,112	3	128	1024	Gen
264	Llama-3.2-1B	973,146,112	3	1024	128	Gen	265	Llama-3.2-3B	2,818,747,392	4	32	128	Gen
266	Llama-3.2-3B	2,818,747,392	4	64	128	Gen	267	Llama-3.2-3B	2,818,747,392	4	128	8	Gen
268	Llama-3.2-3B	2,818,747,392	4	128	32	Gen	269	Llama-3.2-3B	2,818,747,392	4	128	128	Gen
270	Llama-3.2-3B	2,818,747,392	4	128	1024	Gen	271	Llama-3.2-3B	2,818,747,392	4	1024	128	Gen
272	Llama-3.2-3B	2,818,747,392	3	32	128	Gen	273	Llama-3.2-3B	2,818,747,392	3	64	128	Gen
274	Llama-3.2-3B	2,818,747,392	3	128	8	Gen	275	Llama-3.2-3B	2,818,747,392	3	128	32	Gen
276	Llama-3.2-3B	2,818,747,392	3	128	128	Gen	277	Llama-3.2-3B	2,818,747,392	3	128	1024	Gen
278	Llama-3.2-3B	2,818,747,392	3	1024	128	Gen	279	Llama-3.1-8B	6,979,588,096	4	32	128	Gen
280	Llama-3.1-8B	6,979,588,096	4	64	128	Gen	281	Llama-3.1-8B	6,979,588,096	4	128	8	Gen
282	Llama-3.1-8B	6,979,588,096	4	128	32	Gen	283	Llama-3.1-8B	6,979,588,096	4	128	128	Gen
284	Llama-3.1-8B	6,979,588,096	4	128	1024	Gen	285	Llama-3.1-8B	6,979,588,096	4	1024	128	Gen
286	Llama-3.1-8B	6,979,588,096	3	32	128	Gen	287	Llama-3.1-8B	6,979,588,096	3	64	128	Gen
288	Llama-3.1-8B	6,979,588,096	3	128	8	Gen	289	Llama-3.1-8B	6,979,588,096	3	128	32	Gen
290	Llama-3.1-8B	6,979,588,096	3	128	128	Gen	291	Llama-3.1-8B	6,979,588,096	3	128	1024	Gen
292	Llama-3.1-8B	6,979,588,096	3	1024	128	Gen							