Compression Does Not Equally Hurt: Characterizing Calibration Drift in Quantized and Distilled LLMs

Anonymous ACL submission

Abstract

Quantization and knowledge distillation are the dominant methods for compressing large language models before deployment. Their effects on accuracy are well documented; their effects on *calibration* — whether model confidence tracks empirical accuracy — are not. We evaluate calibration drift across a grid of 77 cells measuring Expected Calibration Error before and after compression.[1] Of these, 53 are clean non-baseline cells — compressed cells passing our anomaly criteria — on which all CCD-based claims rest. The headline finding is that compression's calibration effect is sign-mixed rather than uniformly harmful: 32 of 53 cells show *improved* calibration, with the negative-CCD fraction varying by task (50% on NLI/CSQA, 64% on QA, 82% on XSum). We introduce the Compression Calibration Degradation (CCD) metric and propose an interpretive framework: overconfident baselines soften under logit perturbation; well-calibrated ones accumulate noise. The framework predicts a sign-flip in CCD as a function of baseline overconfidence and that the signed effect is detectable only on tasks with sufficient baseline-ECE variance. Both predictions are consistent with the data (pooled $\beta_1 = -2.005$, $p < 10^{-8}$; per-task slopes negative on 3 of 4 tasks). At fixed deployment size, sequence-level distillation matches or improves same-size quantization on the joint accuracy-calibration frontier in 3 of 4 tasks after temperature scaling.

1 Introduction

Large language model deployment increasingly relies on *quantization* to reduce VRAM footprint and *knowledge distillation* to reduce inference cost. The accuracy effects of these methods are well documented, but a less well-characterized question is how compression affects *calibration*: whether the model's stated confidence reliably tracks its empirical accuracy. Calibration matters whenever downstream systems consume model probabilities for routing, abstention, retrieval re-ranking, or human-in-the-loop review; a model whose calibration silently degrades under compression poses a quiet failure mode that accuracy-only benchmarks miss.

This paper presents a systematic empirical characterization of compression's effect on calibration. We evaluate 6 instruction-tuned LLMs across 4 NLP tasks spanning distinct output structures, with up to 5 compression settings per (model, task) pair. The nominal $6 \times 4 \times 5 = 120$ design is not fully exercised ($3{\times}20 + 1{\times}17$: three tasks cover all 6 models, while XSum has 17 cells because QwQ-32B is excluded as a reasoning-oriented model unsuitable for abstractive summarization): several method$\times$model combinations are infeasible on our hardware (notably INT8 for Llama-3.3-70B), and Gemma-4-31B serves only as a multimodal sanity baseline at NF4. After excluding three measurements failing the anomaly criteria of §5, our analysis rests on 53 *clean non-baseline cells*—i.e., valid experimental settings where a compressed model is evaluated against its NF4 baseline counterpart. The central empirical observation is that compression's effect on Expected Calibration Error is *non-monotonic* and *sign-mixed*: 32 cells improve relative to baseline, while 21 degrade. The sign distribution also varies systematically by task: XSum improves in 82

To organize this heterogeneity, we propose an interpretive framework grounded in softmax perturbation geometry (§4). The framework makes two qualitative predictions: (P1) compression should improve calibration when the baseline overconfidence gap is large; (P2) the signed effect should be empirically detectable only on tasks where baseline ECE varies enough for the directional term to

[1] Anonymous code and result JSONs: https://anonymous.4open.science/r/Quantization-vs-distillation-6FFB/readme.md.

dominate task-dependent noise.

Both predictions are consistent with the data. Per-task slopes of CCD on baseline overconfidence are negative on three of four tasks (Table 1), with detectability ordered by within-task baseline-ECE range: significant on QA ($r = -0.82$, range 0.50) and CSQA ($r = -0.68$), directionally consistent on XSum ($r = -0.42$), undetectable on NLI ($r = +0.09$, baselines clustered narrowly). The pooled regression aggregates to $\beta_1 = -2.005$ ($p < 10^{-8}$); LOTO analysis shows QA carries most of the pooled statistical power, which is itself predicted by P2. We treat the framework as an organizing principle for the observed patterns rather than as a predictive law; measurement limitations, particularly on the sequence-log-prob confidence proxy, are discussed in §9 and §9.

On the practical side, we compare same-size deployment options: quantized Llama-3.1-8B (NF4, INT8, GGUF Q3–Q6) versus a Llama-3.1-8B student distilled from a Llama-3.3-70B teacher via sequence-level QLoRA. After temperature scaling, the distilled student matches or improves on same-size quantized baselines on 3 of 4 tasks; on XSum, the calibration improvement is well outside the n=500 binomial noise floor (disjoint 95% CIs).

Contributions. (1) A large empirical characterization of LLM calibration drift under quantization and distillation across multiple compression families and structurally different tasks (§6). (2) The CCD metric, a normalized cross-task measure of compression-induced calibration drift (§3). (3) An interpretive framework with two qualitative predictions — a sign-flip in CCD as a function of baseline overconfidence, and heterogeneity in detectability tied to baseline-ECE variance — both consistent with the data (§4). (4) A practitioner-relevant comparison showing sequence-level distillation matches or improves same-size quantization in 3 of 4 tasks at fixed deployment size (§6.3). All result JSONs, regression scripts, and anomaly-cell raw outputs are released with the codebase.

2 Related Work

LLM compression. Post-training quantization for LLMs is dominated by `bitsandbytes`-style methods (Dettmers et al., 2023, 2022) and structured calibration-aware approaches (Frantar et al., 2023; Lin et al., 2024). The GGUF k-quant family has become the de facto deployment format for consumer hardware. Sequence-level distillation for LLMs was popularized by Kim and Rush (2016) and adopted broadly for instruction-tuned models (Gu et al., 2024). Most existing work evaluates these methods on task accuracy; calibration is rarely a primary metric.

Calibration in NLP and in LLMs. The modern framing of neural-network miscalibration is Guo et al. (2017), which introduced temperature scaling. Desai and Durrett (2020) extended the analysis to pretrained transformers; Jiang et al. (2021) characterized LLM-specific patterns. Recent work has examined whether LLMs can self-report calibration (Kadavath et al., 2022; Tian et al., 2023) or whether semantic equivalence over multiple generations provides better uncertainty estimates (Kuhn et al., 2023). The most closely related prior work is Liu et al. (2023), which studies quantization-on-calibration for smaller models. No prior work, to our knowledge, has systematically characterized calibration drift across LLM-scale models, multiple compression methods, and tasks with structurally different output spaces.

Theoretical analysis. The softmax Jacobian identity used in Theorem 1 is classical (Bishop, 2006). The observation that the $p(1-p)$ factor governs sensitivity appears in the context of vanishing gradients and label smoothing (Müller et al., 2019). We are not aware of prior work connecting this asymmetry to compression-induced calibration drift specifically. The closest theoretical antecedent is work on quantization noise as implicit regularization (Hubara et al., 2017), focused on generalization rather than calibration.

3 Preliminaries and Compression Calibration Degradation

3.1 Calibration of Probabilistic Classifiers

We consider a model $f : \mathcal{X} \to \Delta^{K-1}$ that maps inputs to probability distributions over K classes (for classification tasks) or, in the case of generative tasks, induces a confidence estimate over its output sequence. Following Guo et al. (2017), the model is *perfectly calibrated* if $\Pr(\hat{y} = y \mid \hat{p}(\hat{y} \mid x) = c) = c$ for all $c \in [0, 1]$, where $\hat{y} = \arg\max_k f(x)_k$ and $\hat{p}(\hat{y} \mid x) = \max_k f(x)_k$. We estimate calibration via the Expected Calibration Error with M equal-width bins:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \big|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\big|,$$

with $\mathrm{acc}(B_m)$ and $\mathrm{conf}(B_m)$ the empirical accuracy and mean confidence in bin B_m. We use $M = 15$ and additionally report MCE and Brier score (for classification).

For generative tasks (extractive QA and abstractive summarization) we use the length-normalized sequence log-probability, mapped to $[0, 1]$ via a sigmoid, as the model's confidence. Correctness is determined by exact-match for QA and by a ROUGE-1 threshold ($\text{ROUGE-1} > 0.1$) for summarization. This is a known-imperfect proxy; we discuss its limitations explicitly in §9. We checked sign preservation of CCD on the summarization cells across $\theta \in \{0.05, 0.10, 0.15, 0.20\}$ and found that CCD signs are preserved in every tested cell, although absolute magnitudes can vary by up to $\sim 30\%$.

3.2 The CCD Metric

Comparing raw ECE values across heterogeneous tasks is misleading: an ECE of 0.10 on a 3-class NLI problem indicates substantially worse calibration than the same value on a free-form generation task with a noisy confidence proxy. To enable cross-task comparison of compression effects, we introduce:

$$\mathrm{CCD}(M_c, M_b) = \frac{\mathrm{ECE}(M_c) - \mathrm{ECE}(M_b)}{\max(\mathrm{ECE}(M_b), \epsilon)}, \quad (1)$$

with $\epsilon = 10^{-8}$, where M_b is an uncompressed baseline and M_c its compressed counterpart. CCD has three properties relevant to our analysis: sign encodes direction ($\mathrm{CCD} > 0$ indicates calibration worsened); magnitude is relative to baseline scale; baselines are self-referential ($\mathrm{CCD}(M_b, M_b) = 0$).

Choice of denominator. We normalize by baseline ECE rather than by a theoretical ideal because practitioners deploy *from* the baseline; the practitioner question is "how much worse relative to what I would have shipped?" CCD inherits the binning artifacts of ECE and is undefined when baseline ECE is exactly zero (we use ϵ for stability). When baseline ECE is very small (< 0.02), small absolute changes produce large CCD magnitudes; we flag affected cells.

Operational baseline. We use the NF4 4-bit quantization as the operational baseline rather than full BF16, reflecting deployment reality: serving a 70B model in BF16 requires $\sim$140 GB of GPU memory. NF4 introduces less than 0.1% perplexity degradation relative to BF16 (Dettmers et al., 2023), and ECE differences between the two are within our measurement noise floor (< 0.02).

4 An Interpretive Framework

This section develops a framework for interpreting the empirical patterns we observe in §6. The framework rests on a classical sensitivity bound for the softmax top-1 probability, which we restate as Theorem 1, and a heuristic decomposition (Heuristic Decomposition 1) whose qualitative implications we test against the data.

4.1 Softmax Perturbation Geometry

The following sensitivity bound for the softmax top-1 probability is standard — it is a direct consequence of the softmax Jacobian identity (Bishop, 2006) and the same $p(1-p)$ scaling that appears in the gradient analysis underlying confidence regularization and label smoothing (Müller et al., 2019; Pereyra et al., 2017). We restate it here in the form we use, with the elementary derivation given in Appendix A for completeness; we make no claim of novelty for this result.

Theorem 1 (Softmax Sensitivity; restatement)**.** *Let $z \in \mathbb{R}^V$ be a logit vector and $p = \mathrm{softmax}(z)$, with top-1 index $i^\star = \arg\max_i p_i$. For any perturbation $\Delta z \in \mathbb{R}^V$ with $\|\Delta z\|_\infty \le \delta$:*

$$|\Delta p_{i^\star}| \le 2\, p_{i^\star}(1 - p_{i^\star})\, \delta + O(\delta^2). \quad (2)$$

The novelty of §4 is not Theorem 1 itself, but the combination of (i) instantiating this classical bound in the LLM-compression setting and (ii) using it to motivate the calibration-budget decomposition in §4.2, whose qualitative predictions we then test empirically.

Corollary 1 (Sensitivity Asymmetry)**.** *The softmax perturbation magnitude is maximized at $p_{i^\star} = 1/2$ and tends to zero as $p_{i^\star} \to 1$ or $p_{i^\star} \to 0$.*

Implication for calibration. A uniformly overconfident model sits in the low-sensitivity region: the same logit perturbation that would shift an uncertain prediction barely moves its softmax output. The perturbation that does occur is biased toward softening, paralleling observations from label smoothing and confidence regularization (Müller et al., 2019; Pereyra et al., 2017). This is the geometric origin of the sign-flip phenomenon: compression of an overconfident baseline produces small, softening perturbations that move predictions toward the diagonal of the reliability diagram,

improving calibration. A well-calibrated baseline sits in the high-sensitivity region; perturbation is larger and approximately mean-zero, which inflates ECE on net.

4.2 The Calibration Budget

We present the consequence for ECE as a candidate decomposition rather than a fully rigorous theorem. The proof mainly relies on three idealizations (bin-membership stability; symmetry of quantization perturbation; pointwise-to-population concentration) that preclude a tight statement. We use it heuristically, to generate qualitative predictions we then test against the data.

Heuristic Decomposition 1 (Calibration Budget)**.** *Under the perturbation model above,*

$$\mathrm{ECE}(M_c) \lesssim \mathrm{ECE}(M_b) + \Delta_{\mathrm{comp}}(\delta, \mathcal{T}, M_b), \quad (3)$$

where the drift term decomposes as

$$\Delta_{\mathrm{comp}} \approx -c_1(\mathcal{T}) \cdot \left[\mathrm{ECE}(M_b) - \mathrm{ECE}^\star_{\mathcal{T}}\right] + c_2(\mathcal{T}, \delta), \quad (4)$$

with $c_1(\mathcal{T}) > 0$ *and* $\mathrm{ECE}^\star_{\mathcal{T}} \geq 0$ *a task-specific reference floor.*

The intuition: Theorem 1 implies that under a quantization perturbation, the expected probability shift is biased toward softening when baseline confidences are systematically above the accuracy line. The magnitude scales with the overconfidence gap $\mathrm{ECE}(M_b) - \mathrm{ECE}^\star_{\mathcal{T}}$. The second term captures the directionless component, analogous to noise-induced calibration degradation (Nixon et al., 2019; Minderer et al., 2021): even on a perfectly-calibrated baseline, quantization adds noise and inflates ECE through its effect on bin boundaries. Setting $\Delta_{\mathrm{comp}} = 0$ in (4) yields the sign-flip threshold

$$\mathrm{ECE}(M_b) > \mathrm{ECE}^\star_{\mathcal{T}} + \frac{c_2(\mathcal{T}, \delta)}{c_1(\mathcal{T})} \Rightarrow \mathrm{CCD} < 0. \quad (5)$$

Two qualitative predictions. The framework makes two predictions we can test against the data: **(P1)** *Cells with larger baseline overconfidence gaps should exhibit more-negative CCD.* This is the direct sign-flip prediction: at fixed task structure, the directional term in (4) scales linearly with the gap. **(P2)** *The signed effect should be empirically detectable only on tasks where baseline ECE varies sufficiently.* When baseline ECE varies little within a task, the directional term in (4) contributes little cross-cell variance, and the noise term c_2 dominates. Per-task slopes should therefore be steepest (and statistical power highest) on tasks with the widest baseline-ECE range.

These predictions are qualitative, not quantitative. We do not claim they hold pointwise or that they yield a tight predictive model of CCD. We test their qualitative content in §4.3.

Table 1: Per-task univariate slope of CCD on overconfidence gap, with the within-task range of baseline ECE. The framework predicts steeper (more-negative) slopes on tasks with wider baseline-ECE ranges; the per-task ordering is consistent with this prediction.

Task	n	ECE_b range	slope	r	p
MultiNLI	14	0.16	+0.58	+0.09	0.77
CommonsenseQA	14	0.09	−3.62	−0.68	0.008
SQuAD 2.0	14	0.50	−1.86	−0.82	< 0.001
XSum	11	0.18	−2.04	−0.42	0.20

4.3 Testing the Framework's Predictions

Heuristic Decomposition 1 motivates the linear form

$$\mathrm{CCD} \approx \beta_0 + \beta_1[\mathrm{ECE}(M_b) - \mathrm{ECE}^\star_{\mathcal{T}}] + \beta_2\tau + \beta_3 \log(\mathrm{bpw}) + \varepsilon, \quad (6)$$

where $\tau = \log K$ is output-cardinality and bpw is bits-per-weight. The framework predicts $\beta_1 < 0$ (the only signed prediction).

Per-task slopes (Prediction P2). We compute univariate slopes of CCD on baseline overconfidence within each task, alongside the within-task range of baseline ECE (Table 1). Three of four tasks show the predicted negative slope. The ordering matches the framework's prediction: tasks with wider baseline-ECE ranges (QA: span 0.50; XSum: span 0.18) show stronger negative slopes than tasks with narrower ranges (CSQA: span 0.09; NLI: span 0.16 but with most baselines clustered between 0.078 and 0.144). NLI yields a null result — consistent with Prediction P2 since the directional term in (4) cannot dominate the noise term when baselines cluster near the task floor.

Pooled regression (Prediction P1). Aggregating across all 53 clean non-baseline cells, we fit (6) with $\mathrm{ECE}^\star_{\mathcal{T}}$ set to the per-task minimum baseline ECE. Bootstrap 95% CIs (B=2000):

- $\beta_1 = -2.005 \in [-2.82, -1.46]$, $p < 10^{-8}$ (overconfidence gap, sign as predicted by P1).

- $\beta_2 = -0.004 \in [-0.019, +0.011]$, $p = 0.20$ (task cardinality, n.s.).
- $\beta_3 = +0.024 \in [-0.21, +0.25]$, $p = 0.53$ (bits-per-weight, n.s.).

In-sample $R^2 = 0.487$; 5-fold CV $R^2 = 0.331 \pm 0.382$; sign-prediction accuracy 75.5%. The signed coefficient on the overconfidence gap is consistent with Prediction P1.

Leave-one-task-out sensitivity. Because the per-task slopes show that QA carries the widest baseline-ECE range, we expect QA to provide most of the pooled regression's statistical power. We confirm this directly. Holding out NLI, CSQA, or XSum from the pooled regression leaves β_1 negative and significant (range -1.46 to -1.98, $p < 0.01$); holding out SQuAD 2.0 reduces β_1 to -0.82 ($p = 0.37$, $R^2 = 0.07$). This concentration of statistical power in a single task is a *prediction* of the framework (P2), not a contradiction of it: the task with the widest baseline-ECE range is exactly where the signed effect should be most resolvable. We discuss the implications for the QA confidence proxy in §9.

Regression diagnostics and robustness. We fit (6) on $n = 53$ clean compressed cells, with $\mathrm{ECE}^{\star}_{\mathcal{T}}$ defined as the per-task minimum baseline ECE. Robustness checks using alternative τ encodings ($K \in \{50000, 1000, 100, 50\}$), binary classification-vs-generative indicators, and interaction terms yield the same qualitative result: β_1 remains stable and negative, while β_2 is consistently non-significant.

Table 2: Pooled regression results. Bootstrap confidence intervals use $B = 2000$ resamples. The pooled coefficient aggregates per-task slopes.

Predictor	Est.	95% CI	r
overconf. gap	-2.01	$[-2.82, -1.46]$	$-.695$
$\tau = \log K$	-0.004	$[-0.019, +0.011]$	$-.180$
$\log(\mathrm{bpw})$	$+0.024$	$[-0.21, +0.25]$	$+.089$
intercept	$+0.105$	$[-0.18, +0.41]$	

Summary. Both qualitative predictions are consistent with the data across multiple independent observations — per-task slopes (P2), the pooled regression coefficient sign (P1), and LOTO sensitivity (P2). We do not claim the framework yields a tight predictive law; the moderate cross-validation R^2 is consistent with its qualitative status.

5 Experimental Setup

Models. Six instruction-tuned LLMs: Llama-3.1-8B-Instruct, Llama-3.3-70B-Instruct, Qwen-2.5-14B-Instruct, QwQ-32B, Gemma-2-27B-it, and Gemma-4-31B-it. Llama-3.3-70B serves as teacher in within-family distillation; Qwen-2.5-14B as student in MoE→dense and cross-family distillation. Gemma-4-31B is evaluated under its `AutoProcessor` pipeline.

Compression methods. NF4 (operational baseline), INT8, GGUF Q3_K_M / Q4_K_M / Q6_K, and sequence-level distillation via QLoRA. Teachers generate greedy responses on a training subset; students are fine-tuned on (x, y_t) pairs with LoRA rank 16, alpha 32, and early stopping on a 10% validation split. Distillation settings: Llama-3.3-70B → Llama-3.1-8B, Gemma-4-31B → Gemma-E4B, Qwen3-30B-A3B → Qwen-2.5-14B (MoE→dense), and Gemma-2-27B → Qwen3-2B.

Tasks and evaluation. We evaluate on MultiNLI, CommonsenseQA, SQuAD 2.0, and XSum. Quantized evaluations use n=500 examples per cell; distillation evaluations use n=1000–2000. Classification uses first-token scoring. QA and summarization use greedy decoding with confidence from length-normalized mean log-probability; summarization ECE uses a binary ROUGE-1>0.1 correctness proxy. Calibration uses 15 equal-width bins with temperature scaling fit by L-BFGS-B on NLL ($T \in [10^{-4}, 100]$). All experiments run on RTX A6000 GPUs.

Statistical precision of ECE. For n=500 and M=15 bins, analytical propagation of per-bin binomial variance gives $\sigma_{\mathrm{ECE}} \approx 0.015$–$0.020$. We therefore treat cell-level differences below ≈ 0.04 (roughly $|\mathrm{CCD}| \approx 0.2$ at baseline ECE 0.20) as noise and focus on aggregated trends (§4.3). For distillation evaluations (n=1000–2000), σ_{ECE} drops to ≈ 0.010–0.015.

Anomaly criteria. Three cells are excluded (Appendix C): Gemma-4-31B SQuAD 2.0 (ECE $= 0.752$), and Llama-3.1-8B XSum NF4/INT8 (ROUGE-1 $= 0.160/0.302$), all proxy failures. Remaining Llama-3.1-8B XSum cells are retained, but their CCD values are omitted from headline analysis because the baseline is anomalous. The Qwen3-30B-A3B→Qwen-2.5-14B distilled QA cell (EM $= 0.243$) is flagged but retained.

6 Results

Figure 1 shows reliability diagrams for six representative classification cells. All cells show the characteristic LLM overconfidence pattern (curves below the diagonal) before temperature scaling; after scaling, most cells approach the diagonal closely.

6.1 The Sign-Mixed Pattern

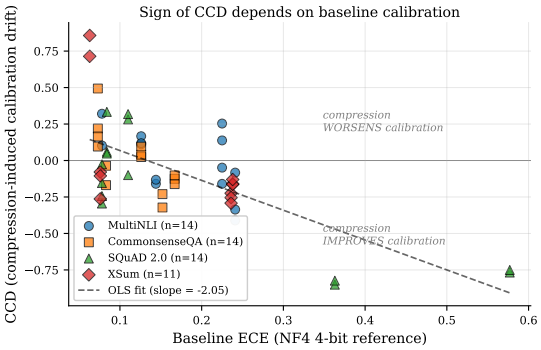

Figure 2: CCD vs. baseline ECE for all 53 clean non-baseline cells, colored by task. Cells with the highest baseline ECEs (QwQ-32B and Llama-3.3-70B on SQuAD 2.0) sit at the bottom-right, showing the most-improved calibration after compression — consistent with the framework's Prediction P1.

Figure 2 plots CCD against baseline ECE for all 53 clean cells. Cells with the highest baseline ECEs (QwQ-32B and Llama-3.3-70B on SQuAD 2.0) sit at the bottom-right, showing the most-improved calibration after compression.

Table 3: Distribution of CCD signs by task. The negative-CCD fraction varies systematically across tasks; per-task slopes (Table 1) and within-task baseline-ECE variation provide additional structure.

Task	n	CCD > 0	CCD < 0	Mean ECE_b
MultiNLI	14	7 (50%)	7 (50%)	0.175
CSQA	14	7 (50%)	7 (50%)	0.117
SQuAD 2.0	14	5 (36%)	9 (64%)	0.198
XSum	11	2 (18%)	9 (82%)	0.161

Table 3 shows the negative-CCD fraction is highest for XSum (82%) and lowest for NLI and CommonsenseQA (50% each); SQuAD 2.0 lies between at 64%.

MultiNLI. The best-calibrated baseline on this task is QwQ-32B NF4 (ECE 0.078), and the highest accuracy is achieved by Gemma-4-31B NF4 (acc 0.836). Compression effects are mixed: Llama-3.1-8B suffers the worst degradation (Q4_K_M CCD $= +1.01$, ECE $0.225 \rightarrow 0.282$, $\Delta\text{ECE} = 0.057$, exceeds noise floor), while Qwen-2.5-14B Q4_K_M improves calibration (CCD $= -0.41$, $\Delta\text{ECE} = -0.099$) alongside an 11-point accuracy gain. This pair — one model hurt, one helped, on the same task — illustrates that the overconfidence-gap mechanism, where it operates, operates at the model level, not just the task level: Qwen-2.5-14B NF4 had the highest baseline ECE on NLI (0.241), so its overconfidence gap was large enough to cross the sign-flip threshold even on a task where the aggregate per-task slope is null.

CommonsenseQA. QwQ-32B is uniformly the strongest. Its compressed cells (Q4–Q6) report ECE within the noise floor of its NF4 baseline (0.069–0.080 vs. baseline 0.083), so we cannot statistically distinguish a calibration improvement from no change; the qualitative observation is that QwQ-32B tolerates mild compression without measurable calibration loss. Llama-3.1-8B is the weakest model on this task: Q6_K yields CCD $= +0.49$ ($\Delta\text{ECE} = 0.036$, just above noise floor), the worst degradation on CSQA.

SQuAD 2.0. The defining pattern is dramatic baseline ECE for high-accuracy models: Llama-3.3-70B NF4 (ECE 0.577, EM 0.610) and QwQ-32B NF4 (ECE 0.363, EM 0.626) are severely overconfident at the sequence-log-prob level. Quantization collapses this: QwQ-32B Q6_K achieves CCD $= -0.85$ and Llama-3.3-70B Q4_K_M achieves CCD $= -0.77$. However, both improvements coincide with EM drops of 24 and 15 points; the proxy cannot disentangle logit softening from accuracy collapse. Temperature scaling is a strictly better recalibration strategy, recovering calibration without the accuracy cost.

XSum. 10 of 12 cells (excluding the two anomalous Llama-3.1-8B cells) show CCD < 0, the strongest task-level signal in our grid. The exception is Llama-3.3-70B (CCD $= +0.85$ and $+0.70$ for Q4/Q3), whose XSum NF4 baseline (ECE $= 0.063$) is the best-calibrated on the task — near-zero overconfidence gap means the noise term c_2 dominates, consistent with Equation 5.

6.2 Temperature Scaling

Figure 3 shows ECE before vs. after temperature scaling for all 40 classification cells. 39 of 40 cells fall below the diagonal (mean ECE $0.142 \rightarrow 0.049$). $T^\star > 1$ in every cell, confirming

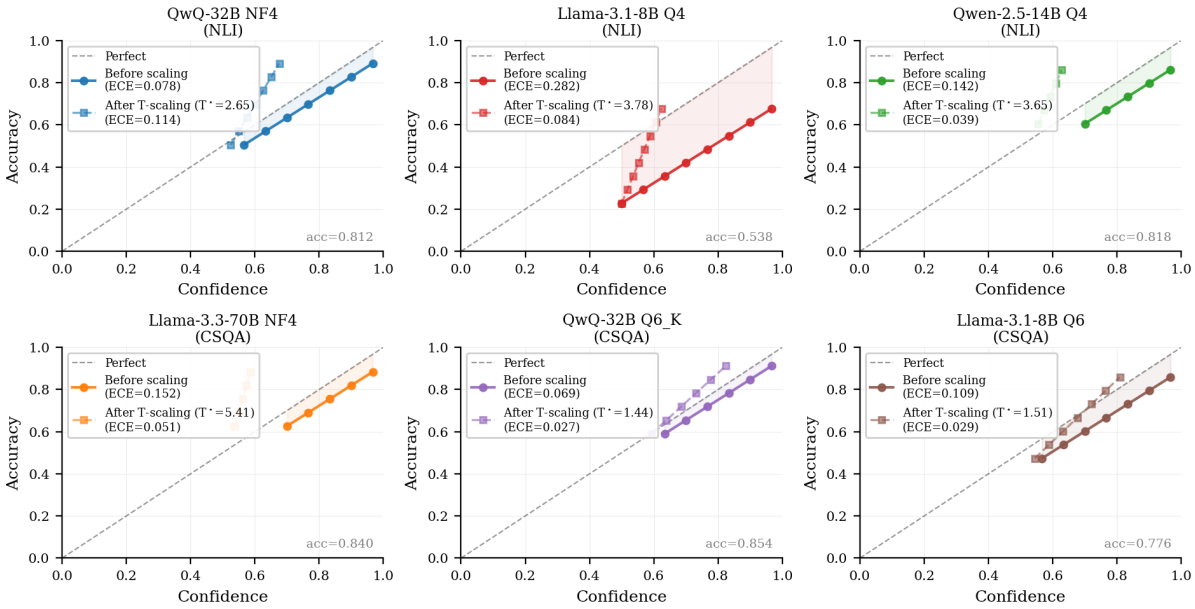

Figure 1: Reliability diagrams for six representative classification cells (3 NLI, 3 CSQA). Solid circles: pre-temperature-scaling; open squares: post-scaling. Most cells move substantially closer to the diagonal after scaling, reflecting systematic LLM overconfidence. Panels span the range from near-well-calibrated baselines to strongly overconfident cells where compression improves calibration (negative CCD).

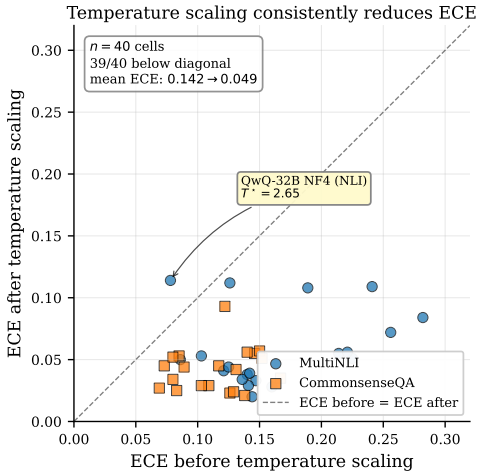

Figure 3: Temperature scaling effectiveness across all 40 classification cells. 39 of 40 fall below the $y = x$ diagonal; the lone outlier (QwQ-32B NF4 NLI) was already best-calibrated pre-scaling and over-correction at $T^\star = 2.65$ slightly inflates its ECE.

Table 4: Llama-3.1-8B at fixed deployment size, with analytical 95% CIs on ECE. * flagged as anomalous (§5). Quantized cells use n=500; distilled uses n=1000–2000.

Task	Metric	NF4	Q4	Distilled
NLI	Acc	0.648	0.538	**0.692**
	ECE	$0.225_{[.18,.27]}$	$0.282_{[.24,.33]}$	$0.221_{[.20,.24]}$
	ECE@$T^\star$	$\mathbf{0.051}_{[.01,.09]}$	$0.084_{[.04,.13]}$	$0.119_{[.10,.14]}$
CSQA	Acc	0.764	0.776	**0.781**
	ECE	$0.073_{[.04,.11]}$	$0.080_{[.04,.12]}$	$0.094_{[.07,.12]}$
	ECE@$T^\star$	$0.045_{[.01,.08]}$	$0.034_{[.00,.07]}$	$\mathbf{0.023}_{[.00,.05]}$
QA	EM	0.426	0.326	**0.676**
	ECE	$0.078_{[.04,.12]}$	$0.066_{[.03,.11]}$	$0.500_{[.48,.52]}$
XSum	R-1	0.160^*	0.160^*	**0.296**
	ECE	$0.236^*_{[.20,.27]}$	$0.176_{[.14,.21]}$	$\mathbf{0.069}_{[.05,.09]}$

systematic LLM overconfidence.

6.3 Distillation versus Quantization at Fixed Deployment Size

A natural practitioner question: *at a fixed deployment-size budget, should I quantize the large model or distill it into a smaller one?* We address this using our Llama-3.3-70B→Llama-3.1-8B experiment: both the quantized and distilled Llama-3.1-8B deploy at the same VRAM footprint.

We make the following claim, supported by the analytical CIs in Table 4: on XSum, the distilled student's calibration improvement is well outside the n=500 binomial noise floor (NF4 ECE 0.236 [0.20, 0.27] vs. distilled 0.069 [0.05, 0.09], disjoint 95% CIs), in addition to a substantial accuracy gain (R-1 0.160 → 0.296). On CommonsenseQA after temperature scaling, the distilled student's calibration improvement of 0.022 lies within the noise floor and is consistent in sign across both NF4 and Q4 quantized baselines; we report this as suggestive rather than statistically established. On MultiNLI, distillation wins on accuracy by 4.4

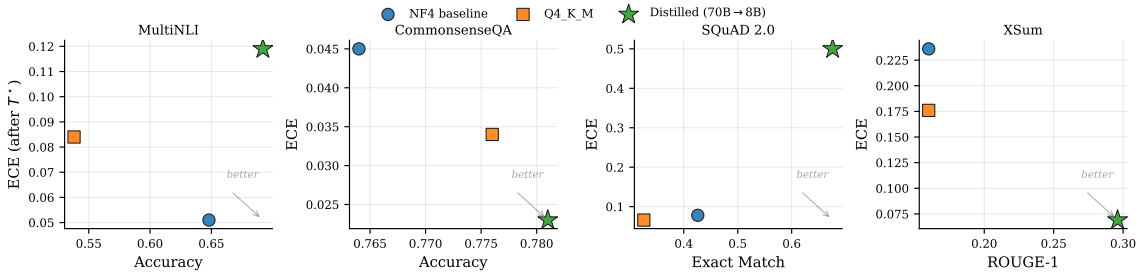

Figure 4: Llama-3.1-8B accuracy-vs-ECE per task (classification: ECE@$T^\star$; generative: raw ECE). Distilled student (green star) matches or improves same-size quantized baselines on 3 of 4 tasks; wins accuracy but loses calibration on MultiNLI after T-scaling; SQuAD 2.0 ECE proxy precludes a confident comparison.

points but *loses* on calibration after T-scaling (CIs of NF4 $[0.01, 0.09]$ and distilled $[0.10, 0.14]$ are essentially disjoint), an accuracy-calibration trade-off. On SQuAD 2.0, distillation dramatically improves Exact Match (+25 points), but the calibration measurement is unreliable due to sequence-log-prob proxy fragility.

Robustness across distillation pairs. Three additional configurations broadly reproduce the same pattern: within-family distillation preserves calibration best, MoE→dense transfer is strong on classification but unstable on QA, and cross-family distillation degrades calibration more noticeably on NLI. Across all four pairs, accuracy gains over same-size quantized baselines are positive in 10 of 12 task-pair cells that pass anomaly criteria.

Per-Task Quantization Results Full tables are in Appendix C. We discuss only cell-level CCDs whose magnitude exceeds 0.10 and whose underlying ECE difference exceeds 0.03 (roughly 2σ); smaller comparisons are reported for completeness but are not treated as individually informative.

7 Discussion

The framework in §4 is qualitative. Heuristic Decomposition 1 rests on three idealizations (bin stability, perturbation symmetry, pointwise-to-population concentration;) and we do not present it as a proven theorem. The empirical consistency we observe supports the framework's qualitative usefulness, not a precise law (moderate CV $R^2 = 0.33$; sign-prediction accuracy 75%). The framework explains the sign-mixed pattern, the per-task ordering of slope strength, the null result on NLI, and sign-flip for high-overconfidence cells; it does not predict cell-level magnitudes or model-specific deviations.

Why QA carries the regression's statistical power. LOTO analysis (§4.3) shows removing QA collapses β_1 to non-significance — a prediction of P2, since QA has the widest baseline-ECE range. However, QA also has the most fragile confidence proxy (§9), so the framework's quantitative fit on QA cannot be cleanly separated from proxy artifacts. The qualitative pattern — negative slopes on three of four tasks ordered by baseline-ECE range — holds independently of QA.

8 Conclusion

Compression does not have a uniform effect on calibration. Across 53 clean non-baseline cells, most compressed models improve rather than degrade ECE, but the sign and magnitude of the effect depend strongly on task and baseline overconfidence. The proposed CCD metric and softmax-perturbation framework organize this heterogeneity: highly overconfident baselines tend to soften under compression, while already well-calibrated baselines are more exposed to noise.

The empirical patterns support this qualitative view. Pooled regression shows a strong negative association between baseline overconfidence and CCD, while per-task analyses show that the effect is easiest to detect when baseline ECE varies substantially within the task. At fixed hardware footprint, sequence-level distillation is also competitive with same-size quantization on the joint accuracy–calibration frontier after temperature scaling.

These findings argue that calibration should be re-measured after any compression step rather than assumed to track accuracy. Compression can improve calibration in useful regimes, but the gains are conditional and must be validated for the target task, confidence proxy, and deployment constraint.

9 Limitations

Confidence proxy. Model confidence on generative tasks uses sigmoid-mapped length-normalized log-probability (Tian et al., 2023), which is fragile: two anomalous cells trace to proxy failures, not compression effects. The SQuAD 2.0 results are particularly affected: large negative CCDs coincide with significant accuracy drops, and the proxy cannot distinguish "logits softened" from "model degraded." A learned semantic calibrator (Kuhn et al., 2023) is the most important methodological extension for future work.

Sample size. ECE at n=500 carries an analytical SE of $\approx$ 0.015–0.020. Individual cell-level CCD comparisons are unreliable; we rely on aggregated patterns over ~26,500 examples and report cell-level CCDs only when $|\mathrm{CCD}| > 0.10$ and $|\Delta\mathrm{ECE}| > 0.03$. Future work should use $n \geq 1500$ with ≥ 3 seeds per cell.

CCD robustness. Conclusions hold under raw ΔECE ($r = -0.69$, $p < 10^{-8}$) and Brier score drift (sign agrees in 82% of classification cells); they are not artifacts of the CCD denominator.

Distillation scope. Experiments use sequence-level KD only; calibration transfer is implicit rather than direct (Hinton et al., 2015; Wang et al., 2020). A sequence-level vs. logit-level KD comparison on calibration remains open.

10 Ethical Statement

This work studies model compression techniques that are already in wide use. All experiments were conducted on publicly available pretrained models and datasets, and all model and dataset artifacts were used in accordance with their stated licenses, terms of use, and redistribution restrictions. We do not introduce new capabilities that raise direct ethical concerns. Compression can reduce inference energy cost, which is a benefit; however, our findings that calibration degrades in some regimes imply that compressed models deployed in high-stakes settings (medical triage, legal routing) should be re-evaluated for calibration after compression, not only for accuracy. We used AI assistants (Claude) for writing assistance and coding support during the preparation of this work.

References

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *arXiv preprint arXiv:2305.14314*.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning (ICML)*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*. NeurIPS Deep Learning Workshop.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. 2017. Quantized neural networks: Training neural networks with low precision weights and activations. *Journal of Machine Learning Research*, 18.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? On the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and Jared Kaplan. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *Preprint*, arXiv:2302.09664.

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 LLM compression and acceleration. *Preprint*, arXiv:2306.00978. Proceedings of Machine Learning and Systems (MLSys).

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. 2023. LLM-QAT: Data-free quantization aware training for large language models. *Preprint*, arXiv:2305.17888.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. *Preprint*, arXiv:2106.07998.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. 2019. When does label smoothing help? *Preprint*, arXiv:1906.02629.

Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. 2019. Measuring calibration in deep learning. *Preprint*, arXiv:1904.01685. CVPR Workshops.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *Preprint*, arXiv:1701.06548.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2022. Prompting GPT-3 to be reliable. *Preprint*, arXiv:2210.09150.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *Preprint*, arXiv:2305.14975.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

A Elementary Derivation of Theorem 1

This result is standard; we present the derivation from the softmax Jacobian for completeness. The argument proceeds in three steps.

Step 1: First-order expansion. The Jacobian of softmax has entries $J_{ij} = p_i(\delta_{ij} - p_j)$ (Bishop, 2006). By Taylor expansion,

$$\begin{aligned}\Delta p_{i^\star} &= p_{i^\star}\left(\Delta z_{i^\star} - \textstyle\sum_j p_j \Delta z_j\right) + R(\Delta z) \\ &= p_{i^\star}(\Delta z_{i^\star} - \overline{\Delta z}) + R(\Delta z).\end{aligned}$$

Step 2: Bounding the linear term. Separating the $i^\star$ contribution from $\overline{\Delta z}$: $\overline{\Delta z} = p_{i^\star}\Delta z_{i^\star} + (1 - p_{i^\star})\overline{\Delta z}_{-i^\star}$, where $\overline{\Delta z}_{-i^\star}$ is the p-weighted average over non-top-1 coordinates. Substituting:

$$\Delta z_{i^\star} - \overline{\Delta z} = (1 - p_{i^\star})(\Delta z_{i^\star} - \overline{\Delta z}_{-i^\star}).$$

Combining with the leading $p_{i^\star}$ factor:

$$p_{i^\star}(\Delta z_{i^\star} - \overline{\Delta z}) = p_{i^\star}(1 - p_{i^\star})(\Delta z_{i^\star} - \overline{\Delta z}_{-i^\star}).$$

Since $|\Delta z_{i^\star}|, |\overline{\Delta z}_{-i^\star}| \le \delta$, the bracket has magnitude $\le 2\delta$, so the linear part is bounded by $2p_{i^\star}(1 - p_{i^\star})\delta$.

Step 3: Remainder. The second-derivative tensor of softmax has bounded entries, so $|R(\Delta z)| \le cV\delta^2$ for some constant $c \le 5/2$. Hence $|\Delta p_{i^\star}| \le 2p_{i^\star}(1 - p_{i^\star})\delta + O(\delta^2)$. □

Tightness. The bound is order-optimal: equality (up to $O(\delta^2)$) holds for the worst-case perturbation $\Delta z_{i^\star} = +\delta$, $\Delta z_j = -\delta$ for $j \ne i^\star$. Typical random quantization perturbations have a smaller absolute constant in expectation, with the same $p(1 - p)$ dependence.

B Proof Sketch for Heuristic Decomposition 1

We present the decomposition as a motivated inequality rather than a fully rigorous theorem; three idealizations preclude a tight proof (bin-membership stability under perturbation; symmetry of quantization perturbation distribution; pointwise-to-population concentration).

Step 1: Decomposing the change in ECE. Writing $\Delta p(x) := p^{(c)}_{i^\star}(x) - p^{(b)}_{i^\star}(x)$ and $r(x) := \mathrm{acc}(B_m(x))$, the reverse triangle inequality gives

$$\begin{aligned}\mathrm{ECE}(M_c) - \mathrm{ECE}(M_b) \le\ & \mathbb{E}_x[\mathrm{sign}(p^{(b)}_{i^\star} - r) \cdot \Delta p] \\ & + \mathbb{E}_x[|\Delta p|].\end{aligned}$$

Step 2: Directional term. For symmetric quantization perturbations, the expected $\Delta p_{i^\star}$ is biased *negatively* (toward $1/V$): the top-1 contribution to

$\overline{\Delta z}$ pulls the average toward itself. When the baseline is more often overconfident than underconfident, $\mathrm{sign}(p_{i^\star}^{(b)} - r)$ is positive in expectation; combined with negative Δp this term becomes negative, with magnitude scaling as $-c_1(\mathcal{T}) \cdot [\mathrm{ECE}(M_b) - \mathrm{ECE}_{\mathcal{T}}^{\star}]$.

Step 3: Noise term. The second term scales with $\delta \cdot \mathbb{E}_x[p_{i^\star}^{(b)}(1 - p_{i^\star}^{(b)})]$, which depends on task structure through the typical $p_{i^\star}$ distribution. This gives $c_2(\mathcal{T}, \delta)$. $\square$

Connection to label smoothing. The negative directional bias under symmetric logit perturbation parallels label smoothing (Müller et al., 2019): quantization can be viewed as implicit label smoothing applied to a trained model — beneficial when overconfident, detrimental when not.

C Full Per-Cell Results

Tables 5–8 report all 77 compression-grid cells ($3 \times 20 + 1 \times 17$): the three classification/extraction tasks each have 20 cells (all 6 models $\times$ up to 5 compression levels, minus hardware-infeasible entries), while XSum has 17 because QwQ-32B is excluded — it is a reasoning model whose chain-of-thought generation style is not appropriate for abstractive summarization evaluation. The 4 Llama-3.3-70B→Llama-3.1-8B distillation cells are in Table 4 of the main paper; remaining distillation cells are released with the codebase. Anomalous cells are marked † and excluded from headline analysis.

Table 5: MultiNLI results. $T^\star$ = optimal temperature scaling parameter.

Model	Method	Performance	ECE	MCE	CCD	$T^\star$
Gemma-2-27B	NF4	acc=0.728	0.126	0.435	0.000	2.12
	Q6_K	acc=0.810	0.140	0.573	$+0.111$	2.25
	Q4_K_M	acc=0.810	0.147	0.698	$+0.167$	2.34
	Q3_K_M	acc=0.808	0.141	0.329	$+0.119$	2.28
Gemma-4-31B	NF4	acc=0.836	0.136	0.441	0.000	3.98
Llama-3.1-8B	NF4	acc=0.648	0.225	0.378	0.000	2.77
	INT8	acc=0.560	0.256	0.599	$+0.138$	3.41
	Q6_K	acc=0.630	0.189	0.311	-0.160	2.99
	Q4_K_M	acc=0.538	0.282	0.597	$+0.253$	3.78
	Q3_K_M	acc=0.552	0.214	0.412	-0.049	2.47
Llama-3.3-70B	NF4	acc=0.804	0.144	0.266	0.000	3.11
	Q4_K_M	acc=0.832	0.121	0.453	-0.160	2.71
	Q3_K_M	acc=0.830	0.125	0.447	-0.132	2.60
QwQ-32B	NF4	acc=0.812	0.078	0.253	0.000	2.65
	Q6_K	acc=0.764	0.103	0.318	$+0.320$	1.50
	Q4_K_M	acc=0.774	0.086	0.385	$+0.103$	1.41
Qwen-2.5-14B	NF4	acc=0.704	0.241	0.486	0.000	3.89
	Q6_K	acc=0.800	0.160	0.476	-0.336	3.55
	Q4_K_M	acc=0.818	0.142	0.348	-0.411	3.65
	Q3_K_M	acc=0.732	0.221	0.576	-0.083	4.44

Table 6: CommonsenseQA results.

Model	Method	Performance	ECE	MCE	CCD	$T^\star$
Gemma-2-27B	NF4	acc=0.736	0.167	0.525	0.000	1.87
	Q6_K	acc=0.816	0.146	0.629	-0.126	2.04
	Q4_K_M	acc=0.816	0.140	0.424	-0.162	2.06
	Q3_K_M	acc=0.814	0.150	0.611	-0.102	2.02
Gemma-4-31B	NF4	acc=0.846	0.122	0.402	0.000	3.13
Llama-3.1-8B	NF4	acc=0.764	0.073	0.249	0.000	1.36
	INT8	acc=0.780	0.089	0.725	$+0.219$	1.40
	Q6_K	acc=0.776	0.109	0.680	$+0.493$	1.51
	Q4_K_M	acc=0.776	0.080	0.220	$+0.096$	1.32
	Q3_K_M	acc=0.778	0.085	0.197	$+0.164$	1.34
Llama-3.3-70B	NF4	acc=0.840	0.152	0.777	0.000	5.41
	Q4_K_M	acc=0.826	0.117	0.599	-0.230	2.01
	Q3_K_M	acc=0.838	0.103	0.460	-0.322	1.72
QwQ-32B	NF4	acc=0.850	0.083	0.462	0.000	2.03
	Q6_K	acc=0.854	0.069	0.632	-0.169	1.44
	Q4_K_M	acc=0.844	0.080	0.456	-0.036	1.43
Qwen-2.5-14B	NF4	acc=0.844	0.126	0.389	0.000	3.91
	Q6_K	acc=0.846	0.129	0.480	$+0.024$	3.70
	Q4_K_M	acc=0.838	0.131	0.457	$+0.040$	3.82
	Q3_K_M	acc=0.826	0.138	0.454	$+0.095$	3.71

Table 7: SQuAD 2.0 results. No temperature scaling for generative tasks.

Model	Method	EM / F1	ECE	MCE	CCD
Gemma-2-27B	NF4	0.530 / 0.580	0.110	0.272	0.000
	Q6_K	0.586 / 0.628	0.141	0.548	+0.282
	Q4_K_M	0.568 / 0.613	0.145	0.380	+0.318
	Q3_K_M	0.560 / 0.613	0.099	0.179	−0.100
Gemma-4-31B	NF4†	0.776 / 0.833	0.752	0.829	0.000
Llama-3.1-8B	NF4	0.426 / 0.475	0.078	0.122	0.000
	INT8	0.342 / 0.389	0.076	0.479	−0.026
	Q6_K	0.336 / 0.386	0.055	0.187	−0.295
	Q4_K_M	0.326 / 0.374	0.066	0.469	−0.154
	Q3_K_M	0.356 / 0.402	0.059	0.878	−0.244
Llama-3.3-70B	NF4	0.610 / 0.691	0.577	0.695	0.000
	Q4_K_M	0.462 / 0.534	0.134	0.438	−0.768
	Q3_K_M	0.436 / 0.501	0.144	0.350	−0.750
QwQ-32B	NF4	0.626 / 0.681	0.363	0.776	0.000
	Q6_K	0.386 / 0.432	0.054	0.309	−0.851
	Q4_K_M	0.402 / 0.449	0.064	0.284	−0.824
Qwen-2.5-14B	NF4	0.468 / 0.517	0.084	0.361	0.000
	Q6_K	0.462 / 0.510	0.112	0.403	+0.333
	Q4_K_M	0.464 / 0.513	0.089	0.395	+0.055
	Q3_K_M	0.452 / 0.500	0.088	0.481	+0.043

Table 8: XSum results. ROUGE-1 (R-1) reported for task performance. Anomalous cells marked †.

Model	Method	R-1 / BERTScore	ECE	MCE	CCD
Gemma-2-27B	NF4	0.315 / 0.894	0.076	0.215	0.000
	Q6_K	0.323 / 0.896	0.070	0.193	−0.080
	Q4_K_M	0.325 / 0.896	0.068	0.188	−0.111
	Q3_K_M	0.330 / 0.897	0.056	0.173	−0.265
Gemma-4-31B	NF4	0.301 / 0.890	0.094	0.241	0.000
Llama-3.1-8B	NF4†	0.160 / 0.847	0.236	0.312	0.000
	INT8†	0.302 / 0.888	0.048	0.109	−0.798
	Q6_K	0.163 / 0.849	0.183	0.298	−0.227
	Q4_K_M	0.160 / 0.849	0.176	0.287	−0.256
	Q3_K_M	0.164 / 0.848	0.167	0.275	−0.294
Llama-3.3-70B	NF4	0.309 / 0.890	0.063	0.198	0.000
	Q4_K_M	0.256 / 0.876	0.117	0.251	+0.846
	Q3_K_M	0.256 / 0.876	0.108	0.239	+0.704
Qwen-2.5-14B	NF4	0.160 / 0.852	0.238	0.304	0.000
	Q6_K	0.155 / 0.853	0.200	0.261	−0.158
	Q4_K_M	0.157 / 0.853	0.199	0.259	−0.162
	Q3_K_M	0.155 / 0.853	0.207	0.268	−0.130

D Regression Diagnostics

We fit (6) on $n = 53$ clean non-baseline cells. $\mathrm{ECE}^{\star}_{\mathcal{T}}$ is set to the per-task minimum baseline ECE: 0.078 (NLI), 0.073 (CSQA), 0.078 (QA), 0.063 (XSum). Bits-per-weight: NF4 = 4.0, INT8 = 8.0, Q6_K = 6.5, Q4_K_M = 4.5, Q3_K_M = 3.5.

Table 9: Pooled regression results. Bootstrap CIs from B=2000 resamples. The pooled coefficient aggregates the per-task slopes in Table 1.

Predictor	Estimate	95% CI	Pearson r
overconf. gap	-2.005	$[-2.82, -1.46]$	-0.695
$\tau = \log K$	-0.004	$[-0.019, +0.011]$	-0.180
log(bpw)	$+0.024$	$[-0.21, +0.25]$	$+0.089$
intercept	$+0.105$	$[-0.18, +0.41]$	

Table 10: Leave-one-task-out sensitivity. β_1 remains negative and significant when NLI, CSQA, or XSum is held out; it collapses when SQuAD 2.0 is held out, consistent with the framework's Prediction P2 (§4.2) that statistical power concentrates on tasks with wider baseline-ECE variation.

Held-out task	n	β_1	p
(none, full)	53	-2.005	$< 10^{-8}$
MultiNLI	39	-1.98	$< 10^{-4}$
CommonsenseQA	39	-1.69	0.001
SQuAD 2.0	39	-0.82	0.37
XSum	42	-1.46	0.002

Table 11: Robustness of regression coefficients to alternative encodings of task structure. The overconfidence-gap coefficient β_1 is stable across all variants; the task-structure coefficient is never significant.

τ encoding	β_1	β_2	R^2
$\log K$, K_g=50000 [main]	-2.005	$-0.004^{\text{n.s.}}$	0.487
$\log K$, K_g=1000	-2.005	$-0.007^{\text{n.s.}}$	0.487
$\log K$, K_g=100	-2.004	$-0.014^{\text{n.s.}}$	0.488
$\log K$, K_g=50	-2.004	$-0.018^{\text{n.s.}}$	0.488
binary (clf/gen)	-2.006	$-0.037^{\text{n.s.}}$	0.486
+ interaction	-1.885	$-0.003^{\text{n.s.}}$	0.487

E Model Selection Justification

The six models were selected to span three practically relevant axes: (1) **parameter scale** (8B, 14B, 27–32B, 70B), covering both on-device and server-side deployment regimes; (2) **model family** (Llama, Qwen, Gemma), ensuring findings are not artifacts of a single architecture or training recipe; and (3) **architectural variant**, including a chain-of-thought reasoning model (QwQ-32B) and a multimodal model (Gemma-4-31B) to test robustness at the boundary of standard instruction-tuned behavior. Llama-3.3-70B is included as both a standalone evaluation target and a distillation teacher, enabling a direct comparison of quantized-large vs. distilled-small at matched deployment size. QwQ-32B is excluded from XSum because its chain-of-thought generation style produces outputs structurally different from standard abstractive summaries, making ROUGE-1-based correctness and ECE estimates unreliable for this model on this task. No model was excluded from classification or extractive QA.

F Summarization Threshold Sensitivity

For XSum we compute ECE using a binary correctness indicator: a generated summary is treated as "correct" if its ROUGE-1 F1 with the reference exceeds a threshold θ. The headline results in §6 use $\theta = 0.1$, following Si et al. (2022).

We verified sign preservation on representative cells across $\theta \in \{0.05, 0.10, 0.15, 0.20\}$. For the Llama-3.3-70B XSum series (Q4_K_M, Q3_K_M) the CCD sign remains positive across the tested range; for the Gemma-2-27B and Qwen-2.5-14B series the sign remains negative. Absolute CCD magnitudes vary by up to 30% across thresholds, but the qualitative direction in each cell is robust. A full per-example threshold sweep over all summarization cells is available in the released code.