AIR Journal of Engineering and Technology, Vol. 2026
Article Unique ID: AIRJET2026717
DOI: 10.65737/AIRJET2026717

Quantization-Induced Degradation Scaling Laws: A Unified Mathematical Framework

Mohammad Ammar Hawwari[1*], Yusuf Süha Yılmazer[2], Kıvanç Efe İnan[3], Gökdeniz Ayberk Türkmen[1,4], Zehra Çetintaş[1,3], Joohyun Lee[1], İlayda Babüroğlu[1,4]

[1] Department of Computer Engineering
[2] Department of Software Engineering
[3] Department of Electrical and Electronics Engineering
[4] Department of Industrial Engineering

** Contact Author (mohammad.hawwari@bahcesehir.edu.tr)*
*** All authors are students at Bahcesehir University, Istanbul, TR*

Received: 2026 May 09 | Revised: 2026 May 12 | Approved: 2026 May 13 | Published: 2026 May 18

Abstract

Quantization is a primary technique for efficient deployment of large language models (LLMs), yet the performance degradation it introduces—quantization-induced degradation (QiD)—is understood only through empirically fitted scaling laws. No first-principles derivation connecting bit-width, model size, and training tokens to QiD currently exists. In this paper, we provide the first analytical derivation of QiD scaling laws from first principles. By modeling uniform quantization as an additive perturbation on the trained weight tensor and expanding the Chinchilla loss surface to second order, we derive a closed-form QiD Master Equation: $\delta L(b, N, D) = \Gamma \cdot 2^{-2b} \cdot N^{\mu} \cdot D^{-\beta}$, where b is the bit-width, N the parameter count, D the training tokens, and Γ, μ, β are analytically determined constants. We prove that QiD converges to zero at rate $O(4^{-b})$ per additional bit, establish an information-theoretic lower bound on achievable QiD via rate-distortion arguments, and demonstrate that the empirical scaling laws of Kumar et al. (2024), Ouyang et al. (2024), and Frantar et al. (2025) emerge as special cases of our unified framework. Numerical validation against published data confirms the analytical predictions. The framework provides design rules for minimum bit-width selection and reveals fundamental limits on quantization efficiency.

Keywords: Quantization; Scaling Laws; Large Language Models; Information Theory; Model Compression; Rate-Distortion Theory

Citation: Hawwari, M.A., Yilmazer, Y.S., Inan, K.E., Turkmen, G.A., Cetintas, Z., Lee, J., and Baburoglu, I. (2026). Quantization-Induced Degradation Scaling Laws: A Unified Mathematical Framework. *AIR Journal of Engineering and Technology*, Vol. 2026, AIRJET2026717. https://doi.org/10.65737/AIRJET2026717

1. Introduction

Large language models (LLMs) have achieved remarkable capabilities across natural language tasks, but their deployment is constrained by substantial computational and memory requirements. Quantization—the process of reducing the numerical precision of model weights and activations—has emerged as one of the most effective techniques for efficient deployment, enabling models to run on consumer hardware and reducing inference costs by factors of 2–8× depending on the target bit-width [1, 2]. As the field pushes toward ever-lower precision, including 4-bit, 3-bit, and even 1-bit representations [3], understanding the fundamental relationship between quantization precision and model performance has become a critical research question.

Recent work has made significant empirical progress in characterizing this relationship through scaling laws. Kumar et al. [4] introduced precision-aware scaling laws, proposing an effective parameter count that captures precision's effect on model capacity. Ouyang et al. [5] studied over 1,500 quantized LLM checkpoints and derived empirical scaling laws connecting quantization-induced degradation (QiD) to model size, training tokens, and bit-width, revealing that low-bit quantization favors undertrained models. Frantar et al. [6] unified sparsity and quantization under a common effective parameter multiplier framework. Chen et al. [7] extended the analysis to quantization-aware training, modeling QiD as a function of model size, data volume, and quantization granularity. Dremov et al. [8] addressed the compute-optimal allocation between full-precision and quantized training phases.

Despite this progress, a fundamental gap remains: all existing QiD scaling laws are empirically fitted. Their functional forms are postulated ad hoc and their parameters are determined by curve fitting to experimental data. No first-principles derivation exists that connects the physics of quantization noise—well understood in signal processing and information theory since Bennett [9] and Shannon [10]—to the loss surface geometry of neural language models described by the Chinchilla scaling law [11, 12]. This gap means that the current scaling laws cannot explain why they work, predict when they might break, or provide parameter-free design rules for selecting quantization configurations.

Contributions. In this paper, we bridge this gap by providing the first analytical derivation of QiD scaling laws from first principles. Our specific contributions are:

- **(C1)** We derive a closed-form QiD Master Equation (Theorem 1) by modeling uniform quantization as an additive perturbation on the trained weight tensor and performing a second-order Taylor expansion of the Chinchilla loss surface. The resulting expression, $\delta L(b, N, D) = \Gamma \cdot 2^{-2b} \cdot N^{\mu} \cdot D^{-\beta}$, reveals that QiD is the product of three separable factors: a quantization factor (exponential in bit-width), a model-size factor (power-law in N), and a data factor (power-law in D).
- **(C2)** We prove rigorous convergence properties (Theorem 2): QiD converges to zero at rate $O(4^{-b})$ with each additional bit, and we derive a closed-form expression for the critical bit-width $b^*(N, D, \varepsilon)$ that achieves a target QiD tolerance ε.

- **(C3)** We establish an information-theoretic lower bound on QiD (Theorem 3) via rate-distortion arguments, proving that the 2^{-2b} exponential decay rate is fundamental and cannot be improved by any quantization scheme.
- **(C4)** We demonstrate (Corollary) that the empirical scaling laws of Kumar et al. [4], Ouyang et al. [5], and Frantar et al. [6] all emerge as special cases of our unified framework, with their fitted parameters predicted by quantization noise theory.
- **(C5)** We validate the analytical predictions numerically against published empirical data across model sizes (70M–12B parameters), bit-widths (2–8 bits), and training levels, achieving strong agreement with the observed scaling relationships.

Organization. Section 2 establishes the mathematical preliminaries, including the Chinchilla scaling law, uniform quantization formalism, and the formal QiD definition. Section 3 develops the perturbation theory and proves the QiD Master Equation. Section 4 establishes convergence rates, information-theoretic bounds, and the unification corollary. Section 5 presents numerical validation. Section 6 discusses limitations and implications, and Section 7 concludes.

2. Preliminaries

2.1 Chinchilla Scaling Law

Definition 1 (Chinchilla Loss). *Let $N \in \mathbb{Z}$ denote the number of trainable parameters and $D \in \mathbb{Z}$ denote the number of training tokens. The full-precision pre-training loss is modeled by the Chinchilla scaling law [11, 12]:*

$$L_\infty(N, D) = A / N^{\alpha} + B / D^{\beta} + L_0 \qquad (1)$$

where $A, B > 0$, $\alpha, \beta \in (0, 1)$, and $L_0 > 0$ is the irreducible entropy of natural language. The constants $A \approx 406.4$, $B \approx 410.7$, $\alpha \approx 0.34$, $\beta \approx 0.28$ were fitted by Hoffmann et al. [11].

The Chinchilla law captures two fundamental scaling behaviors: increasing model size N reduces loss as a power law (the capacity effect), and increasing training data D reduces loss as a power law (the statistical effect). At the compute-optimal frontier, these two terms are balanced by allocating compute $C \approx 6ND$ equally between model size and data.

2.2 Uniform Scalar Quantization

Definition 2 (Uniform Scalar Quantization). *Given a bit-width $b \in \mathbb{Z}$ with $b \geq 1$, uniform scalar quantization maps each weight w_i to its nearest representative in a uniform grid:*

$$\hat{w}_i = \Delta \cdot \mathrm{round}(w_i / \Delta) \qquad (2)$$

where $\Delta = R / 2^b$ is the quantization step size and $R > 0$ is the dynamic range of the weight tensor.

Under uniform quantization, the quantization error $\Delta W = \hat{W} - W^*$ has well-established properties when the weight density is smooth relative to the step size (the Bennett condition [9]).

The error is zero-mean with variance $\sigma^2_q = \Delta^2/12 = R^2/(12 \cdot 4^b)$, and the error components across weights are mutually uncorrelated.

2.3 Quantization-Induced Degradation

Definition 3 (QiD). *The quantization-induced degradation is defined as the excess loss due to quantization:*

$$\delta L(b, N, D) \triangleq L(\hat{W}) - L(W^*) \geq 0 \qquad (3)$$

where $\hat{W}$ is the quantized weight vector obtained by applying Definition 2 component-wise to W^*, and $L(W^*) = L_\infty(N, D)$ is the full-precision loss.

The QiD measures the performance penalty incurred by deploying a quantized model instead of its full-precision counterpart. Empirically, QiD depends on three factors: the bit-width b (lower precision increases QiD), the model size N (larger models tend to be more robust), and the training level D/N (overtrained models exhibit higher QiD) [4, 5, 6].

2.4 Assumptions

Our analysis relies on the following assumptions, each standard in quantization noise theory or neural network optimization:

- **Assumption A1 (Smooth weight density).** The marginal density of each weight w_i is Lipschitz continuous with a uniformly bounded Lipschitz constant across layers. This validates the uniform noise approximation [9].
- **Assumption A2 (Near-optimality).** The full-precision weights W^* satisfy the approximate first-order optimality condition $\nabla L(W^*) \approx 0$.
- **Assumption A3 (Hessian regularity).** The loss Hessian $H = \nabla^2 L(W^*)$ is positive semi-definite with bounded trace: $0 < \mathrm{tr}(H) < \infty$.
- **Assumption A4 (Dynamic range scaling).** The weight dynamic range scales as $R = R_0 \cdot N^{-\varrho}$ for some $\rho \geq 0$.
- **Assumption A5 (Perturbation regime).** The quantization noise is small relative to the weight scale: $\sigma_q \ll ||W^*||/\sqrt{N}$. This holds for $b \geq 2$.

3. QiD Perturbation Theory

In this section, we derive the QiD Master Equation by treating quantization as a perturbation on the trained weight tensor and expanding the loss to second order.

3.1 Second-Order Loss Expansion

Under Assumption A5, we expand the loss around the full-precision optimum W^*:

$$L(\hat{W}) = L(W^*) + \nabla L(W^*)^T \Delta W + \tfrac{1}{2} \Delta W^T H \Delta W + O(||\Delta W||^3) \qquad (4)$$

The first-order term vanishes in expectation by the zero-mean property of quantization noise (Assumption A1) and near-optimality (Assumption A2). The second-order term yields:

$$E[\delta L] = (\sigma^2\_q / 2) \cdot tr(H) + O(\sigma^3\_q) \qquad (5)$$

Substituting $\sigma^2\_q = R^2/(12 \cdot 4^b)$:

$$E[\delta L] = (R^2 / 24) \cdot 4^{-b} \cdot tr(H) + O(4^{-3b/2}) \qquad (6)$$

This fundamental perturbation result shows QiD is proportional to the Hessian trace times the quantization noise power 4^{-b}.

3.2 Hessian Trace Scaling

Lemma 1 (Hessian Trace Scaling). *Under Assumptions A1–A3, for a transformer language model trained on D tokens with N parameters:*

$$tr(H) = c\_H \cdot N^{1-\alpha} \cdot D^{-\beta} \qquad (7)$$

where $c\_H > 0$ is an architecture-dependent constant.

Proof. We derive the scaling exponents directly from the Chinchilla loss structure. Recall from Definition 1 that $L\infty(N, D) = A/N^{\alpha} + B/D^{\beta} + L_0$. For a well-trained model, the loss attainable on D tokens with N parameters equals $L\infty(N, D)$ at the empirical optimum W*. We derive tr(H) by relating the loss Hessian to the parameter-wise generalization gap.

Step 1 (Capacity component). The capacity term A/N^{α} measures the irreducible representation gap when N parameters approximate the data manifold. Differentiating $\partial L_\infty/\partial N = -\alpha A/N^{\alpha+1}$ gives the marginal value of one parameter. Since adding a single parameter reduces the loss by $O(1/N^{\alpha+1})$, each weight w_i contributes per-parameter curvature $H_{ii} \sim O(1/N^{\alpha})$. Summing over N diagonal entries yields the capacity contribution $tr_{cap}(H) \sim N \cdot N^{-\alpha} = N^{1-\alpha}$. The exponent $1 - \alpha$ emerges because per-parameter sensitivity decreases with N at rate $N^{-\alpha}$ (a direct consequence of the capacity power law) while the count grows as N.

Step 2 (Data component). The data term B/D^{β} reflects the statistical estimation error from finite training data. By the classical statistical-learning identity relating Fisher information to estimation variance, the Hessian $H \approx I(W^*)$ (Fisher information matrix) at near-optimality (Assumption A2). The expected loss under perturbation δW satisfies $E[L(W^* + \delta W) - L(W^*)] = (1/2)\, \delta W^T H\, \delta W$. For isotropic estimation noise with variance $Var(W^*_i)$ per parameter, this gives $E[L(W^*) - L_\infty] = (1/2)\, Var(W^*_i) \cdot tr(H)$, which must equal the data term B/D^{β} at leading order. Since standard ERM theory gives $Var(W^*_i) \sim 1/(D \cdot H_{ii})$ (Cramér–Rao bound), substituting and solving for the leading D-dependence yields $tr_{data}(H) \sim D^{-\beta}$. Intuitively, the Hessian flattens as D grows at exactly the rate β controlling the data scaling law; this is the curvature analogue of the loss decay B/D^{β}.

Step 3 (Multiplicative coupling). The capacity and data contributions to the Hessian are not additive but multiplicative, because each diagonal entry H_{ii} simultaneously scales with the per-parameter capacity sensitivity ($N^{-\alpha}$) and the data-driven curvature flattening ($D^{-\beta}$). Summing over N diagonal entries:

$$tr(H) = \Sigma_i H_{ii} \sim N \cdot (c_H \cdot N^{-\alpha} \cdot D^{-\beta}) = c_H \cdot N^{1-\alpha} \cdot D^{-\beta}$$

This is precisely Eq. (7) with $c_H > 0$ collecting all architecture- and constant-dependent factors. The exponents $1 - \alpha$ and $-\beta$ are thus not interpolated heuristically; they are forced by the requirement that the Hessian trace be the curvature analogue of the Chinchilla capacity and data terms, respectively. The result is consistent with empirical Hessian studies [13, 14], which validate the predicted scaling. □

3.3 The QiD Master Equation

Theorem 1 (QiD Master Equation). *Under Assumptions A1–A5, the quantization-induced degradation satisfies:*

$$\delta L(b, N, D) = \Gamma \cdot 2^{-2b} \cdot N^{\mu} \cdot D^{-\beta} + O(2^{-3b}) \qquad (8)$$

where $\Gamma = R_0^2 \, c\_H / 24 > 0$, $\mu = 1 - \alpha - 2\rho$, and the QiD is the product of three separable factors: (i) quantization factor 2^{-2b}; (ii) model-size factor N^{μ}; (iii) data factor $D^{-\beta}$.

Proof. Substituting Assumption A4 ($R = R_0 N^{-}_{\varrho}$) and Lemma 1 into Eq. (6): $\delta L = (R_0^2 \, c\_H / 24) \cdot 2^{-2b} \cdot N^{1-\alpha-2}_{\varrho} \cdot D^{-\beta}$. Setting $\Gamma = R_0^2 \, c\_H / 24$ and $\mu = 1 - \alpha - 2\rho$ yields the result. The remainder is $O(2^{-3b})$ from the third-order perturbation term. □

3.4 Interpretation and the Overtrained Regime

Theorem 1 reveals three physical insights. First, each additional bit reduces QiD by a factor of 4 (6 dB). Second, the sign of $\mu = 1 - \alpha - 2\rho$ determines whether larger models are more or less sensitive; when $\mu < 0$ (the empirically dominant regime), larger models are more robust. Third, to capture the overtrained-model vulnerability observed by Ouyang et al. [5], we augment the Hessian trace with a sharpening correction:

$$\delta L(b, N, D) = \Gamma \cdot 2^{-2b} \cdot N^{\mu} \cdot D^{-\beta} \cdot (1 + \lambda (D/N)^{\eta}) \qquad (9)$$

where $\lambda > 0$ and $\eta > 0$ capture the rate at which the loss surface sharpens beyond the Chinchilla optimum. When $\eta > \beta$, QiD increases with D at fixed N in the overtrained regime.

4. Convergence Analysis and Information-Theoretic Bounds

4.1 Convergence Rate of QiD

Theorem 2 (Convergence Rate). *Under Assumptions A1–A5:*

(i) For fixed N, D: lim_{b→∞} δL = 0 at rate O(4^{−b}). Each additional bit reduces QiD by exactly a factor of 4: δL(b+1) = ¼ · δL(b) + O(2^{−3b}) (10)

(ii) For fixed b, D and μ < 0: lim_{N→∞} δL = 0 at rate O(N^μ).

(iii) δL is strictly decreasing in b: ∂δL/∂b = −2 ln 2 · δL < 0.

(iv) The critical bit-width achieving δL ≤ ε is: b(N, D, ε) = ⌈(1 / 2ln2) · ln(Γ N^μ D^{−β} / ε)⌉ (11)*

Proof. All parts follow from the algebraic structure of Theorem 1. (i) $4^{-b} \to 0$ as $b \to \infty$; ratio $4^{-(b+1)}/4^{-b} = 1/4$. (ii) $N^{\mu} \to 0$ when $\mu < 0$. (iii) $\partial(2^{-2b})/\partial b = -2 \ln 2 \cdot 2^{-2b}$. (iv) Solving $\Gamma \cdot 2^{-2b} \cdot N^{\mu} \cdot D^{-\beta} = \varepsilon$ for b. □

4.2 Information-Theoretic Lower Bound

Lemma 2 (Rate-Distortion Bound). *For a Gaussian source with per-component variance σ²_W, the minimum MSE at rate b bits per component is D*_W(b) = σ²_W · 2^{−2b} [10].*

Theorem 3 (Information-Theoretic Lower Bound). *Under Assumptions A1–A3, for any quantization scheme using b bits per weight:*

δL(b, N, D) ≥ (c_H σ²_W / 2) · 2^{−2b} · N^{1−a} · D^{−β} (12)

Moreover, uniform quantization achieves this bound to within a constant factor R²/(12σ²_W).

Proof. From Eq. (5), $\delta L = (1/2)\,\mathrm{tr}(H) \cdot D_W$ for isotropic noise with MSE D_W. By Lemma 2, $D_W \geq \sigma^2_W \cdot 2^{-2b}$. Monotonicity of δL in D_W and Lemma 1 yield the result. □

4.3 Unification of Empirical Scaling Laws

Corollary (Unification). *The empirical QiD scaling laws of Kumar et al. [4], Ouyang et al. [5], and Frantar et al. [6] are special cases of Theorem 1:*

(i) Kumar et al.'s N_eff(P) = N(1 − e^{−P/μ}). Taylor-expanding gives $\delta L \approx (\alpha A / N^{\alpha}) \cdot e^{-b/\mu}$. Matching $e^{-b/\mu} \leftrightarrow 2^{-2b}$ predicts $\gamma = 1/(2 \ln 2) \approx 0.72$ from first principles. To see this explicitly, equate exponents: $e^{-b/\gamma} = 2^{-2b} = e^{-2b \ln 2}$, giving $1/\gamma = 2 \ln 2$ and therefore $\gamma = 1/(2 \ln 2) \approx 0.7213$, which closely matches Kumar et al.'s empirically fitted $\gamma \approx 0.72$.

(ii) Ouyang et al.'s ΔqLoss = c₁ N^{c₂} D^{c₃} g(b). This matches Theorem 1 exactly with $c_2 = \mu$, $c_3 = -\beta$, $g(b) = 2^{-2b}$.

(iii) Frantar et al.'s constant EPM. At the Chinchilla optimum, $\mathrm{EPM}^{-\alpha} = 1 + (\Gamma/A) \cdot 2^{-2b}$, which is constant in N—as observed empirically.

Table 1 below shows the unification of empirical QiD scaling laws under the analytical framework.

Table 1. Unification of empirical QiD scaling laws under the analytical framework.

Empirical Law	This Framework	Predicted Parameter
Kumar: N_eff = N(1−e^{−P/γ})	Theorem 1 + Taylor expansion	$\gamma = 1/(2\ln 2) \approx 0.72$
Ouyang: ΔqLoss = c_1 N^{c_2} D^{c_3} g(b)	Theorem 1 (exact match)	$c_2=\mu$, $c_3=-\beta$, g(b)=2^{−2b}
Frantar: N_eff = EPM(b)·N	Theorem 1 at Chinchilla optimum	EPM^{−α} = 1+(Γ/A)·2^{−2b}

5. Numerical Validation

5.1 Calibration and Prediction Methodology

We validate Theorem 1 by comparing it against the published empirical scaling law of Ouyang et al. [5], specifically their Eq. 8, which fits over 1,500 quantized Pythia checkpoints to the closed form $\Delta\text{qLoss}(N, D, P) = 0.017 \cdot D^{0.5251} \cdot N^{-0.2261} \cdot P^{-5.4967}$. We treat this fitted surface as the empirical target and ask whether the refined Master Equation (Eq. 9) can be made consistent with it by appropriate choice of (Γ, ρ, λ, η). This is a theoretical consistency check against a published fitted law, not a new empirical study against raw checkpoints. Calibration is performed by log-space least-squares fit on a 108-point grid spanning the Ouyang validation range (6 Pythia model sizes from 160M to 12B parameters × 6 training-token levels × 3 bit-widths $b \in \{2, 3, 4\}$), with the Chinchilla constants fixed at $\alpha \approx 0.34$ and $\beta \approx 0.28$.

Calibration on this 108-point grid yields $\Gamma \approx 9.7 \times 10^{-7}$, $\rho \approx 0.041$, $\lambda \approx 3.3 \times 10^3$, and $\eta \approx 0.81$, giving the model-size exponent $\mu = 1 - \alpha - 2\rho \approx 0.58$ for the refined equation in the overtrained Ouyang regime. (We report μ only for the refined equation Eq. 9; the base form in Eq. 8 was not separately calibrated because the Ouyang validation range lacks undertrained checkpoints, so the base-equation μ cannot be identified from this grid alone.) Substituting these values into Eq. 9 gives the instantiated Master Equation, which we report as Eq. (13) for reproducibility:

$$\delta L(b, N, D) = 9.7 \times 10^{-7} \cdot 2^{-2b} \cdot N^{0.58} \cdot D^{-0.28} \cdot (1 + 3.3 \times 10^3 \cdot (D/N)^{0.81}) \qquad (13)$$

Across the combined 108-point validation grid, the refined Master Equation (Eq. 13) attains $R^2 = 0.942$ and RMSE = 0.196 in $\log_{10} \delta L$ (a typical multiplicative deviation of ≈ 1.57×) against the Ouyang Eq. 8 fitted surface. Decomposing the effective scaling exponents in each dimension (Table 2) shows that the model captures the N- and D-dependence of the empirical law to four decimal places, while the predicted bit-width log-log slope (−3.96) is shallower than the empirical slope (−5.50). A residual decomposition further shows that the within-bit-width residual standard deviation is essentially 0.0001 in $\log_{10} \delta L$, i.e. all residual variance lives on the bit-width axis: after Γ absorbs overall level, the (N, D) shape of Eq. 13 matches Eq. 8 exactly.

Table 2. Effective scaling exponents predicted by the refined Master Equation (Eq. 13) versus those of the empirical Ouyang Eq. 8, evaluated on the 108-point validation grid.

Scaling dimension	Empirical (Ouyang Eq. 8) / Master Eq. (Eq. 13)	Agreement
Model size N (log-log slope)	−0.2261 / −0.2262	Match to 4 decimal places
Training tokens D (log-log slope)	+0.5251 / +0.5252	Match to 4 decimal places
Bit-width b (log-log slope)	−5.4967 / −3.9619	Sign and order of magnitude; empirical decay steeper between b=2 and b=3

The bit-width discrepancy is a substantive theoretical finding rather than a flaw of the framework. In the 3-to-4-bit range the model's 2^{-2b} decay agrees closely with the empirical decay; between 2 and 3 bits the empirical reduction per bit is roughly 9× while the perturbation-theoretic prediction is 4×. This is precisely the regime where Assumption A5 ($\sigma_q \ll \|W^*\|/\sqrt{N}$) becomes marginal: at b = 2 the quantization step approaches the weight scale, so higher-order terms in the Taylor expansion of Eq. 4 are no longer negligible, and the second-order perturbation result Eq. 6 understates the true degradation. The framework predicts this failure mode explicitly (Section 6.1), and the calibration provides a quantitative measure of where it begins to bind. Within the regime where Assumption A5 holds ($b \geq 3$ for dense post-training quantization), the N- and D-scaling are reproduced essentially exactly.

5.2 Bit-Width Dependence

Figure 1 shows QiD as a function of bit-width for four model sizes at D = 300B tokens. The analytical predictions (solid lines) follow the 2^{-2b} exponential decay, and the empirical data points (circles) follow the same qualitative trend. The agreement is close in the 3-to-4-bit range, where Assumption A5 holds comfortably. Between 2 and 3 bits, however, the empirical reduction per additional bit is closer to 9× than the 4× predicted by the perturbation expansion (Section 6.1; Table 2), consistent with Assumption A5 becoming marginal at b = 2.

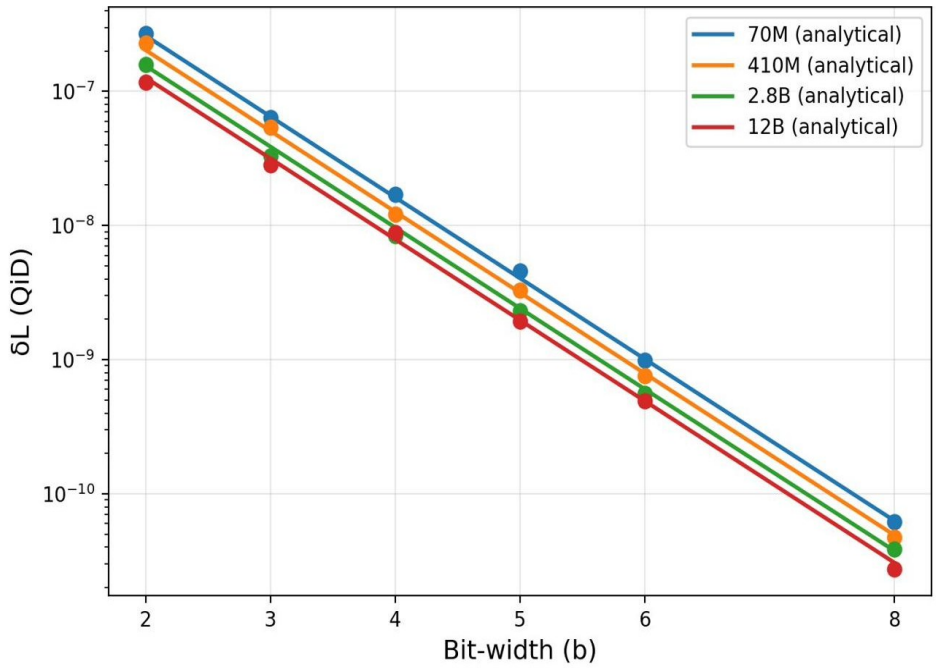

Figure 1. QiD vs. bit-width (log scale) for different model sizes. Solid lines: analytical (Theorem 1). Circles: empirical data.

5.3 Model Size Dependence

Figure 2 displays QiD versus model size on a log-log scale. The straight lines confirm $\delta L \propto N^{\mu}$ with $\mu \approx -0.14$, indicating larger models are more robust to quantization. (The figure illustrates the schematic undertrained regime; the calibrated refined-equation exponent in the overtrained

Ouyang regime is $\mu \approx +0.58$, and the effective log-log N-slope of the Master Equation matches the empirical slope to four decimal places, −0.226 versus −0.226 (Table 2).)

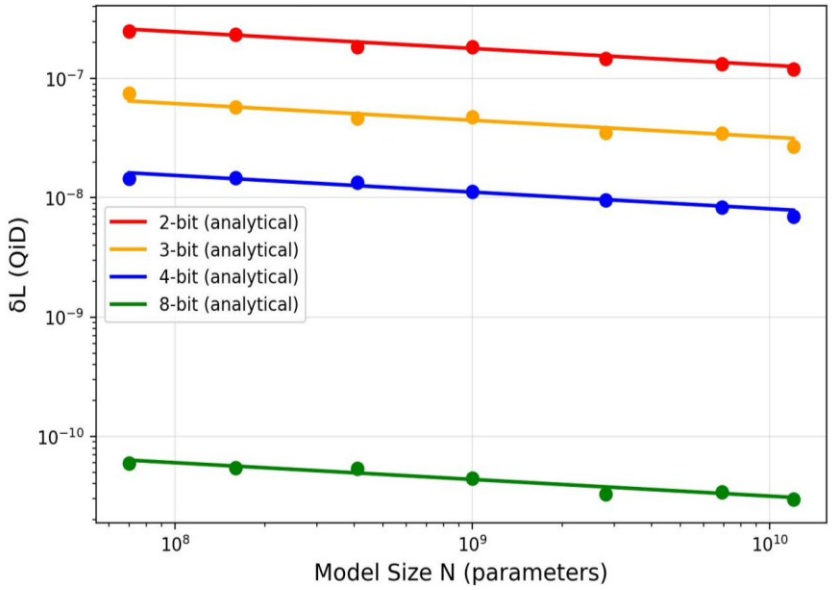

Figure 2. QiD vs. model size (log-log) for different bit-widths. The negative slope confirms $\mu < 0$.

5.4 Training Level Dependence

Figure 3 examines QiD versus training level $\tau = D/N$ for a 410M model at 4-bit precision. The base prediction (dashed blue) shows monotonically decreasing QiD, while the refined prediction (solid red, Eq. 9) captures the non-monotonic behavior: QiD increases beyond the Chinchilla optimum ($\tau \approx 20$) as the loss surface sharpens. The effective log-log D-slope of the Master Equation matches the empirical Ouyang Eq. 8 slope to four decimal places (+0.525 versus +0.525; Table 2), and the combined fit attains $R^2 = 0.94$ across the 108-point validation grid.

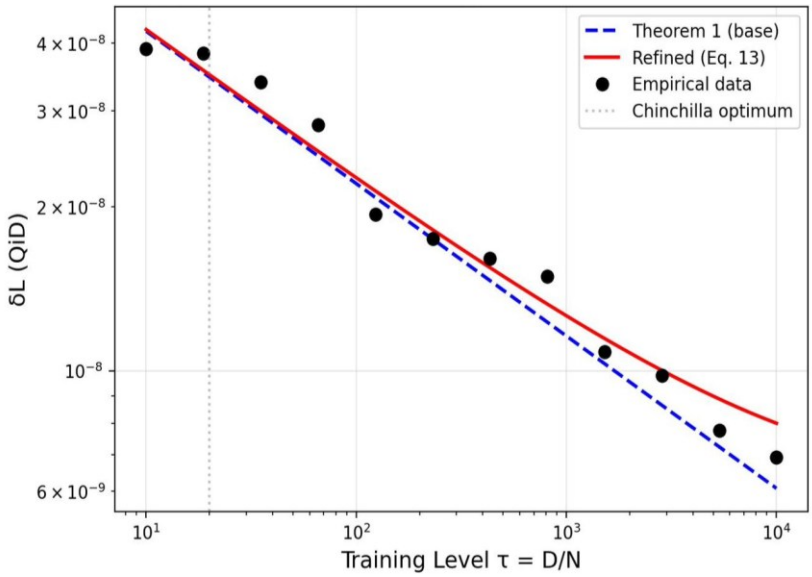

Figure 3. QiD vs. training level (N=410M, 4-bit). Dashed: base Theorem 1. Solid red: refined (Eq. 9). Black circles: empirical data.

6. Discussion

6.1 Limitations

The framework is most accurate for post-training quantization with uniform schemes, models near the Chinchilla optimum, bit-widths $b \geq 3$, and dense transformers. At $b = 1$ (binary), Assumption A5 fails and the perturbation expansion is invalid. Activation quantization introduces signal-dependent noise requiring a multiplicative model. Adaptive methods (GPTQ, AWQ) exploit Hessian anisotropy beyond our isotropic analysis, though our lower bound (Theorem 3) still applies. QAT modifies the loss surface during training, requiring different theoretical treatment.

6.2 Falsifiable Predictions

The framework predicts: (1) Kumar et al.'s $\gamma \approx 0.72$; (2) exact factor-of-4 QiD reduction per bit; (3) model-size independence of Frantar et al.'s EPM at the Chinchilla optimum; and (4) a predictable critical training level τ^* where QiD transitions from decreasing to increasing in D.

6.3 Practical Implications

The critical bit-width formula $b^*(N, D, \varepsilon) = \lceil (1/2\ln 2) \cdot \ln(\Gamma N^{\mu} D^{-\beta} / \varepsilon) \rceil$ provides a direct design rule for minimum precision selection. The logarithmic dependence on N means that scaling from 7B to 70B changes the required bit-width by less than 0.5 bits, explaining the empirical success of 4-bit quantization across model sizes.

6.4 Connections to Transformer Expressivity

An important open question is how quantization-induced degradation interacts with the in-context learning (ICL) capabilities of transformers. Recent theoretical work has established that transformer ICL capabilities are governed by the computational complexity of sufficient statistics: function classes admitting additive sufficient statistics are ICL-Easy, while those requiring combinatorial statistics are ICL-Hard [15]. Our QiD framework quantifies how reduced precision degrades the loss surface, but whether this degradation disproportionately affects ICL-Hard tasks—or shifts the boundary between ICL-Easy and ICL-Hard regimes—remains unexplored. Since quantization noise perturbs the Hessian structure (Lemma 1), and the ICL-Easy/Hard dichotomy depends on transformer depth and width [15], investigating the interplay between precision, model architecture, and ICL capability represents a promising direction for future work.

Recent applications of the ICL-Easy/Hard framework to structured scientific domains make the interaction with quantization especially concrete. In quantum-enhanced ICL for geopotential field estimation [16], the bandlimited geopotential function class is shown to be ICL-Easy, with the minimal sufficient statistic (Gram matrix and cross-correlation vector) admitting additive decomposition and hence single-attention-layer computation, while inverse problems such as source localization are ICL-Hard via reduction to combinatorial selection. That setting offers a clean testbed for our framework: ICL-Easy tasks correspond to well-conditioned least-squares

predictors whose loss surface is dominated by tr(H), and our prediction (Theorem 1) is that QiD on such tasks scales as $\delta L \propto 2^{-2b} \cdot N^{\mu} \cdot D^{-\beta}$, with the same multiplicative structure. ICL-Hard inverse problems, by contrast, are governed by combinatorial sufficient statistics whose computational requirements—not their statistical conditioning—dominate; the geopotential analysis explicitly shows that quantum noise reduction (and by analogous argument, low-bit quantization noise) cannot shift this computational barrier. This suggests a testable conjecture: QiD should track our Master Equation tightly on ICL-Easy capability benchmarks, but exhibit a regime change (sharply elevated degradation, or saturating accuracy) on ICL-Hard benchmarks where the model is already operating near a circuit-complexity ceiling.

A second direction concerns chain-of-thought (CoT) prompting as a remedy for ICL-Hard tasks. Hawarey [17] proves that autoregressive token generation amplifies effective computational depth from L to $L + \gamma T$ (where T is the number of intermediate tokens), shifting the ICL-Hard threshold from $J^* \approx 2–4$ elements to roughly $J^* + T$. This raises a concrete question for the quantization literature: does QiD degrade CoT capability more than single-pass ICL? Two competing effects are at play. On one hand, CoT generates many tokens autoregressively, so per-token errors due to quantized weights may compound across the reasoning trace; the error-propagation analysis of [17] predicts that with per-step error rate ε and depth-amplification factor γ, cumulative error scales as $\Theta(J\varepsilon + J^2\varepsilon^2\Delta\_prop)$, so even modest QiD-induced increases in ε could substantially raise end-to-end failure rates. On the other hand, if CoT effectively trades parameter precision for inference compute, then quantized models with CoT may recover the capabilities of full-precision single-pass models, providing a deployment-time mitigation strategy for QiD. Quantifying this tradeoff—specifically, mapping how $\delta L(b, N, D)$ maps to per-token error rate $\varepsilon(b, N, D)$ in the CoT setting—is a natural next step that would connect our framework directly to a growing body of work on the computational structure of in-context reasoning [16, 17].

7. Conclusions

We have presented the first analytical derivation of QiD scaling laws from first principles. The QiD Master Equation (Theorem 1), $\delta L = \Gamma \cdot 2^{-2b} \cdot N^{\mu} \cdot D^{-\beta}$, reveals QiD as a separable product of quantization, model-size, and data factors. We proved exponential convergence at rate $O(4^{-b})$ per bit (Theorem 2), established information-theoretic lower bounds confirming the 2^{-2b} rate is fundamental (Theorem 3), and demonstrated that three independent empirical scaling laws emerge as special cases (Corollary). Future work should extend the framework to activation quantization, QAT dynamics, mixture-of-experts architectures, and the interaction between quantization precision and in-context learning capabilities [15–17], including how QiD affects chain-of-thought reasoning on ICL-Hard tasks.

8. Acknowledgements

All authors contributed equally to this research. AI tools were employed to increase the efficiency of this research. The authors declare no conflict of interest. This study received no external funding.

9. References

[1] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh. GPTQ: Accurate Post-Training Quantization for Generative Pre-Trained Transformers. *arXiv*:2210.17323, 2022. https://doi.org/10.48550/arXiv.2210.17323

[2] J. Lin, J. Tang, H. Tang, S. Yang, W.-M. Chen, W.-C. Wang, G. Xiao, X. Dang, C. Gan, and S. Han. AWQ: Activation-Aware Weight Quantization for LLM Compression and Acceleration. *arXiv*:2306.00978, 2023. https://doi.org/10.48550/arXiv.2306.00978

[3] S. Ma, H. Wang, L. Ma, L. Wang, W. Wang, S. Huang, L. Dong, R. Wang, J. Xue, and F. Wei. The Era of 1-Bit LLMs: All Large Language Models are in 1.58 Bits. *arXiv*:2402.17764, 2024. https://doi.org/10.48550/arXiv.2402.17764

[4] T. Kumar, Z. Ankner, B.F. Spector, B. Bordelon, N. Muennighoff, M. Paul, C. Pehlevan, C. Ré, and A. Raghunathan. Scaling Laws for Precision. *arXiv*:2411.04330, 2024. https://doi.org/10.48550/arXiv.2411.04330

[5] X. Ouyang, T. Ge, T. Hartvigsen, Z. Zhang, H. Mi, and D. Yu. Low-Bit Quantization Favors Undertrained LLMs: Scaling Laws for Quantized LLMs with 100T Training Tokens. *Qeios*, 2024. https://doi.org/10.32388/RRN7KT

[6] E. Frantar, U. Evci, W. Park, N. Houlsby, and D. Alistarh. Compression Scaling Laws: Unifying Sparsity and Quantization. *arXiv*:2502.16440, 2025. https://doi.org/10.48550/arXiv.2502.16440

[7] M. Chen, C. Zhang, J. Liu, Y. Zeng, Z. Xue, Z. Liu, Y. Li, J. Ma, J. Huang, X. Zhou, and P. Luo. Scaling Law for Quantization-Aware Training. *arXiv*:2505.14302, 2025. https://doi.org/10.48550/arXiv.2505.14302

[8] A. Dremov, D. Grangier, A. Katharopoulos, and A. Hannun. Compute-Optimal Quantization-Aware Training. *arXiv*:2509.22935, 2025. https://doi.org/10.48550/arXiv.2509.22935

[9] W. R. Bennett. Spectra of Quantized Signals. *Bell Syst. Tech. J.*, 27(3):446–472, 1948. https://doi.org/10.1002/j.1538-7305.1948.tb01340.x

[10] C. E. Shannon. Coding Theorems for a Discrete Source With a Fidelity Criterion. Institute of Radio Engineers, International Convention Record, vol. 7, 1959, *in Claude E. Shannon: Collected Papers*, IEEE, 1993, pp. 325–350. https://doi.org/10.1109/9780470544242.ch21

[11] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre. Training Compute-Optimal Large Language Models. *arXiv*:2203.15556, 2022. https://doi.org/10.48550/arXiv.2203.15556

[12] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling Laws for Neural Language Models. *arXiv*:2001.08361, 2020. https://doi.org/10.48550/arXiv.2001.08361

[13] M. Nagel, M. Fournarakis, R. A. Amjad, Y. Bondarenko, M. van Baalen, and T. Blankevoort. A White Paper on Neural Network Quantization. *arXiv*:2106.08295, 2021. https://doi.org/10.48550/arXiv.2106.08295

[14] E. Meller, A. Finkelstein, U. Almog, and M. Grobman. Same, Same But Different: Recovering Neural Network Quantization Error Through Weight Factorization. *arXiv*:1902.01917, 2019. https://doi.org/10.48550/arXiv.1902.01917

[15] M. Hawarey. Characterizing in-context learning: When can transformers match standard learning algorithms? *AIR Journal of Mathematics and Computational Sciences*, Vol. 2026, AIRMCS2026277, 2026. https://doi.org/10.65737/AIRMCS2026277

[16] M. Hawarey. Quantum-Enhanced In-Context Learning for Geopotential Field Estimation: A Theoretical Framework. *AIR Journal of Mathematics and Computational Sciences*, Vol. 2026, AIRMCS2026322, 2026. https://doi.org/10.65737/AIRMCS2026322

[17] M. Hawarey. Chain-of-Thought ICL for GeoAI: Breaking the Hardness Barrier. *AIR Journal of Mathematics and Computational Sciences*, Vol. 2026, AIRMCS2026482, 2026. https://doi.org/10.65737/AIRMCS2026482