

Displacement Is Not Direction: Evaluating Fidelity Metrics for Quantized LLM Deployment

Miloš Nikolić^{1,2}, Ali Hadi Zadeh^{1,2}, Enrique Torres Sanchez^{1,2}, Andreas Moshovos^{1,2,3}

¹ByteShape, ²University of Toronto, ³Vector Institute for Artificial Intelligence

Correspondence: contact@byteshape.com

Abstract

Fidelity metrics such as per-token KL divergence (KLD) against a high-precision reference are often used as low-cost proxies for benchmark quality. We test this practice on 28 Qwen3.6-35B-A3B and 41 Devstral-Small-2-24B quantizations across downstream benchmarks. KLD strongly correlates with benchmark score over the full cohorts ($\rho = -0.72$ on Qwen and $\rho = -0.86$ on Devstral, both with $p < 0.001$). However, among quantizations retaining at least 99% of measured BF16 composite quality, KLD provides no dependable ranking signal ($\rho = +0.11$, $p = 0.71$, $n = 14$ on Qwen; $\rho = +0.02$, $p = 0.93$, $n = 16$ on Devstral). This collapse persists across measurement variants spanning KLD aggregation, perplexity formulations, top-1 agreement, measurement corpora, and context lengths. At the per-prompt level, KLD provides only weak failure-prediction power and fails as a cross-model router across four benchmarks, selecting the better response on only 40.0–49.1% of disagreement prompts and reducing score relative to the better static model on all four. We trace the collapse to a structural decomposition: within the KLD-defined silent zone, KLD primarily measures the *volume* of disagreement with the reference (composite $\rho = +0.94$, $p < 0.001$ on Qwen; $\rho = +0.55$, $p = 0.03$ on Devstral), while its relationship to the *direction* of those disagreements is weak and task-conditional.

1 Introduction

As large language models (LLMs) become more capable, the motivation to deploy them across settings ranging from edge devices to data centers continues to grow. Much of this progress has been accompanied by rapidly increasing parameter counts, from several billion to a trillion. Larger models, however, require correspondingly greater memory capacity and bandwidth, compute, and energy.

Quantization and compression offer a practical way to reduce these deployment costs, increasing throughput in data centers and often making edge deployment possible. Yet selecting a quantized model is not straightforward. For Qwen3.6-35B-A3B (Qwen, 2026) alone, the community has released hundreds of GGUF quantizations from different publishers, using different calibration recipes and exhibiting widely varying output quality. Running a full benchmark suite on every candidate is expensive, so practitioners often rely on lower-cost fidelity metrics, particularly KL divergence (KLD) and perplexity (PPL) relative to a BF16 reference.

This work examines the relationship between fidelity metrics, such as PPL and KLD, and downstream benchmark quality. At first glance, these metrics appear to work: across the full cohort, KLD and PPL correlate strongly with composite benchmark score, with Spearman $\rho(\text{KLD}, \text{composite quality}) = -0.72$ on Qwen3.6-35B-A3B and -0.86 on Devstral-Small-2-24B (both with $p < 0.001$), and similar values across every fidelity-metric variant we tested.

However, this signal is not uniform across the quality range. It is driven primarily by clearly degraded quantizations, where KLD and PPL span a wide range and benchmark quality has already fallen substantially. When we restrict the analysis to near-baseline candidates, those retaining at least 99% of measured BF16 composite quality and therefore most relevant to practitioners, we find no dependable KLD ranking signal, with $\rho = +0.11$ ($p = 0.71$) on Qwen and $\rho = +0.02$ ($p = 0.93$) on Devstral. The correlation remains non-significant through a 2.0% quality drop on Devstral and through 2.5% on Qwen. We use the term *silent zone* for the near-baseline regime in which KLD and PPL continue to measure distance from the reference but no longer provide a dependable ranking signal for downstream quality. We examine this regime under two complementary definitions:

a quality-defined view based on retained BF16 composite quality and a KLD-defined view, in which models form a dense low-divergence cluster followed by a progressively degraded tail. Beyond the silent zone, in the *lossy zone*, KLD/PPL remain useful as coarse degradation detectors because the models have already suffered significant quality loss.

Inside the silent zone, fidelity metrics fail at three levels: they fail to (1) reliably rank models, (2) separate correct from incorrect responses, or (3) route a prompt to the better of two models when models disagree. We explain this with a volume–direction decomposition of score differences: KLD tracks how often the quantized model and reference differ in correctness, but task quality also depends on whether those differences are improvements or regressions.

In short, we make the following contributions:

1. **Silent-zone vs lossy-zone characterization:** We identify a near-baseline *silent zone* where fidelity metrics like KLD and PPL lose ranking power, and a degraded *lossy zone* where they remain useful as coarse predictors (Section 4).
2. **Decompose score into volume and direction:** We build on the work of Dutta et al. (2024) by extending the observed flips into *leapfrogs* (quantized model is correct where the reference is not) and *drops* (the reverse), and derive a closed-form decomposition $\text{score} = \text{score}_{\text{ref}} + \text{vol} \cdot (2f - 1)/N$. KLD primarily tracks volume; whether it also tracks direction is task-conditional (Section 5).
3. **Per-prompt KLD has no useful model-selection signal:** KLD only weakly separates a model’s own passes from its failures and fails to route between models, both in detailed LiveCodeBench pairings and across four prompt-level benchmarks (Section 4.4).
4. **The collapse is metric-invariant:** The non-significant silent-zone correlation ($|\rho| \leq 0.25$, all $p > 0.10$) holds across 14 measurement variants, including PPL ratio, $\Delta \log \text{PPL}$, top-1 agreement, longer contexts, and response-only measurement on held-out instruction data (Section 4.2).

2 Related Work

Quantization is widely used to reduce the memory, bandwidth, and compute cost of deploying large

language models. To preserve quality at lower bit-widths, prior work has introduced practical post-training quantization methods such as GPTQ, AWQ, and SmoothQuant (Frantar et al., 2023; Lin et al., 2024; Xiao et al., 2023). Further compression can be achieved with mixed-precision quantization, where precision is assigned at finer granularity than the whole network, including by tensor, layer, channel, or feature (Zheng et al., 2026; Nikolić et al., 2024). More recent studies evaluate the accuracy-performance trade-offs of quantized LLMs across model families, bit-widths, hardware targets, and downstream tasks (Lee et al., 2025; Jaiswal et al., 2024; Kurtic et al., 2025). These works show that quantized-model quality cannot be summarized by model size alone, and that the practical value of a quantized model depends on the interaction between quality, throughput, hardware, and workload. Our work studies a narrower question: given multiple quantized variants of the same model, can inexpensive fidelity metrics such as KLD and PPL be used to rank their downstream task quality?

2.1 Beyond PPL and KLD

The limitations of PPL and related fidelity metrics for evaluating compressed models have been noted before. Jaiswal et al. (2024) show that pruned and quantized models can preserve perplexity while losing downstream capabilities, and introduce a broader benchmark suite for compressed models. Deiseroth et al. (2024) similarly argue that perplexity and aggregate accuracy do not fully capture generation degradation, and propose divergent-token metrics as a more generation-sensitive diagnostic for pruning and quantization. However, even though fidelity metrics are imperfect, their low computational cost keeps them attractive, especially in open-source model comparison. We build on this background, but focus specifically on model selection: whether fidelity metrics can distinguish among near-baseline quantized models, separate correct from incorrect responses, or route a prompt to the better of two quantized models.

2.2 Flips

Dutta et al. (2024) present the closest prior art to our work. They introduce *flips*: the percentage of multiple-choice answers that change correctness between a baseline model and a quantized model. They study the relationship between KLD, flips, and MMLU accuracy across quantized models, showing that accuracy alone can hide substantial

Table 1: Two-cohort experimental setup. Zone statistics use the KLD-defined partition.

Item	Qwen cohort	Devstral cohort
Reference model	Qwen3.6-35B-A3B	Devstral-Small-2-24B
Architecture	MoE (35B/3B active)	dense (24B)
Specialization	general	code-specialized
Cohort size	28 quants	41 quants
Quant sources	5	3
bpw range	2.17–6.14	1.88–9.84
Silent-zone τ	$\text{KLD} \leq 0.064$	$\text{KLD} \leq 0.027$
Silent/lossy split	17 / 11	16 / 25

behavior changes after quantization.

We extend this framework in five ways:

1. We extend the analysis beyond MMLU and multiple-choice accuracy to a broader set of tasks, including coding, tool calling, instruction following, and math reasoning.
2. We split flips into *leapfrogs* (quantized model correct, reference wrong) and *drops* (reference correct, quantized model wrong), and decompose score as $\text{score}_{\text{ref}} + \text{vol} \cdot (2f - 1)/N$.
3. We show that KLD consistently tracks disagreement volume, while its relationship to direction depends on the task and quality regime.
4. We characterize the silent zone of near-baseline low-KLD quants, a regime where we find no dependable KLD–quality ranking signal.
5. We evaluate KLD at the per-prompt level and as a routing signal, showing that it provides only a weak response-correctness signal and fails to select the better model across four prompt-level benchmarks.

Additionally, Dutta et al. (2024) introduce top-margin analysis (the probability gap between the model’s first and second choices). This is a per-prompt instability indicator. Our analysis focuses on whether fidelity to the reference predicts downstream task quality among quantized models.

3 Methodology

We evaluate two quantized-model cohorts with deliberately different architectures (Table 1): 28 GGUF quantizations of Qwen3.6-35B-A3B (Qwen, 2026), a mixture-of-experts model, and 41 quantizations of the dense, code-specialized Devstral-Small-2-24B-Instruct-2512 (Mistral AI, 2025). All fidelity metrics and leapfrog/drop counts are computed relative to the corresponding BF16 reference.

The cohorts draw from multiple public quantization sources and span a broad range of compression, from near-reference high-bpw variants to aggres-

sively compressed low-bpw ones. We use bits per weight (bpw) as a size proxy, not a quality metric: it sets the memory footprint, while benchmark scores and throughput determine practical quality.

3.1 Downstream Quality Evaluation

We evaluate downstream quality on benchmarks spanning coding (HumanEval (Chen et al., 2021), LiveCodeBench v6 (Jain et al., 2025)), instruction following (IFEval (Zhou et al., 2023)), tool calling (BFCL v3 (Patil et al., 2025)), math reasoning (GSM8K (Cobbe et al., 2021)), math and visual reasoning (GSM8K-v (Yuan et al., 2025)), and multi-task knowledge (MMLU (Hendrycks et al., 2021)); the Devstral cohort adds MATH-500 (Lightman et al., 2024). All benchmarks run in non-thinking mode, except GSM8K-v, which on the thinking-capable Qwen3.6 cohort we run in both modes.

We report per-benchmark scores and a composite (unweighted mean of normalized scores), oriented so higher is better and compared against the BF16 reference; per-prompt pass/fail outcomes feed the leapfrog/drop and routing analyses. Qwen MMLU is reported separately and excluded from the Qwen composite because of a prompt-format artifact, not a knowledge deficit (Appendix A). All main-cohort GGUF evaluations use the EvalScope framework (Team, 2024), serving each model with llama.cpp (Gerganov, 2023) (build b8855 for Qwen3.6 and b7744 for Devstral). The separate AWQ/GPTQ analysis uses vLLM (Section 4.3).

All reported correlation coefficients are Spearman rank correlations, with two-sided p -values reported without adjustment for multiple comparisons.

3.2 Fidelity Metrics

Our default KLD is $\text{KL}(p_{\text{BF16}} \parallel p_{\text{quant}})$ over next-token distributions. Response-only measurements score only response tokens (excluding the prompt); two diagnostics use this form with top- k truncation, defined in Appendix B: the validation-set variants in Table 3 (a held-out instruction mixture, Appendix D) and the per-prompt analyses (benchmark prompt-response pairs).

We compute PPL and KLD with llama-perplexity from the same builds. The default and headline metric is mean KLD on WikiText (Merity et al., 2017) at 512-token context, which defines the KLD-space silent-zone split. We also test 8192-token context and the response-only variants above to assess whether the

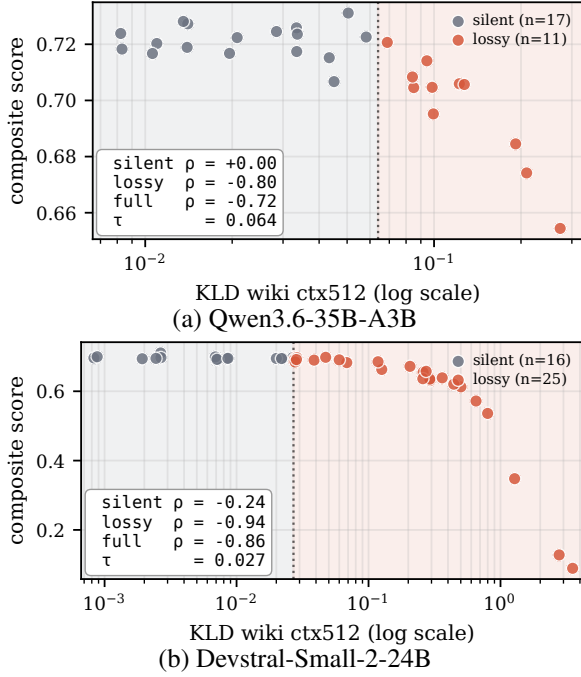


Figure 1: The silent-zone collapse on both cohorts, with Spearman correlation coefficients $\rho(\text{KLD}, \text{composite score})$ for the full cohort and the KLD-defined silent zone.

result depends on context length, measurement corpus, vocabulary support, or response-only scoring.

We report full-cohort results and analyze the near-baseline *silent zone* under separate quality-defined and KLD-defined partitions.

4 The Silent Zone

Our first finding is that KLD provides no dependable ranking signal among near-baseline quantizations. Defining near-baseline models as those retaining at least 99% of the corresponding measured BF16 composite score gives 14 Qwen and 16 Devstral quantizations. Within these sets, the KLD–quality correlation is $\rho = +0.11$ ($p = 0.71$) on Qwen and $\rho = +0.02$ ($p = 0.93$) on Devstral, compared with full-cohort correlations of -0.72 and -0.86 . The correlation remains non-significant through a 2.0% quality drop on Devstral and through 2.5% on Qwen (Appendix C).

4.1 The silent-zone collapse in KLD space

This structure is clearer in KLD space: models form a dense low-KLD cluster followed by a progressively degraded tail (Figure 1). We call the low-KLD, near-baseline region the *silent zone*, and the degraded tail the *lossy zone*. Under our default WikiText KLD measurement, the silent-zone

Table 2: Spearman $\rho(\text{KLD}, \text{benchmark score})$ by zone, on both cohorts. * $p < 0.05$, † $p < 0.10$.

Benchmark	Qwen3.6-35B-A3B			Devstral-Small-2-24B		
	silent	lossy	full	silent	lossy	full
HumanEval	+0.20	−0.91*	−0.19	−0.14	−0.81*	−0.63*
LiveCodeBench	+0.08	−0.87*	−0.65*	−0.20	−0.96*	−0.92*
IFEval	+0.18	−0.11	−0.26	−0.10	−0.87*	−0.73*
BFCL_v3	−0.24	−0.25	−0.71*	−0.12	−0.90*	−0.75*
GSM8K	+0.11	−0.61*	−0.43*	−0.05	−0.96*	−0.89*
GSM8K-v	−0.43†	−0.57†	−0.83*	−0.14	−0.88*	−0.74*
GSM8K-v think	−0.51*	−0.84*	−0.83*	—	—	—
MATH-500	—	—	—	+0.30	−0.95*	−0.80*
MMLU	−0.54*	+0.50	+0.39*	−0.22	−0.98*	−0.89*
Composite	+0.00	−0.80*	−0.72*	−0.24	−0.94*	−0.86*

boundaries are $\tau = 0.064$ for Qwen and $\tau = 0.027$ for Devstral, containing 17 and 16 models, respectively. Within the silent zone, the KLD–quality correlation is $+0.00$ on Qwen and -0.24 ($p = 0.36$) on Devstral. The boundaries are descriptive and specific to the model and KLD setup.

The low-KLD regime extends beyond these limits. Expanding the boundary keeps the within-zone correlation non-significant through $n = 21$ ($\tau = 0.0944$, $\rho = -0.35$, $p = 0.12$) on Qwen and $n = 24$ ($\tau = 0.060$, $\rho = -0.33$, $p = 0.11$) on Devstral. Significance appears only after additional degraded quantizations enter the comparison.

The pattern spans two cohorts with different architectures (MoE vs. dense), training profiles (general vs. code-specialized), and sizes (35B/3B active vs. 24B). The near-baseline result is also not driven by a single publisher: leaving each publisher out in turn gives $\rho \in [-0.26, +0.29]$ on Qwen and $\rho \in [-0.37, -0.12]$ on Devstral. It remains non-significant in the largest single-family subset, the 14 Devstral models using K-quant formats ($\rho = -0.27$, $p = 0.34$).

In the lossy zone, the negative correlation is strong: $\rho = -0.80$ ($p = 0.003$) for Qwen and $\rho = -0.94$ ($p < 0.001$) for Devstral, stronger than the full-cohort relationship in either case. Thus, the strong correlation observed over the full cohort is driven primarily by the inclusion of substantially degraded low-bit quants rather than by reliable ranking within the near-baseline candidates.

Table 2 breaks the pattern down by benchmark and shows that the silent-zone collapse is broad rather than driven by any single task.

Across most benchmarks included in the composite, the lossy zone shows a substantially stronger negative ρ than the silent zone. In the silent zone, correlations are generally non-significant or inconsistent in sign. The only included Qwen benchmark with a detectable silent-zone signal at $p < 0.05$ is GSM8K-v thinking. Qwen MMLU is reported separately but excluded from the composite because

Table 3: The KLD-defined silent-zone collapse holds across 14 measurement variants, on both cohorts. Significance: $*p < 0.05$.

Metric variant	Qwen3.6-35B-A3B		Devstral-Small-2-24B	
	full	silent	full	silent
KLD mean wiki c512	-0.72*	+0.00	-0.86*	-0.24
KLD median wiki c512	-0.69*	+0.03	-0.86*	-0.24
KLD p90 wiki c512	-0.71*	+0.01	-0.86*	-0.24
KLD p99 (tail) wiki c512	-0.75*	-0.11	-0.86*	-0.21
mean $\Delta \log$ PPL wiki c512	-0.68*	+0.09	-0.87*	-0.25
PPL ratio (q/base) wiki c512	-0.68*	+0.09	-0.87*	-0.25
PPL diff wiki c512	-0.68*	+0.09	-0.87*	-0.25
top-1 agreement % wiki c512	+0.70*	-0.04	+0.86*	+0.25
KLD wiki c8k (long context)	-0.72*	+0.00	-0.86*	-0.23
PPL ratio wiki c8k	-0.69*	+0.09	-0.86*	-0.18
top-1 agreement % wiki c8k	+0.69*	-0.03	+0.86*	+0.24
KLD val-wiki c512	-0.71*	+0.03	-0.86*	-0.17
KLD val-response top-k=60	-0.73*	-0.03	-0.86*	-0.21
KLD val-response top-k=20	-0.73*	-0.03	-0.87*	-0.21

of the prompt-format artifact (Appendix A).

4.2 Collapse is metric-invariant

We next test whether the collapse persists across fidelity-metric formulations. We evaluate 14 variants spanning KLD aggregation, PPL formulation, top-1 agreement, context length, measurement corpus, top- k truncation, and response-only measurement. Table 3 shows that all 14 variants preserve the full-cohort signal but collapse in the silent zone, with $|\rho| \leq 0.25$ ($p > 0.10$) on both cohorts. Response-only KLD on a task-aligned validation set shows the same pattern, suggesting that the collapse is not an artifact of WikiText, context length, or aggregation choice.

4.3 Preliminary evidence beyond GGUF

Our main cohorts consist of public GGUF quantizations, raising the question of whether the silent-zone behavior is specific to the llama.cpp quantization ecosystem. While quantization methods differ in how they choose scales, groups, and calibration criteria, they share the same basic goal of approximating a high-precision model with lower-precision weights. More broadly, the volume-direction explanation does not depend on quantization itself. Pruning, LoRA, fine-tuning, and other model modifications can also move a model away from a reference distribution. A fidelity metric can measure the size of that change, but not necessarily whether it helps or hurts the downstream task. We therefore expect the same distinction between displacement and direction to apply more broadly, although this requires empirical validation.

We preliminarily test four vLLM-served Qwen variants (three AWQ and one GPTQ) from four publishers. All are near baseline and span only 0.54 percentage points in downstream quality. Across

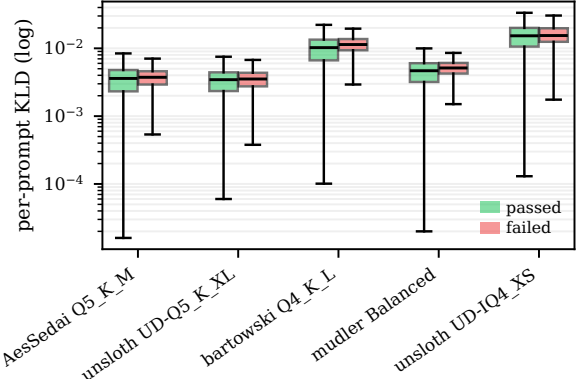


Figure 2: Per-prompt KLD distribution on LiveCodeBench, split by each model’s own pass (green) / fail (red) outcomes, for five KLD-defined silent-zone Qwen quants.

six pairs, KLD orders quality correctly in only 2, including an inversion for the pair with the largest KLD difference.

Four models are too few to establish that the silent-zone failure generalizes beyond GGUF. This result is preliminary evidence that the same ranking limitation may also occur among near-baseline AWQ and GPTQ models. A larger cross-format study, as well as evaluation of pruning, LoRA, and fine-tuning, is needed to test that hypothesis.

Appendix H lists the models and their measured KLD and composite scores.

4.4 Per-prompt KLD analysis

To test whether aggregate KLD hides a stronger prompt-level signal, we evaluate KLD directly on benchmark prompts. We compute KLD on each response and compare it with the corresponding benchmark pass/fail outcome. If KLD had useful predictive power in the *silent zone*, we should be able to identify which prompts a given quantized model will fail.

We compute per-prompt KLD on LiveCodeBench (LCB) for five silent-zone Qwen quantizations and split each model’s prompts into those it passed and those it failed. Figure 2 shows substantial overlap: the passed and failed interquartile ranges overlap for all five models. Failed prompts have slightly higher KLD on average, with failed-to-passed geometric-mean ratios from 1.08 to 1.22 (13% average increase), but this small aggregate shift provides little prompt-level separation.

A stronger test is whether per-prompt KLD can choose between two quantized models on the same prompt. We pair ByteShape IQ4_XS-3.93bpw (KLD 0.069) against each of five near-baseline quantizations on LiveCodeBench and ask:

Table 4: Per-prompt KLD routing on LiveCodeBench. Each model is paired with ByteShape IQ4_XS-3.93bpw, and evaluation is restricted to prompts where exactly one model passes. Routing succeeds when the model with lower per-prompt KLD is the passer.

Pairing vs. ByteShape IQ4_XS-3.93bpw	disagrees	wins	$\rho(p)$
AesSedai-Q5_K_M	190	47.4%	+0.033 (0.53)
Unsloth-UD-Q5_K_XL	177	44.1%	+0.109 (0.040)
mudler-Balanced	170	42.9%	+0.126 (0.020)
Bartowski-Q4_K_L	163	42.3%	+0.128 (0.020)
Unsloth-UD-IQ4_XS	180	49.4%	-0.009 (0.86)
BF16-fails: AesSedai-Q5_K_M vs. ByteShape IQ4_XS-3.93bpw	71	40.8%	+0.159 (0.059)

Table 5: Per-prompt KLD routing across four Qwen benchmarks. Routing accuracy is the fraction of disagreement prompts on which the lower-KLD model is better. Oracle headroom and routed change are score differences relative to the better static model.

Benchmark	Lower-KLD wins	Oracle headroom	Routed change
LiveCodeBench	45.3%	+7.55 pp	-1.57 pp
HumanEval	40.0%	+1.59 pp	-0.98 pp
GSM8K	46.0%	+0.85 pp	-0.17 pp
IFEval	49.1%	+4.62 pp	-0.44 pp

on prompts where exactly one of the two passes, does routing to the lower per-prompt KLD model identify the passer more than half the time?

Table 4 shows that the answer is no. Lower-KLD-wins accuracy is in [42.3%, 49.4%] across the five pairings, never reaching 50%. The signed correlation $\rho(\text{KLD}, \text{correctness})$ is positive (wrong-direction) in four and essentially zero ($\rho = -0.01$) in the fifth. On the BF16-fail slice (the prompts where a quantized model could leapfrog the reference), accuracy drops further to 40.8%.

Here, ρ is the Spearman correlation over both models’ stacked (KLD, pass) observations on the disagreement prompts: a negative value favors lower-KLD routing, while a positive value points in the opposite direction.

A broader routing experiment applies the same five model pairings across four prompt-level benchmarks. For each benchmark, we pool disagreement prompts across pairings to compute lower-KLD routing accuracy and compare the routed result against the better static model and an oracle that selects the correct model whenever the pair disagrees.

Table 5 shows that lower-KLD routing does not outperform the better static model on any of the four benchmarks. This is not simply because the models rarely differ in useful ways: the oracle has positive headroom on every task, reaching +7.55 percentage points on LiveCodeBench and +4.62 on IFEval. KLD fails to identify the direction of these disagreements and instead reduces score on every benchmark. Per-prompt KLD therefore con-

Table 6: $\rho(\text{KLD}, \cdot)$ for volume and direction f in the KLD-defined silent zone and full cohort. $*p < 0.05$, $^\dagger p < 0.10$.

Benchmark	Qwen3.6-35B-A3B				Devstral-Small-2-24B			
	silent		full		silent		full	
	vol	f	vol	f	vol	f	vol	f
HumanEval	+0.51*	+0.02	+0.76*	-0.31	+0.73*	+0.35	+0.91*	-0.17
LiveCodeBench	+0.50*	+0.22	+0.82*	-0.59*	—	—	—	—
IFEval	+0.42 [†]	+0.17	+0.67*	-0.19	+0.20	-0.15	+0.81*	-0.71*
BFCL_v3	+0.91*	-0.14	+0.96*	+0.44*	+0.51*	+0.17	+0.84*	-0.57*
GSM8K	+0.38	+0.04	+0.59*	-0.48*	+0.58*	-0.04	+0.92*	-0.82*
GSM8K-v	+0.62*	-0.35	+0.88*	-0.81*	+0.55*	-0.31	+0.90*	-0.77*
GSM8K-v think	+0.66*	-0.52*	+0.84*	-0.83*	—	—	—	—
MATH-500	—	—	—	—	+0.19	+0.38	+0.78*	-0.80*
MMLU	+0.66*	-0.27	+0.72*	+0.37 [†]	+0.84*	-0.67*	+0.98*	-0.90*
Composite	+0.94*	-0.14	+0.97*	-0.70*	+0.55*	-0.41	+0.92*	-0.86*

tains a weak failure signal, but not a dependable model-selection signal.

5 Disagreement Volume and Direction

What does KLD actually measure, if not quality? We propose a decomposition that explains it.

For a benchmark of N prompts where the reference (BF16) passes N_p and a candidate quant passes N_q , define: (1) **leapfrogs** ℓ : prompts where the quant passes and the reference fails; (2) **drops** d : prompts where the quant fails and the reference passes; and (3) **volume** $\text{vol} = \ell + d$: the total disagreement count. For $\text{vol} > 0$, define **direction** $f = \ell/\text{vol}$ as the fraction of disagreements that are improvements; when $\text{vol} = 0$, set $f = 1/2$ by convention. Since $N_q = N_p + \ell - d$, we obtain:

$$\text{score}_q = \text{score}_{\text{ref}} + \frac{\text{vol} \cdot (2f - 1)}{N} \quad (1)$$

where score_q and $\text{score}_{\text{ref}}$ are benchmark scores of the quant and BF16. Score is determined by *how many* prompts the quant flips (volume) and *in which direction* ($2f - 1$). That is, volume is the number of flips, while direction captures their balance between leapfrogs and drops. The identity is exact for benchmarks whose official score is one binary outcome per prompt (HumanEval, LCB, IFEval, GSM8K, and both GSM8K-v modes), and approximate for BFCL_v3, MMLU, and MATH-500, whose official scores macro-average over sub-categories with a weighting different from the per-prompt count we use for leap/drop accounting.

KLD measures displacement, not direction:

Table 6 and Figure 3 show this decomposition on both cohorts. In both Qwen and Devstral, KLD consistently tracks disagreement volume: in the silent zone, $\rho(\text{KLD}, \text{volume})$ is positive on almost every benchmark, with Qwen spanning $\rho \in [+0.38, +0.91]$ and a composite $\rho = +0.94$ ($p < 0.001$). On Devstral, the per-benchmark range is wider, with two small-magnitude cases

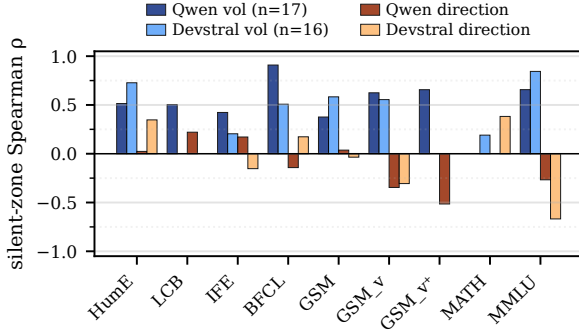


Figure 3: KLD-defined silent-zone Spearman $\rho(\text{KLD}, \cdot)$ per benchmark for volume (blue) and direction (orange).

on IFEval and MATH-500, and the composite correlation is weaker at $+0.55$ ($p = 0.03$). Direction behaves differently. Its correlation with KLD is weak and task-dependent in the silent zone, but becomes strongly negative over the full cohort once lossy-zone models enter the comparison. These results explain the full-cohort KLD–quality correlation: degraded models both move farther from the reference and tend to move in a harmful direction. Once we control for volume, the residual KLD signal on composite score does not reach significance in either cohort: partial $\rho = -0.35$ ($p = 0.17$) for Qwen and $\rho = -0.20$ ($p = 0.46$) for Devstral¹.

Why task matters: The task-dependent direction signal may reflect differences in how tasks tolerate token-level deviations. Math reasoning often depends on locally determined token choices, where small deviations can compound through the solution. Code correctness is more structural: many programs with different token sequences can implement the same specification, so moving away from the BF16 token distribution does not necessarily imply failure. On code-specialized models, where the BF16 distribution is more task-aligned, drift can become more consistently harmful.

In short, KLD measures how far the quantized model moves from the reference. Score depends on whether that movement goes in a useful direction.

6 Practical Implications

For deployment, our results suggest separating *screening* from *ranking*. Fidelity metrics remain useful for filtering clearly damaged candidates in the lossy zone. Among near-baseline models, lower KLD/PPL does not reliably mean better downstream quality. Ranking should rely on downstream evidence, together with practical constraints such

¹We regress raw KLD and composite quality separately on disagreement volume, then compute Spearman correlation between the residuals.

as throughput, memory use, and hardware fit.

That evaluation may not require running every candidate on the full benchmark suite. We evaluate two preliminary ways to reduce its cost:

1. **Canary evaluation:** select the prompts that most divide previously evaluated models, using cross-model correctness variance as the informativeness score, then evaluate new candidates only on that smaller set. In leave-one-publisher-out evaluation on Qwen, a 308-prompt canary, 10% of our four-benchmark suite, reaches $\rho \approx 0.52$ with the complete four-benchmark ranking and identifies the best silent-zone candidate in 76% of trials.
2. **Estimate volume and measure direction:** use WikiText KLD to predict disagreement volume, then measure direction by evaluating prompts that the BF16 reference fails. In leave-one-publisher-out evaluation on Qwen, measuring leapfrogs on 214 reference-failure prompts, approximately 7% of the four-benchmark suite, reaches $\rho \approx 0.70$ with the full ranking and identifies the best silent-zone candidate in 91% of trials (1,000 trials; 95% CI [89%, 93%]).

The two approaches use evaluation budget differently. The selected canary measures downstream quality directly on prompts identified from previous model outcomes. The second approach uses KLD for the component it predicts well, disagreement volume, and spends downstream evaluations on the missing direction signal. These results are preliminary: both experiments cover one model family, four benchmarks, and a family-specific leave-one-publisher-out setting (Appendix F).

7 Conclusion

We evaluated KLD and PPL as proxies for downstream quality of quantized LLMs. Across two cohorts, KLD appears predictive over the full quality range, but this signal is driven mainly by degraded low-bpw variants. In the near-baseline silent zone, we find no dependable KLD ranking signal across model families and metric variants, and per-prompt KLD does not rescue model selection. The volume–direction decomposition explains why: KLD tracks how often a quantized model disagrees with the reference, but benchmark quality depends on whether those disagreements help or hurt. KLD measures displacement, not direction. For deployment, fidelity metrics should filter damaged candidates, not rank near-baseline ones.

8 Limitations

Our results should be read as a study of fidelity metrics for quantized-model selection, not as a claim that KLD, PPL, and other fidelity metrics are useless. The failure mode we identify is specifically their inability to provide a dependable ranking signal among near-baseline candidates.

The silent-zone thresholds are descriptive and cohort-specific: they depend on the reference model, quantization methods, benchmark suite, and KLD setup. We therefore do not propose $\tau = 0.064$ or $\tau = 0.027$ as general decision rules. The near-baseline subsets are also relatively small ($n \approx 14\text{--}17$), and the wide confidence intervals in Appendix E should be interpreted accordingly. Our main cohort analysis covers two model families, per-prompt routing experiments are currently limited to Qwen, and the non-GGUF check includes only four models. We did not test whether the same failure mode extends to other forms of model modification, such as pruning or distillation.

Our reduced-evaluation experiments are limited to our Qwen family silent-zone candidates and four benchmarks with prompt-level outcomes. Both canary selection and KLD-to-volume calibration use previously evaluated models from the same family, so their performance may not transfer directly to a new model family or deployment distribution.

References

- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. [Evaluating large language models trained on code](#). *Preprint*, arXiv:2107.03374.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- Björn Deiseroth, Max Meuer, Nikolas Gritsch, Constantin Eichenberg, Patrick Schramowski, Matthias Aßenmacher, and Kristian Kersting. 2024. [Divergent token metrics: Measuring degradation to prune away LLM components – and optimize quantization](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6764–6783, Mexico City, Mexico. Association for Computational Linguistics.
- Abhinav Dutta, Sanjeev Krishnan, Nipun Kwatra, and Ramachandran Ramjee. 2024. [Accuracy is not all you need](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Elias Frantar, Saleh Ashkboos, Torsten Hoefer, and Dan Alistarh. 2023. [OPTQ: Accurate post-training quantization for generative pre-trained transformers](#). In *The Eleventh International Conference on Learning Representations*.
- Georgi Gerganov. 2023. [llama.cpp](#).
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. [LiveCodeBench: Holistic and contamination free evaluation of large language models for code](#). In *The Thirteenth International Conference on Learning Representations*.
- Ajay Kumar Jaiswal, Zhe Gan, Xianzhi Du, Bowen Zhang, Zhangyang Wang, and Yinfei Yang. 2024. [Compressing LLMs: The truth is rarely pure and never simple](#). In *The Twelfth International Conference on Learning Representations*.
- Eldar Kurtic, Alexandre Noll Marques, Shubhra Pandit, Mark Kurtz, and Dan Alistarh. 2025. [“Give me BF16 or give me death”? Accuracy-performance trade-offs in LLM quantization](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26872–26886, Vienna, Austria. Association for Computational Linguistics.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with PagedAttention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP ’23*, pages 611–626, New York, NY, USA. Association for Computing Machinery.
- Jemin Lee, Sihyeong Park, Jinse Kwon, Jihun Oh, and Yongin Kwon. 2025. [Exploring the trade-offs: Quantization methods, task difficulty, and model size in large language models from edge to giant](#). In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 8113–8121. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe.

2024. [Let's verify step by step](#). In *International Conference on Learning Representations*.
- Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. [AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration](#). In *Proceedings of Machine Learning and Systems*, volume 6, pages 87–100.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. [Pointer sentinel mixture models](#). In *International Conference on Learning Representations*.
- Mistral AI. 2025. [Devstral-Small-2-24B-Instruct-2512](#).
- Miloš Nikolić, Ghouthi Boukli Hacene, Ciaran Bannan, Alberto Delmas Lascorz, Matthieu Courbariaux, Omar Mohamed Awad, Isak Edo Vivancos, Yoshua Bengio, Vincent Gripon, and Andreas Moshovos. 2024. [BitPruning: Learning bitlengths for aggressive and accurate quantization](#). In *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5.
- Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. 2025. [The Berkeley function calling leaderboard \(BFCL\): From tool use to agentic evaluation of large language models](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 48371–48392. PMLR.
- Qwen. 2026. [Qwen3.6-35B-A3B: Agentic coding power, now open to all](#).
- ModelScope Team. 2024. [EvalScope: Evaluation framework for large models](#).
- Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. [SmoothQuant: Accurate and efficient post-training quantization for large language models](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 38087–38099. PMLR.
- Fan Yuan, Yuchen Yan, Yifan Jiang, Haoran Zhao, Tao Feng, Jinyan Chen, Yanwei Lou, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yuet-ing Zhuang. 2025. [GSM8K-V: Can vision language models solve grade school math word problems in visual contexts](#). *Preprint*, arXiv:2509.25160.
- Zhen Zheng, Xiaonan Song, and Chuanjie Liu. 2026. [MixLLM: LLM quantization with global mixed-precision between output-features and highly-efficient system design](#). In *Ninth Conference on Machine Learning and Systems*.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. [Instruction-following evaluation for large language models](#). *Preprint*, arXiv:2311.07911.

A Qwen MMLU Format Artifact

We exclude Qwen MMLU from the Qwen composite quality score because of a prompt-format artifact rather than a lack of knowledge. MMLU is administered with a five-shot prompt that instructs the model to output only the final answer choice. Qwen3.6 often begins reasoning instead of following this answer-only format, causing the benchmark’s answer extraction to fail and the response to be scored as incorrect. Manual inspection suggests that many such responses nevertheless contain the correct reasoning.

This behavior is already present in the BF16 reference and persists across the quantized cohort, so the resulting low MMLU scores cannot be interpreted as quantization-induced degradation. We therefore report Qwen MMLU separately (Table 2) and omit it from the Qwen composite, while retaining MMLU in the Devstral composite, where this behavior does not occur.

B Top- k and Response-Only KLD Measurement

This appendix specifies the response-only top- k KLD measurements used for KLD val-response in Table 3 and for the per-prompt analyses in Section 4.4.

B.1 Full-vocabulary KLD

At a single position, an autoregressive model defines a next-token distribution over the vocabulary V . Let $p \equiv p_{\text{BF16}}$ be the distribution of the BF16 reference and $q \equiv p_{\text{quant}}$ that of the candidate quant for the same context (the notation of Section 3). The Kullback–Leibler divergence of q from p is

$$\text{KL}(p \parallel q) = \sum_{i \in V} p(i) \log \frac{p(i)}{q(i)}, \quad (2)$$

which is 0 iff $q \equiv p$ and grows as q relocates probability mass. The expectation is taken under the reference p , so the metric penalizes the quantized model precisely where the reference places mass. Our headline metric is the mean of (2) over all scored positions. All logarithms are natural; full-vocabulary KLD and the top- k score are therefore reported in nats per scored token.

B.2 Top- k KLD

Under truncated decoding, most vocabulary tokens have negligible or zero sampling probability, so

their contribution to full-vocabulary KLD may have little direct bearing on the generated output. We therefore also evaluate a top- k score that emphasizes the tokens each model ranks most highly. We retain and renormalize each model’s top- k probabilities as follows.

Let $\mathcal{S}_p$ and $\mathcal{S}_q$ be the index sets of the k highest-probability tokens under p and q respectively. Each side is renormalized over *its own* support,

$$\begin{aligned}\tilde{p}(i) &= \frac{p(i)}{Z_p}, \quad i \in \mathcal{S}_p, \\ \tilde{q}(i) &= \frac{q(i)}{Z_q}, \quad i \in \mathcal{S}_q,\end{aligned}\tag{3}$$

with $Z_p = \sum_{j \in \mathcal{S}_p} p(j)$ and $Z_q = \sum_{j \in \mathcal{S}_q} q(j)$; the normalizers $\log Z_p$ and $\log Z_q$ are computed in log-space for numerical stability. Each truncated distribution is zero outside its own support. The top- k divergence is summed over the *reference’s* support $\mathcal{S}_p$,

$$\text{KL}_k(p \parallel q) = \sum_{i \in \mathcal{S}_p} \tilde{p}(i) (\log \tilde{p}(i) - \log \tilde{q}(i)).\tag{4}$$

The subtlety is the term $\log \tilde{q}(i)$ when a reference-important token $i \in \mathcal{S}_p$ is absent from the quant’s own top- k set, $i \notin \mathcal{S}_q$: the quant has effectively abandoned a token the reference cares about. Since $\tilde{q}$ is renormalized over $\mathcal{S}_q$, it assigns zero mass to any such token, so $\log \tilde{q}(i) = -\infty$ and the corresponding term, and hence the entire divergence, would blow up to $+\infty$. To keep the metric finite, we replace this $-\infty$ with a fixed log-probability floor,

$$\log \tilde{q}(i) = \begin{cases} \log q(i) - \log Z_q, & i \in \mathcal{S}_q, \\ c_{\min}, & i \notin \mathcal{S}_q, \end{cases}\tag{5}$$

with floor $c_{\min} = -10$. When $\mathcal{S}_p \neq \mathcal{S}_q$, we use “top- k KLD” as shorthand for this floored asymmetric divergence score. Because the floored candidate values are not renormalized over a common support with the reference, it is not the ordinary KL divergence between two distributions on the same truncated vocabulary. We use the same floor for every model, cohort, and value of k .

The floor caps the contribution of a missing token at $\tilde{p}(i) (\log \tilde{p}(i) - c_{\min})$, so a single abandoned token incurs a large but bounded penalty rather than diverging. This design has two properties:

- **Exact reduction.** When the two top- k sets coincide, $\mathcal{S}_p = \mathcal{S}_q$, every term uses the renormalized

$\log \tilde{q}(i)$ and (4) reduces to the plain renormalized top- k KLD.

- **Mass-relocation sensitivity.** When the supports diverge, the floored terms expose the case where the quant moved mass *outside* the reference’s top- k , which a renormalized-only metric would hide.

In addition to focusing the measurement on high-probability tokens, the top- k restriction has a large practical benefit: the two-pass protocol of Section B.3 must cache the reference distribution at every scored position, and storing only k tokens instead of the full vocabulary shrinks this footprint by orders of magnitude. For Qwen3.6-35B-A3B, whose vocabulary is $|V| = 248,320$ tokens, a full next-token distribution stored as 16-bit quantized log-probs occupies ≈ 485 KB per position; the top-60 record (a 32-bit index and a 32-bit log-prob per token) takes only 480 bytes, a $\sim 1000\times$ reduction, rising to $\sim 3100\times$ at $k = 20$. Over a million scored response tokens this is the difference between roughly 500 GB and under half a gigabyte of cached reference logits.

The same per-position quantity (4) can be evaluated either over a plain-text corpus (e.g. WikiText) or, as below, over response tokens only. We report it at $k \in \{20, 60\}$, with $k = 60$ as the default stored support; the smaller report point reuses the prefix of the sorted top-60 list.

B.3 Response-only (SFT-style) KLD

For instruction-tuned evaluation we care about divergence on the model’s *answer*, not on the prompt it was merely conditioned on. Averaging KLD over every position dilutes the answer-region signal with prompt tokens the model never had to generate. The response-only protocol measures divergence **only on response-token predictions**, while keeping the full prompt in the conditioning context.

Each corpus sample is tokenized as a prompt followed by an on-policy BF16 continuation. The same fixed prompt–response sequence is used for the BF16 reference and every quantized candidate. Writing the concatenated sequence as

$$x = \underbrace{[x_1, \dots, x_m]}_{\text{prompt}}, \underbrace{[x_{m+1}, \dots, x_n]}_{\text{response}},\tag{6}$$

the sample is decoded as a single sequence, so the full prompt populates the KV-cache. However, divergence is evaluated only at positions whose next

token is a response token: with position t predicting x_{t+1} , this is the set $\{m, m+1, \dots, n-1\}$. The prompt positions $\{1, \dots, m-1\}$ contribute context but no divergence terms.

The metric is computed in two passes over the same corpus. *Pass 1* (BF16 reference) stores, at each scored response position, the reference’s top- k indices and their full-softmax log-probabilities, and optionally the full distribution, when the full-vocabulary KLD of (2) is also desired. *Pass 2* (quant) runs the candidate on the identical token sequences, recomputes its own top- k at each scored position, and evaluates $\text{KL}_k(p \parallel q)$ from (4) against the stored reference record. Aggregating over all response tokens of all samples,

$$\overline{\text{KL}_k} = \frac{1}{N_{\text{resp}}} \sum_s \sum_{t=m_s}^{n_s-1} \text{KL}_k(p_t^{(s)} \parallel q_t^{(s)}), \quad (7)$$

where s indexes samples, $p_t^{(s)}$ and $q_t^{(s)}$ are the reference and quantized-model next-token distributions at position t of sample s , and $N_{\text{resp}} = \sum_s (n_s - m_s)$ is the total response-token count. This response-only corpus mean is reported as KLD val-response in Table 3.

Per-prompt benchmark protocol. The per-prompt analyses use the same response-only top- k score with $k = 60$, but retain one mean per benchmark prompt rather than aggregating across the corpus. For each prompt, the BF16 model’s generated continuation is teacher-forced through BF16 and every quantization, so all candidates are scored on identical response tokens. The final end-of-turn token is included for IFEval and HumanEval.

Routing aggregation. For each model pairing, disagreement prompts are those on which exactly one model passes, and the router selects the model with lower per-prompt KLD. Exact KLD ties are excluded, although none occur. The reported lower-KLD win rate pools disagreement events across the five pairings, whereas oracle headroom and routed change are equal-weight means of pair-level score differences relative to each pair’s better static model. The reported ρ stacks both models’ (KLD, pass) observations over the disagreement set before computing Spearman correlation.

C Quality-Defined Near-Baseline Analysis

We define near-baseline quantizations by their retained downstream quality, report the 99% quality-

Table 7: KLD–quality correlation among quantizations retaining at least 99% of measured BF16 composite quality. Significance: $*p < 0.05$.

Quantity	Qwen	Devstral
Full-cohort ρ	−0.72*	−0.86*
<i>Near-baseline ($\geq 99\%$ of BF16)</i>		
Near-baseline quantizations n	14	16
Maximum KLD in subset	0.069	0.047
Silent-zone ρ	+0.11	+0.02
Silent-zone p	0.71	0.93

retention result used in the main text, and examine sensitivity to the quality threshold and fidelity metric.

C.1 Definition and quality threshold

Let S_q denote the composite score of quantized model q , and let S_{BF16} denote the measured composite score of the corresponding BF16 reference. For a maximum relative quality drop δ , expressed as a fraction, we define q as near-baseline if:

$$S_q \geq (1 - \delta)S_{\text{BF16}}. \quad (8)$$

Our primary quality threshold is $\delta = 0.01$, corresponding to retention of at least 99% of the BF16 composite score. This gives 14 Qwen and 16 Devstral near-baseline quantizations (Table 7).

C.2 Sensitivity to the quality bar

Table 8 varies the maximum relative quality drop δ from 0.5% to 3.0%. The KLD–quality correlation remains non-significant through $\delta = 2.0\%$ on Devstral and through $\delta = 2.5\%$ on Qwen. It becomes significant at $\delta = 2.5\%$ on Devstral ($\rho = -0.47$, $p = 0.02$) and $\delta = 3.0\%$ on Qwen ($\rho = -0.55$, $p < 0.01$), as progressively degraded quantizations enter the comparison set.

C.3 Metric-invariance is preserved

Repeating the analysis under the quality-defined silent zone across the measurement variants in Table 3 gives the same qualitative result (Table 9). KLD aggregations, PPL formulations, top-1 agreement, long-context variants, and response-only variants all remain non-significant (Qwen $|\rho| \leq 0.20$; Devstral $|\rho| \leq 0.04$). The result therefore holds across metric formulations and under both silent-zone definitions.

Table 8: Sensitivity of the quality-defined silent zone to the quality-retention bar δ (maximum relative drop from BF16). n is the number of near-baseline quantizations. The correlation remains non-significant through $\delta = 2.5\%$ on Qwen and $\delta = 2.0\%$ on Devstral; the primary $\delta = 1\%$ threshold is bold.

δ (%)	Qwen			Devstral		
	n	ρ	p	n	ρ	p
0.5	9	-0.13	0.73	7	-0.64	0.12
1.0	14	+0.11	0.71	16	+0.02	0.93
1.5	18	-0.07	0.80	23	-0.30	0.16
2.0	18	-0.07	0.80	23	-0.30	0.16
2.5	20	-0.25	0.29	26	-0.47	0.02
3.0	24	-0.55	< 0.01	26	-0.47	0.02

Table 9: Metric-invariance under the quality-defined silent zone (composite $\geq 99\%$ of BF16). Significance: $*p < 0.05$.

	Qwen3.6-35B-A3B		Devstral-Small-2-24B	
	full	$\geq 99\%$	full	$\geq 99\%$
KLD mean wiki c512	-0.72*	+0.11	-0.86*	+0.02
KLD median wiki c512	-0.69*	+0.17	-0.86*	+0.02
KLD p90 wiki c512	-0.71*	+0.14	-0.86*	+0.02
KLD p99 (tail) wiki c512	-0.75*	+0.04	-0.86*	+0.01
mean Δ log PPL wiki c512	-0.68*	+0.09	-0.87*	+0.01
PPL ratio (q/base) wiki c512	-0.68*	+0.09	-0.87*	+0.01
PPL diff wiki c512	-0.68*	+0.09	-0.87*	+0.01
top-1 agreement % wiki c512	+0.70*	-0.20	+0.86*	-0.02
KLD wiki c8k (long context)	-0.72*	+0.11	-0.86*	+0.04
PPL ratio wiki c8k	-0.69*	+0.15	-0.86*	+0.00
top-1 agreement % wiki c8k	+0.69*	-0.20	+0.86*	-0.04
KLD val-wiki c512	-0.71*	+0.16	-0.86*	+0.02
KLD val-response top- $k=60$	-0.73*	+0.12	-0.86*	+0.04
KLD val-response top- $k=20$	-0.73*	+0.12	-0.87*	+0.04

D Internal Validation Dataset

The KLD val-wiki c512 and KLD val-response rows in Table 3 use the held-out validation slices summarized in Table 10. The val-wiki variant renders each sample as a single plain-text sequence and computes full-vocabulary mean KLD over all tokens at 512-token context. The val-response variants instead score only the tokens of an on-policy BF16 continuation, using the response-only top- k protocol defined in Appendix B. Together, these measurements test whether the silent-zone collapse persists under a prompt distribution closer to instruction-following use and under both full-sequence and response-only scoring.

For each reference model, we build a separate validation slice by stratified sampling from source datasets covering instruction following, multi-turn chat, code generation, math reasoning, tool calling, and multilingual prompts. The slices are not used for calibrating any quantized model in this study;

Table 10: Internal validation slices used for the KLD val-wiki c512 and KLD val-response measurements in Table 3.

Property	Qwen	Devstral
Samples (scored)	1,662	1,364
Median prompt length (tokens)	174	168
Tool-bearing samples (%)	9.3	20.8
Thinking-mode samples (%)	24.8	—

Table 11: Bootstrap 95% confidence intervals for headline Spearman correlations. Intervals are percentile bootstrap intervals over quantized models.

Analysis	Cohort	n	ρ	95% CI
Full cohort	Qwen	28	-0.72	[-0.89, -0.39]
	Devstral	41	-0.86	[-0.95, -0.68]
KLD-defined silent zone	Qwen	17	+0.00	[-0.53, +0.54]
	Devstral	16	-0.24	[-0.63, +0.33]
Lossy zone	Qwen	11	-0.80	[-1.00, -0.29]
	Devstral	25	-0.94	[-0.99, -0.81]
Quality-defined silent zone	Qwen	14	+0.11	[-0.48, +0.64]
	Devstral	16	+0.02	[-0.50, +0.59]
KLD-defined silent-zone volume	Qwen	17	+0.94	[+0.75, +0.99]
	Devstral	16	+0.55	[+0.02, +0.88]
KLD-defined silent-zone partial ρ controlling for volume	Qwen	17	-0.35	[-0.75, +0.33]
	Devstral	16	-0.20	[-0.60, +0.49]

they serve exclusively as measurement corpora for the validation-set KLD measurements against the corresponding BF16 reference.

E Confidence Intervals for Headline Correlations

To further estimate uncertainty for the headline Spearman correlations, we perform a nonparametric bootstrap over the quantized models. For each KLD-quality row, each bootstrap replicate samples n models with replacement from the corresponding set and recomputes $\rho(\text{KLD}, \text{composite})$. We used $B = 10,000$ bootstrap replicates with fixed seed 0 and report percentile 95% confidence intervals. We computed the intervals on the same data and zone definitions as the corresponding point estimates.

Composite volume pools raw disagreement counts across the seven benchmarks with prompt-level outcomes in each cohort, without normalization by benchmark size. Composite direction is the pooled leapfrog count divided by pooled volume. Devstral LiveCodeBench remains in the quality composite but is omitted from these quantities because only aggregate scores are available. For the volume rows, each bootstrap replicate recomputes the KLD-volume correlation. For the partial-correlation rows, we regress raw KLD and compos-

Table 12: Leave-one-publisher-out reduced-evaluation results on the Qwen KLD-defined silent-zone models. ρ is the mean Spearman correlation with the ranking obtained from all 3,079 prompts in the four-benchmark suite. Top-1 is the frequency with which the method selects the true best candidate. Regret and MAE are measured in percentage points.

Method	Prompts	ρ	Top-1	Regret	MAE
KLD alone	0	-0.48	0/200 (0%)	2.25	0.81
Uniform subsample	308	+0.36	55/200 (27.5%)	0.98	1.75
Reference-stratified	308	+0.45	61/200 (30.5%)	0.91	1.18
Selected canary	308	+0.52	151/200 (75.5%)	0.29	0.58
KLD volume + leapfrogs	214	+0.70	910/1000 (91.0%)	0.10	0.83

ite quality separately on volume, with intercepts, and compute the Spearman correlation between the residuals. Both ordinary-least-squares regressions are refit within every replicate.

Because the silent-zone subsets are small, these intervals should be interpreted as uncertainty estimates rather than evidence that the true correlation is exactly zero. The purpose of this analysis is to test whether the strong full-cohort KLD–quality relationship is also supported inside the near-baseline region.

Table 11 supports the same interpretation as our main results. The full-cohort and lossy-zone KLD–quality intervals exclude zero on both cohorts. In contrast, the KLD–quality intervals for both the KLD-defined silent zone and the quality-defined silent zone are wide and straddle zero. Thus, the silent-zone result should be read as an absence of detectable KLD–quality ranking signal in the near-baseline region, not as a precise estimate that the true correlation is exactly zero. The volume correlations remain positive, especially on Qwen, while the partial correlations controlling for volume also straddle zero. This is consistent with the volume–direction account: KLD primarily tracks displacement volume, but its residual ranking signal for downstream quality is not detectable among near-baseline candidates.

F Reduced-Evaluation Model Selection

We evaluate two reduced-evaluation strategies on Qwen using four benchmarks with prompt-level outcomes: LiveCodeBench (1,055 prompts), HumanEval (164), GSM8K (1,319), and IFEval (541), for 3,079 prompts in total. Ranking and model-selection results are computed over the KLD-defined Qwen silent-zone ($\tau = 0.064$) models. BF16 supplies the reference outcomes but is not a candidate.

We use five leave-one-publisher-out folds: AesSedai (24 training and 4 held-out models, of which 3 are silent-zone models), Bartowski (24/4, with 4 silent), ByteShape (18/10, with 2 silent), mudler (24/4, with 3 silent), and Unsloth (22/6, with 5 silent). Predictions for the held-out silent-zone models are pooled across folds, and each trial’s Spearman correlation is computed over all 17 pooled predictions. Prompt selection and model fitting use training-publisher outcomes only.

We compare against KLD alone and two sampling baselines. For each benchmark, KLD alone fits a decreasing isotonic mapping from log Wiki-Text KLD to benchmark score using all training-publisher models in the fold, then predicts the held-out models from their KLD. Uniform subsampling draws prompts without replacement within each benchmark. Reference-stratified subsampling allocates each benchmark’s budget between prompts the BF16 reference passes and fails in proportion to the two strata’s sizes, then weights the stratum means by their full-benchmark proportions. The two sampling baselines and selected-canary method use the same 308-prompt budget and are evaluated over 200 deterministically seeded trials.

F.1 Selected canary

For each fold and benchmark, the selected-canary method ranks prompts by $\hat{p}(1 - \hat{p})$, where $\hat{p}$ is the pass rate among training-publisher models. It selects the highest-variance prompts. Across trials, randomness breaks ties among prompts with equal variance. Every held-out model in a fold receives the same selected prompts. A per-benchmark linear calibration from canary accuracy to full-benchmark score is fit on the training models, and a candidate’s predicted composite is the equal-weight mean of its four calibrated benchmark estimates. Prompt selection and calibration never use outcomes from the held-out publisher.

Per trial, predictions from all folds are pooled over the 17 silent-zone candidates. We compute Spearman ρ between predicted and complete four-benchmark composites. Top-1 records whether the predicted best is the true best, regret records the percentage-point composite gap between the true best and the selected candidate, and MAE is the mean absolute composite prediction error over the 17 candidates. Table 12 reports the mean of these quantities across trials.

Table 13: Qwen3.6-35B-A3B cohort ($n = 28$, $\tau = 0.064$), sorted by KLD.

Publisher	Model	bpw	KLD	comp	zone
Unsloth	UD-Q5_K_XL	6.14	0.0082	0.724	silent
AesSedai	Q5_K_M	6.06	0.0083	0.718	silent
mudler	Balanced	5.91	0.0106	0.717	silent
Bartowski	Q5_K_L	5.84	0.0110	0.720	silent
Unsloth	UD-Q4_K_XL	5.16	0.0135	0.728	silent
AesSedai	Q4_K_M	5.11	0.0140	0.719	silent
mudler	Quality	5.26	0.0140	0.727	silent
Bartowski	Q4_K_L	5.02	0.0196	0.717	silent
Bartowski	Q4_K_M	4.93	0.0208	0.722	silent
Bartowski	IQ4_XS	4.34	0.0285	0.725	silent
Unsloth	UD-IQ4_NL	4.16	0.0334	0.726	silent
AesSedai	IQ4_XS	4.06	0.0335	0.717	silent
Unsloth	UD-IQ4_XS	4.09	0.0336	0.724	silent
Unsloth	UD-Q3_K_XL	3.89	0.0434	0.715	silent
mudler	Compact	3.99	0.0451	0.707	silent
ByteShape	IQ4_XS-4.15bpw	4.15	0.0504	0.731	silent
ByteShape	Q4_K_S-4.22bpw	4.22	0.0582	0.723	silent
ByteShape	IQ4_XS-3.93bpw	3.93	0.0689	0.721	lossy
Unsloth	UD-IQ3_S	3.15	0.0842	0.708	lossy
AesSedai	IQ3_S	3.13	0.0850	0.705	lossy
ByteShape	Q4_K_S-3.80bpw	3.80	0.0944	0.714	lossy
ByteShape	Q3_K_S-3.39bpw	3.39	0.0985	0.705	lossy
mudler	Mini	3.30	0.0994	0.695	lossy
ByteShape	IQ3_S-3.48bpw	3.48	0.1223	0.706	lossy
ByteShape	IQ3_S-3.00bpw	3.00	0.1273	0.706	lossy
ByteShape	Q3_K_S-2.71bpw	2.71	0.1917	0.685	lossy
ByteShape	Q3_K_S-2.69bpw	2.69	0.2092	0.674	lossy
ByteShape	IQ2_S-2.17bpw	2.17	0.2733	0.654	lossy

F.2 KLD volume with sampled leapfrogs

For benchmark b , we fit an isotonic mapping from WikiText log-KLD to normalized disagreement volume V_b/N_b using training-publisher models only. For each held-out candidate, we estimate its leapfrog rate $\hat{r}_{\text{leap},b}$ by uniformly sampling, without replacement, prompts that the BF16 reference fails. The estimated score is

$$\hat{S}_b = S_{\text{ref},b} + \frac{2N_{\text{fail},b}\hat{r}_{\text{leap},b} - \hat{V}_b}{N_b}. \quad (9)$$

At the 10% grid point, the nominal per-benchmark budgets are 106, 16, 132, and 54 prompts. Because HumanEval and GSM8K contain only 8 and 46 BF16 failures, respectively, the method evaluates $106 + 8 + 46 + 54 = 214$ prompts. Thus, 214 is a fixed count representing 6.95% of the 3,079-prompt suite. The same sampled prompts are applied to every held-out model within a fold.

Over 1,000 deterministically seeded trials, the KLD-volume-plus-leapfrog method gives mean $\rho = 0.699$; the trial-wise 2.5–97.5 percentile interval is $[0.45, 0.87]$. It selects the true best candidate in 910/1000 trials, with a Wilson 95% confidence interval of $[0.89, 0.93]$, and incurs mean regret of 0.10 percentage points. The selected canary is also more effective than uniform or reference-stratified

Table 14: Devstral-Small-2-24B cohort ($n = 41$, $\tau = 0.027$), sorted by KLD.

Publisher	Model	bpw	KLD	comp	zone
Unsloth	Q8_0	8.50	0.0008	0.696	silent
Unsloth	UD-Q8_K_XL	9.84	0.0009	0.700	silent
Unsloth	UD-Q6_K_XL	7.05	0.0019	0.694	silent
Bartowski	Q6_K_L	6.67	0.0025	0.695	silent
Unsloth	Q6_K	6.56	0.0027	0.711	silent
Bartowski	Q6_K	6.56	0.0027	0.697	silent
Bartowski	Q5_K_L	5.83	0.0069	0.699	silent
Unsloth	Q5_K_M	5.69	0.0070	0.690	silent
Bartowski	Q5_K_M	5.69	0.0071	0.692	silent
Unsloth	UD-Q5_K_XL	5.69	0.0072	0.692	silent
Unsloth	Q5_K_S	5.53	0.0084	0.695	silent
Bartowski	Q5_K_S	5.53	0.0086	0.695	silent
Unsloth	UD-Q4_K_XL	4.92	0.0200	0.695	silent
Bartowski	Q4_K_M	4.86	0.0219	0.694	silent
Unsloth	Q4_K_M	4.86	0.0222	0.696	silent
Unsloth	Q4_1	5.04	0.0266	0.697	silent
Unsloth	Q4_K_S	4.60	0.0274	0.691	lossy
Bartowski	IQ4_NL	4.57	0.0278	0.685	lossy
Unsloth	IQ4_NL	4.57	0.0279	0.696	lossy
Bartowski	IQ4_XS	4.33	0.0284	0.691	lossy
Unsloth	IQ4_XS	4.33	0.0286	0.697	lossy
Unsloth	Q4_0	4.58	0.0387	0.690	lossy
ByteShape	IQ4_XS-4.04bpw	4.04	0.0473	0.698	lossy
Unsloth	UD-Q3_K_XL	4.02	0.0600	0.691	lossy
Unsloth	Q3_K_M	3.89	0.0684	0.683	lossy
ByteShape	IQ3_S-3.47bpw	3.47	0.1181	0.685	lossy
Unsloth	Q3_K_S	3.53	0.1260	0.663	lossy
ByteShape	IQ3_S-3.19bpw	3.19	0.2062	0.672	lossy
Unsloth	UD-Q2_K_XL	3.15	0.2579	0.636	lossy
Unsloth	UD-IQ3_XXS	3.19	0.2592	0.655	lossy
ByteShape	IQ3_S-2.96bpw	2.96	0.2733	0.658	lossy
Unsloth	Q2_K_L	3.07	0.2913	0.635	lossy
Unsloth	Q2_K	3.01	0.2936	0.631	lossy
ByteShape	IQ3_S-2.78bpw	2.78	0.3617	0.639	lossy
ByteShape	IQ3_S-2.67bpw	2.67	0.4425	0.620	lossy
Unsloth	UD-IQ2_M	2.79	0.5013	0.613	lossy
ByteShape	IQ2_S-2.43bpw	2.43	0.6525	0.572	lossy
ByteShape	IQ2_S-2.34bpw	2.34	0.7991	0.536	lossy
Unsloth	UD-IQ2_XXS	2.29	1.2795	0.348	lossy
Unsloth	UD-IQ1_M	2.04	2.7823	0.128	lossy
Unsloth	UD-IQ1_S	1.88	3.5199	0.089	lossy

subsampling, but the KLD-volume-plus-leapfrog method achieves the strongest ranking and top-1 results with fewer evaluated prompts.

These results demonstrate family-specific evaluation reduction rather than a transferable recipe. The canary’s informativeness criterion and the KLD-to-volume calibration are both learned from previously evaluated models and may not transfer unchanged to another model family, benchmark collection, or deployment distribution.

G GGUF Model List

Tables 13 and 14 list all public GGUF quantizations in the two cohorts, sorted by KLD.

H Non-GGUF Model List

Table 15 lists the four public non-GGUF quantizations included in this check, sorted by KLD.

Table 15: Qwen3.6-35B-A3B 4-bit non-GGUF quantizations served with vLLM, sorted by KLD.

Publisher	Model	bpw	KLD	comp	zone
cyankiwi	Qwen3.6-35B-A3B-AWQ-4bit	5.43	0.0185	0.7195	silent
btbtyler09	Qwen3.6-35B-A3B-GPTQ-4bit	5.47	0.0215	0.7181	silent
FenomAI	Qwen3.6-35B-A3B-AWQ-4bit	5.31	0.0217	0.7235	silent
QuantTrio	Qwen3.6-35B-A3B-AWQ	5.28	0.0250	0.7229	silent

The KLD measurements use the same Wiki-Text corpus, 512-token context, chunking, and mean aggregation as the GGUF analysis, but use a vLLM-served BF16 reference (Kwon et al., 2023). The measured cross-stack BF16-to-BF16 floor is 0.00547, so the KLD values are directly comparable within this table; the silent-zone labels relative to the GGUF threshold $\tau = 0.064$ should be interpreted descriptively. All four models also retain at least 99% of measured BF16 composite quality.