ACCURACY IS NOT ENOUGH: MEASURING INSTRUCTION-FOLLOWING STABILITY UNDER POST-TRAINING QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training quantization is routinely evaluated by task accuracy alone, leaving open whether the instruction-following behavior instilled during instruction tuning survives aggressive compression. We study this question for INT4 weight-only quantization (the mixed-precision q4_K_M format) of two compact instruction-tuned models, Llama 3 8B Instruct and Phi-3 Mini 3.8B Instruct, across seven prompt structures and three decoding temperatures on GSM8K and the MMLU high-school mathematics subset (12,600 generations). We score every generation both for task accuracy and for strict compliance with an explicit output contract, and we show that lenient, extraction-based parse rate is saturated above 0.99 and cannot detect the effects we report. Three findings emerge. First, task correctness and instruction-level compliance decouple under INT4, and the compliance effect is the larger one: Phi-3 contract compliance falls from 0.107 to 0.055 while its accuracy simultaneously rises from 0.344 to 0.373, so a purely accuracy-based evaluation would record this model as unaffected or improved. Second, the effect is strongly model-dependent: Llama 3 compliance is stable and slightly higher under INT4, and the two models also re-rank prompt effectiveness differently on accuracy. Third, the loss is broad rather than confined to one prompt style: Phi-3 compliance falls under every prompt family, with the largest relative losses in three- and five-shot (-76%, -61%), so no prompting strategy can be assumed to stabilize behavior under compression. We discuss a distributional hypothesis that is consistent with these behaviors but that we do not test directly, and we argue that evaluation of quantized models should treat instruction-following stability as a first-class metric alongside accuracy.

1 INTRODUCTION AND RELATED WORK

Large Language Models (LLMs) have demonstrated strong performance across reasoning, summarization, and code generation, largely enabled by scaling model size and training data. However, this scaling introduces substantial inference cost: LLaMA-70B alone requires roughly 140 GB in FP16, making deployment impractical without multi-GPU infrastructure (Wu et al., 2023; Husom et al., 2025). These constraints motivate compression techniques such as quantization, particularly for edge or resource-limited environments. Yet compression may also perturb the behavioral regularities instilled during instruction tuning, and standard quantization evaluation, which reports task accuracy, is not designed to detect this.

1.1 QUANTIZATION FOR EFFICIENT DEPLOYMENT

Post-training quantization (PTQ) reduces memory and compute by mapping floating-point weights and activations to low-bit representations (Yao et al., 2023; Xiao et al., 2023; Rakka et al., 2025; Wu et al., 2023; Hasan, 2024). INT4 enables up to $4\times$ compression at the cost of increased sensitivity to rounding and outliers (Kurtic et al., 2024; Gong et al., 2024; Li et al., 2026). Practical low-bit inference is supported by techniques such as GPTQ, AWQ, and SmoothQuant (Frantar et al., 2022; Lin et al., 2024; Xiao et al., 2023), by system frameworks such as FlexGen and ZeroQuant-FP (Sheng et al., 2023; Wu et al., 2023), and by improved quantizers such as block-wise optimal

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

float codebooks (Blumenberg et al., 2026) and rotation-based schemes for reasoning models (Liang et al., 2026). Prior work asks how to quantize with less error; we ask what fixed quantization does to instruction-following behavior. We therefore intentionally do not compare quantization algorithms, and instead fix a single 4-bit configuration, the *q4_K_M* GGUF format used by Ollama/llama.cpp, which a recent unified evaluation of llama.cpp formats identifies as a standard compression–quality trade-off point (Kurt, 2026).

Quantization is not behavior-neutral. Prior studies show that arithmetic and multi-step reasoning degrade more under INT4 than classification tasks (Li et al., 2026; Yao et al., 2023), that activation quantization often dominates performance loss (Gong et al., 2024; Husom et al., 2025), and that impacts vary by architecture and workload (Kurtic et al., 2024; Liu et al., 2025). Nor is the effect confined to accuracy. Hong et al. (2024) report that 4-bit quantization largely preserves the trustworthiness of a 13B model, while 3-bit GPTQ causes severe losses in toxicity and out-of-distribution robustness, which they link to degraded instruction following and examine further with MT-Bench, including failures to follow the system prompt and required output format; these risks are not apparent from benign MMLU performance alone. For a 7B source model they report losses above five points in three trust dimensions despite a 1.6-point MMLU drop. Quantized instruction-tuned models from 1B to 405B parameters generally beat smaller FP16 baselines except on instruction following and hallucination detection (Lee et al., 2025), and in-context learning, chain-of-thought reasoning, and instruction following remain largely preserved under 4-bit GPTQ on LLaMA models but degrade severely at 2 bits, as measured by benchmark accuracy and automatic win-rate judgments (Liu et al., 2024). Closest to our work, Li et al. (2026) localize *where* mathematical reasoning breaks under PTQ, finding that errors concentrate in method and execution steps and cascade from the first vulnerable step; and Kurt (2026) benchmark accuracy, instruction following, and throughput across llama.cpp formats for a single Llama-3-class model under fixed prompting. Instruction following is thus a known weak point of quantized models; what is new here is a strict-contract metric, a saturation argument against parse rate, and the prompt–precision–metric interaction. These studies hold the prompt fixed and ask how much capability is lost; we hold the quantizer fixed and ask whether the model still obeys its output contract when it still gets the answer right.

1.2 Prompting, Instruction Following, and Precision Sensitivity

Prompt engineering remains the primary mechanism for steering LLM behavior. Few-shot prompting enables in-context learning (Brown et al., 2020), chain-of-thought improves multi-step reasoning (Wei et al., 2022; Kojima et al., 2022), and instruction tuning with RLHF strengthens adherence and coherence (Ouyang et al., 2022). Instruction following is increasingly measured separately from reasoning, typically with verifiable-constraint benchmarks such as IFEval (Zhou et al., 2023): Huang et al. (2026) show that reasoning models can reason well yet fail to follow format constraints, and evaluate the two on distinct benchmarks. Most such evaluations assume full-precision inference.

Low-bit compression can reshape token distributions and attention dynamics, potentially disrupting prompt effectiveness and instruction adherence, particularly in compact deployment-oriented models such as LLaMA-3-8B and Phi-3-mini (Grattafiori et al., 2024; Abdin et al., 2024). Quantization research typically benchmarks accuracy, and where it measures instruction following it does so under a single fixed prompt (Frantar et al., 2022; Lin et al., 2024; Xiao et al., 2023; Kurt, 2026; Lee et al., 2025); prompting research rarely considers precision effects (Brown et al., 2020; Wei et al., 2022; Kojima et al., 2022; Ouyang et al., 2022), although model behavior is known to be highly sensitive to prompt formatting even at full precision (Sclar et al., 2024); and system optimizations treat prompt behavior as orthogonal to throughput (Kwon et al., 2023; Sheng et al., 2023). What remains open is whether compliance with an explicit output format, and the prompt structures meant to secure it, survive quantization.

1.3 Scope and Contributions

We evaluate LLaMA-3-8B and Phi-3-mini under FP16 and INT4 across seven prompt structures and three decoding temperatures on GSM8K and the high-school mathematics subset of MMLU (Cobbe et al., 2021; Hendrycks et al., 2020), totaling 12,600 inferences, scoring every generation for both accuracy and compliance with an explicit output contract.

Our contributions are threefold. (1) **Correctness–compliance decoupling.** We show that INT4 quantization can degrade adherence to an explicit output contract while leaving answer accuracy intact or even improving it: Phi-3 contract compliance falls by 48.5% relative ($0.107 \rightarrow 0.055$) while its accuracy rises ($0.344 \rightarrow 0.373$). The outputs remain machine-parseable throughout, so this failure class is invisible both to accuracy-only evaluation and to the lenient extraction that accuracy harnesses rely on. (2) **Prompt robustness is a joint property of model, precision, and prompt.** Quantization re-ranks prompt effectiveness, and the re-ranking is model-dependent: on Llama 3 every exemplar-based prompt improves under INT4 and every non-exemplar prompt degrades, whereas on Phi-3 most prompts improve and the richest exemplar prompt is the only one that degrades. A prompting strategy's robustness to quantization therefore cannot be assumed to transfer across precisions or architectures. (3) **A compliance-aware evaluation protocol.** We operationalize instruction-following stability as strict compliance with the output contract, show that lenient parse rate is saturated and cannot detect the effects we report, and validate the metric's sensitivity to its own definition (Appendix A.2). Throughout, we separate what this protocol observes from the distributional hypothesis that might explain it, which we state but do not test.

2 METHODOLOGY

This study examines how post-training quantization changes both (i) reasoning accuracy and (ii) instruction-following behavior in compact, instruction-tuned LLMs. We evaluate two open-weight models, Llama 3 8B Instruct and Phi-3 Mini 3.8B Instruct, selected as a widely deployed class of small instruction-tuned models with different design emphases: 8.0B parameters with an 8192-token context and 4096-dimensional embeddings, and 3.8B with a 4096-token context and 3072-dimensional embeddings, respectively. Each model is evaluated in FP16 and INT4 to isolate precision-induced behavioral differences without retraining or architectural modification. We use the *q4_K_M* scheme as implemented in Ollama/llama.cpp, written INT4 throughout: weight-only and block-wise, with activations unquantized. It is a mixed-precision k-quant rather than uniformly 4-bit, keeping selected tensors above 4 bits, so the memory saving over FP16 is closer to $3\times$ than $4\times$.

2.1 EXECUTION SETUP AND CONTROL VARIABLES

All runs are executed locally via Ollama on an Apple M4 Pro machine with 24 GB unified memory.[1]

A standardized system prompt is applied to all configurations to enforce an output contract: a single final line in the form `Answer:<value>`. We operationalize instruction-level behavioral compliance as adherence to an explicit output contract specified in the system prompt. Violations—including additional explanatory text, missing answer markers, or formatting deviations—are treated as compliance failures rather than stochastic noise, since such deviations introduce integration risk in pipeline-coupled or safety-relevant deployments.

2.2 PROMPTING AND DECODING FACTORS

We vary two user-controllable factors that commonly shape deployed behavior: prompt structure and decoding temperature. Seven prompt families are evaluated: zero-shot, zero-shot chain-of-thought, concise instruction, detailed instruction, and one-, three-, and five-shot demonstrations (full templates in Appendix A.1). The demonstrations are short synthetic items that show the output format, not in-domain problems, so we call these families exemplar-based (format-demonstration) prompts rather than tests of few-shot learning. Following the standard reading of in-context learning, we expected demonstrations to stabilize behavior under low precision by constraining the response distribution toward the shown template. As § 3.4 shows, that expectation holds for accuracy on one model and inverts for compliance on the other, which is itself a finding rather than an assumption carried through the analysis.

Decoding temperature is set to $\{0.0, 0.3, 0.7\}$ to capture deterministic, moderately stochastic, and highly stochastic regimes. Fixed decoding settings are `num_ctx`=2048, `num_batch`=8, and

[1]Model identifiers: *llama3:8b-instruct-fp16*, *llama3:8b-instruct-q4_K_M*, *phi3:3.8b-mini-4k-instruct-fp16*, *phi3:3.8b-mini-4k-instruct-q4_K_M*.

`stream`=False. The context limit is chosen to avoid truncation even under multi-shot prompts. We restrict decoding to temperature variation to avoid confounding stochastic controls.

2.3 Benchmarks and Scoring

We evaluate on GSM8K (Cobbe et al., 2021) and the high-school mathematics subset of MMLU (Hendrycks et al., 2020). GSM8K stresses multi-step arithmetic; MMLU tests symbolic recall in multiple-choice format. For MMLU, answer options (A–D) are appended to each question to maintain consistent formatting across prompts and models. Across 4 model variants $\times$ 7 prompt families $\times$ 3 temperatures, we evaluate 84 configurations. Each uses the first 100 items of the GSM8K test split and the first 50 of the MMLU high-school-mathematics test split, yielding 12,600 inferences. We compute exact-match accuracy via automated extraction. For GSM8K, the parser detects the answer line, extracts numerals/expressions, evaluates arithmetic when present, and applies a 10^{-2} tolerance. For MMLU, the final choice (A–D) is extracted using regular expressions with fuzzy fallback for free-text outputs.

To disentangle numerical correctness from instruction-following behavior, we additionally score every generation for **strict contract compliance**: a response is compliant if and only if it consists of exactly one non-empty line and that line begins with the marker `Answer:`. This enforces the required single-line format and rules out the appended rationale that the contract forbids. We adopt this as our primary adherence metric in place of *parse rate*, the fraction of responses from which any answer can be recovered. Extraction of this kind is standard in accuracy harnesses, and a high parse rate is easily read as adherence. Parse rate uses a lenient extractor and is saturated in our data, exceeding 0.99 in all 84 configurations (Table 3), so it cannot distinguish a bare answer from rationale containing a buried answer; strict compliance does. The same strict/lenient distinction appears in IFEval itself, whose loose metric deliberately strips prefaces and markup that the strict metric counts as violations (Zhou et al., 2023). Because such a metric is definition-sensitive, Appendix A.2 gives the formal definition and reports two looser variants as a robustness check. We also report median response length as a continuous signal independent of the compliance definition. Relative accuracy changes are signed so that positive values indicate degradation under INT4 (Figure 1(b), Tables 1–2), and are computed as the mean of the three temperature-specific values of $(A_{\text{FP16}} - A_{\text{INT4}})/A_{\text{FP16}}$, not as the relative change of the temperature-averaged accuracies; the latter estimator changes magnitudes but no directions (Appendix A.4). Intervals are 95% Wilson intervals computed from the underlying counts; per-cell counts and intervals are in Appendix A.7. Each FP16 generation is matched to the INT4 generation for the same item, prompt and temperature, so precision comparisons use paired McNemar tests. The six primary comparisons (accuracy, strict compliance and correct-but-non-compliant rate, $\times$ two models) are unadjusted for multiplicity; prompt-level tests are secondary and Holm-adjusted (Appendix A.7).

3 Results

We report four findings: (i) quantization's effect on accuracy varies by model and prompt rather than degrading it uniformly (§ 3.1); (ii) accuracy declines with decoding temperature, and the arithmetic–multiple-choice gap widens with it (§ 3.2); (iii) INT4 substantially reduces strict contract compliance for Phi-3 while leaving Llama 3 stable, even where accuracy does not fall (§ 3.3); and (iv) the effect depends jointly on model and prompt, and within Phi-3 is broad, largest in absolute terms under concise instruction and in relative terms under the three- and five-shot prompts that appear protective on accuracy (§ 3.4). Figure 1 summarizes all four.

3.1 Overall Accuracy and Quantization Effects

Pooled over prompts and temperatures, Llama 3 scores 0.199 (FP16) and 0.198 (INT4) and Phi-3 0.344 and 0.373 (Table 3), far below published results for these models. For reference, the Llama 3 8B Instruct model card and the Phi-3 technical report give 79.6% and 82.5% on GSM8K under 8-shot chain-of-thought evaluation (Meta AI, 2024; Abdin et al., 2024); our contract forbids exactly the rationale those settings supply. We attribute the gap to that design rather than to the harness, though a contract-free validation run is still pending (§ 4.3). Quantization does not uniformly degrade accuracy: depending on the prompt and model, INT4 can either reduce or improve perfor-

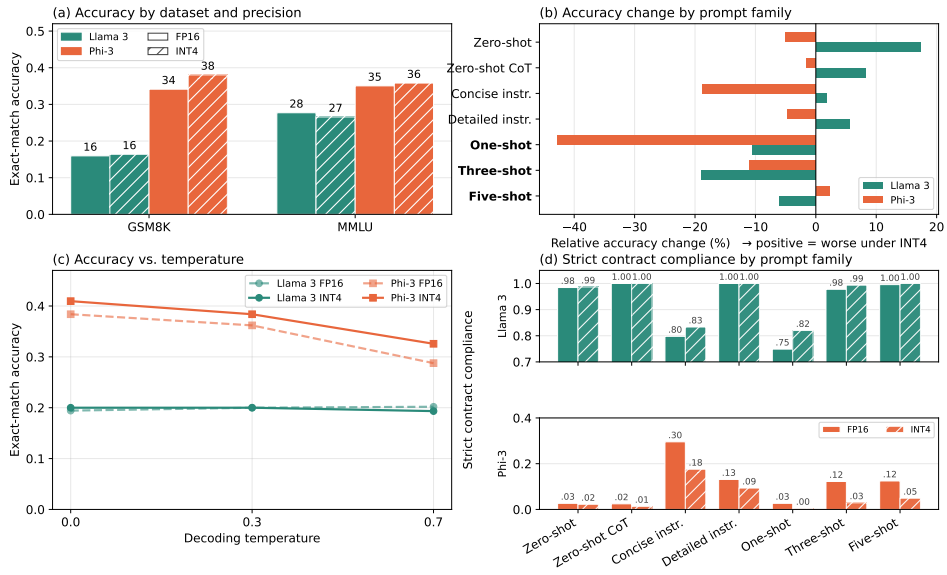

Figure 1: Accuracy and instruction-compliance effects across all 84 configurations. **(a)** Exact-match accuracy by dataset, model, and precision. **(b)** Relative FP16→INT4 accuracy change by prompt family; positive values indicate degradation under INT4. **(c)** Accuracy against decoding temperature; INT4 solid, FP16 dashed. **(d)** Strict contract compliance by prompt family, shown on separate scales for Llama 3 and Phi-3 so that differences at low compliance rates remain visible. Exemplar-based prompt families are shown in bold in (b), where Llama 3 *gains* the most accuracy; Phi-3 *loses* compliance under every family, most in absolute terms under concise instruction and most in relative terms under three- and five-shot. Every value is computed from the per-generation records and matches Tables 1–4.

mance relative to FP16. Table 1 and Table 2 report this per-prompt, averaged over temperatures, under the sign convention of § 2.3.

On Llama 3 (Table 1), the pattern is clean: all three exemplar-based prompts (one-, three-, and five-shot) improve under INT4, while all four non-exemplar prompts (zero-shot, zero-shot CoT, concise instruction, detailed instruction) degrade. This split is consistent with a structural-anchoring account: demonstrations may constrain the response distribution toward a shown template, reducing sensitivity to quantization-induced probability perturbations, whereas prompts without explicit exemplars offer no such anchor.

On Phi-3 (Table 2), the same account only partly holds. One- and three-shot again improve substantially under INT4, consistent with anchoring. However, several non-exemplar prompts improve as well—most notably concise instruction (18.74% relative gain) and zero-shot (4.98% relative gain)—and five-shot, the richest exemplar condition, is the only prompt that degrades (2.22% relative loss). Exemplar structure alone therefore does not explain Phi-3's pattern, which we read as a genuine model-dependence rather than a single universal mechanism (§ 4).

Individual configurations follow the aggregated pattern. Phi-3 one-shot improves under INT4 at all three temperatures, most at T=0.7 (0.180 vs. 0.313); Llama 3 three-shot improves at T=0.3 (0.187 vs. 0.247); and Llama 3 zero-shot degrades most at T=0.7 (0.207 vs. 0.133), the single cell driving much of that row's averaged degradation.

3.2 Temperature Sensitivity and Decoding Behavior

Quantization interacts with temperature. As shown in Table 6 (Appendix A.3), greedy decoding at T=0.0 yields the best stability. Llama 3 INT4 is nearly flat (0.200 → 0.193) while Phi-3 INT4 drops from 0.410 to 0.326. Pooled across models and prompts, GSM8K accuracy falls roughly twice as far as MMLU's, widening the gap monotonically from 4.0 to 7.0 points: multi-step arithmetic is

Table 1: Quantization impact by prompt on Llama 3 (averaged over temperatures). Positive relative change = degradation under INT4 (§ 2.3).

Prompt Type	FP16 (avg acc)	INT4 (avg acc)	Rel. change (%), +=worse
Zero-shot	0.200	0.164	+17.32
Zero-shot CoT	0.204	0.187	+8.22
Concise instruction	0.207	0.202	+1.82
Detailed instruction	0.229	0.216	+5.55
One-shot	0.169	0.187	-10.58
Three-shot	0.191	0.227	-18.89
Five-shot	0.191	0.202	-5.96

Table 2: Quantization impact by prompt on Phi-3 (averaged over temperatures). Positive relative change = degradation under INT4 (§ 2.3).

Prompt Type	FP16 (avg acc)	INT4 (avg acc)	Rel. change (%), +=worse
Zero-shot	0.438	0.460	-4.98
Zero-shot CoT	0.496	0.504	-1.65
Concise instruction	0.360	0.400	-18.74
Detailed instruction	0.340	0.356	-4.74
One-shot	0.242	0.333	-42.75
Three-shot	0.269	0.298	-10.97
Five-shot	0.267	0.260	+2.22

Table 3: Accuracy, parse rate, and strict contract compliance, pooled over prompts and temperatures. Brackets: 95% Wilson intervals. Parse rate is saturated and non-discriminative. Under INT4, Phi-3 accuracy rises while its contract compliance nearly halves; Llama 3 is stable on both. Appendix A.2 reports robustness to the compliance definition.

Model	Precision	Accuracy	Parse rate	Compliance	Median len.
Llama 3	FP16	0.199 [0.185, 0.213]	1.000	0.929 [0.920, 0.938]	10
Llama 3	INT4	0.198 [0.184, 0.212]	1.000	0.948 [0.939, 0.955]	10
Phi-3	FP16	0.344 [0.328, 0.361]	0.993	0.107 [0.097, 0.119]	262
Phi-3	INT4	0.373 [0.356, 0.390]	0.993	**0.055** [0.048, 0.064]	337

more sensitive to sampling stochasticity than constrained symbolic selection, consistent with errors in intermediate steps propagating to the final answer (Li et al., 2026).

3.3 Strict Contract Compliance

Accuracy and compliance move independently, and the compliance effect is the larger one. Table 3 reports both alongside parse rate. Parse rate is saturated, between 0.993 and 1.000 for all four model variants, and changes by at most 0.007 under quantization; measured this way, the models appear uniformly well behaved. Strict contract compliance tells a different story. Phi-3 satisfies the output contract in only 10.7% of FP16 generations (95% interval $[9.7, 11.9]$), and this falls to 5.5% ($[4.8, 6.4]$) under INT4, a 48.5% relative reduction (paired McNemar test, $p{=}7.5{\times}10^{-24}$: 213 generations lose compliance under INT4 and 49 gain it). Llama 3 shows higher compliance under INT4 than FP16 (0.948 vs. 0.929; $p{=}2.3{\times}10^{-8}$). Over the same comparison, Phi-3 accuracy does not fall; it rises slightly, from 0.344 to 0.373 ($p{=}0.0026$), while Llama 3 accuracy is unchanged ($p{=}0.91$). The two Phi-3 accuracy intervals in Table 3 overlap narrowly, so we claim only that accuracy did not fall.

The decoupling is therefore directly observable rather than inferred. Restricting attention to generations that are *correct*, the share that nonetheless violate the contract rises from 89.3% to 93.4% for Phi-3 under INT4, while falling from 12.8% to 9.6% for Llama 3. Expressed over all generations, Phi-3's correct-but-non-compliant rate increases from 0.308 to 0.348 ($p{=}1.8{\times}10^{-5}$), whereas for Llama 3 it moves in the opposite direction, from 0.025 to 0.019 ($p{=}0.0042$); the two models diverge on the same quantity. Median response length provides an independent behavioral signal pointing the same way: Phi-3 grows from 262 to 337 characters under INT4, while Llama 3 remains at 10 characters. We do not claim that verbosity increases with temperature: the per-temperature medians for Phi-3 FP16 in fact *decrease* across the range ($278 \to 262 \to 240$), so that relationship does not hold in our data.

Table 4: Strict contract compliance by prompt family, averaged over temperatures. For Phi-3 every prompt family loses compliance under INT4, with the largest relative losses in the exemplar-based families; one-shot starts from 0.027 and is not interpreted (see text). For Llama 3 no family loses compliance.

	Llama 3			Phi-3		
Prompt Type	FP16	INT4	Rel. (%)	FP16	INT4	Rel. (%)
Zero-shot	0.984	0.987	+0.2	0.027	0.022	−16.7
Zero-shot CoT	1.000	1.000	0.0	0.024	0.013	−45.5
Concise instruction	0.798	0.833	+4.5	0.296	0.176	−40.6
Detailed instruction	1.000	1.000	0.0	0.131	0.093	−28.8
One-shot	0.749	0.820	+9.5	0.027	0.004	−83.3
Three-shot	0.978	0.993	+1.6	0.122	0.029	**−76.4**
Five-shot	0.996	1.000	+0.4	0.124	0.049	**−60.7**

3.4 Model and Prompt Dependence of the Compliance Effect

The compliance loss is not uniform across prompts, and its structure differs from the accuracy results in § 3.1. Table 4 breaks compliance down by prompt family. Quantization shifts Phi-3 compliance broadly: it falls in all seven families, with the largest absolute loss in concise instruction ($0.296 \rightarrow 0.176$) and the largest relative losses in three- and five-shot (-76.4%, -60.7%). The change is separately detectable in three of the seven prompt families after Holm correction, those same three (Appendix A.7); one-shot's -83.3% starts from an FP16 compliance of 0.027 and does not survive correction. For Llama 3, no prompt family loses compliance; the two families with the lowest FP16 compliance, one-shot (0.749) and concise instruction (0.798), both improve under INT4.

This matters for interpretation. On accuracy, exemplar prompting looked protective for Llama 3 (§ 3.1). On compliance, the same exemplar prompts are where Phi-3 degrades most. In our data, format-only demonstrations constrain *what answer* the model produces more reliably than they constrain *how* it formats the response, and under INT4 the two come apart; whether in-domain exemplars would behave the same way is untested. Any claim that exemplar-based prompting stabilizes behavior under quantization is therefore metric-dependent as well as model-dependent, and does not hold as a general prescription. Temperature modulates compliance only weakly by comparison: Phi-3 compliance declines mildly in both precisions ($0.119 \rightarrow 0.087$ FP16; $0.057 \rightarrow 0.052$ INT4), and the INT4–FP16 gap persists at every temperature including greedy decoding, so the effect is not an artifact of sampling stochasticity.

4 Analysis and Discussion

We organize the discussion around a sharp distinction. What we *observe* is precision-dependent behavior: quantization changes both whether the model obeys its output contract and which prompts work best, and it does so differently for the two models. What we *hypothesize* is a distributional explanation for these behaviors, which our output-level measurements cannot confirm.

4.1 What Is Observed

Correctness and compliance decouple, and compliance moves further. The clearest result in this study is that INT4 halves Phi-3's strict contract compliance, from 0.107 to 0.055 (-48.5% relative, $p{=}7.5{\times}10^{-24}$), while its accuracy over the same comparison *rises*, from 0.344 to 0.373 ($p{=}0.0026$). The two metrics therefore do not merely differ in magnitude; they move in opposite directions. An accuracy-only evaluation would report a small improvement, which is what makes this failure class invisible to evaluations that report accuracy alone (Husom et al., 2025). The effect is not universal: Llama 3 compliance is stable and marginally higher under INT4 ($0.929 \rightarrow 0.948$), so the decoupling belongs to particular model–precision pairs rather than to low-bit inference in general. Median response length moves with compliance for Phi-3 ($262 \rightarrow 337$ characters) and is flat for Llama 3, as expected if the violations consist largely of appended rationale; we treat length as

corroborating rather than primary evidence, since it proxies contract violation rather than measuring it. Since the contract forbids the rationale that helps on GSM8K, one shift toward longer responses could in principle lower compliance and raise accuracy together. The records do not support that account: Phi-3's accuracy gain appears in both strata, $0.343 \rightarrow 0.448$ among compliant generations and $0.345 \rightarrow 0.369$ among non-compliant ones, and in an exploratory logistic model the INT4 coefficient is essentially unchanged after adjusting for log response length ($+0.120$), providing no evidence that response length explains the gain; its statistical significance is sensitive to the assumed dependence structure, so we treat the model as exploratory (Appendix A.5).

Prompt robustness is model-dependent, and metric-dependent too. On accuracy, Llama 3 shows a clean split: all three exemplar-based prompts improve under INT4 and all four non-exemplar prompts degrade (Table 1), consistent with exemplars acting as structural anchors (Brown et al., 2020; Wei et al., 2022). On Phi-3 it already fits only partly for accuracy (Table 2), and on compliance it fails outright: Phi-3 loses compliance under INT4 in every prompt family, and the largest relative losses are in the three- and five-shot families, although the largest absolute loss is concise instruction (Table 4). A prompt that preserves or improves Phi-3's answers is therefore no guarantee of preserved formatting. What follows directly from the tables, with no appeal to mechanism, is that the quantization robustness of a prompting strategy is a joint property of model, precision, *and* the metric being optimized, and cannot be assumed to transfer across any of the three.

Greedy decoding protects accuracy but not compliance. INT4 accuracy declines with temperature for both models (§ 3.2), yet Phi-3's INT4–FP16 compliance gap is present at every temperature including $T{=}0$ ($0.119 \rightarrow 0.057$ at greedy decoding). Sampling stochasticity is therefore not the cause of the compliance effect, only a modest additional stressor.

4.2 What Is Hypothesized

We measure neither token-level probability distributions nor hidden activations, so what follows is an interpretive hypothesis, not a result of this study. The common account holds that quantization flattens the token distribution and weakens the dominance of high-probability continuations, which is consistent with instruction-tuned formatting behaviour (Ouyang et al., 2022) eroding before task accuracy does (Gong et al., 2024; Dettmers et al., 2022). One commonly assumed consequence is *not* supported. Structural anchoring predicts that exemplar-rich prompts should be more robust under low precision, because demonstrations narrow the sampled region. Our compliance results invert that prediction without supporting a simpler one. Phi-3's three- and five-shot families lose 76% and 61% of their compliance, more in relative terms than any non-exemplar family, yet the largest absolute loss falls on concise instruction, which contains no exemplars (Table 4): the erosion is broad, not specific to format demonstrations. Llama 3 shows the opposite pattern under one-shot prompting, with compliance rising from 337 to 369 generations under INT4 (Holm-adjusted $p{=}0.000128$), the only prompt-level Llama 3 comparison detectable after correction. Our demonstrations are format-only, so this falsifies anchoring for format demonstrations rather than for in-domain few-shot prompting, and it is compatible with Liu et al. (2024), whose 4-bit in-context learning results are accuracy-scored. Whatever demonstrations constrain, it is not the surface form of the response. That falsified prediction is the main reason we do not present this account as a mechanism for our results; logit-entropy and hidden-state comparisons across precisions would be needed to test it.

4.3 Limitations

Our conclusions are bounded by the design, and two validity checks remain outstanding: seed replication and rendered-prompt validation (Appendix A.8). *Contract design:* the contract forbids rationale where rationale drives accuracy, so compliance and accuracy are in tension by construction; this depresses absolute accuracy and leaves verbosity as an untested alternative account of the decoupling. A contract that permits reasoning before a final answer line is the most informative follow-up. *Prompt delivery:* Phi-3's FP16 compliance of 0.107 may partly reflect how its chat template handles the system role: the documented phi-3-mini chat template contains user and assistant roles but no system role (Abdin et al., 2024). We have not yet verified the rendered prompts or tested user-turn placement. *Statistics:* precision comparisons are paired by item, prompt and temperature, but each

configuration is still a single run with no seed replication at $T>0$, so variance across stochastic seeds is unmeasured, and the primary p-values are unadjusted for multiplicity. *Models and precisions:* Llama 3 8B and Phi-3 3.8B differ in family and scale and are compared at one quantization point (q4_K_M), so we can neither attribute the Phi-3 pattern to architecture, data, or size nor read it as a dose–response; a same-lineage control (e.g., Llama 3.2 3B) and further bit-widths are the natural next experiments. *Scope:* both benchmarks are mathematical; the exemplars are synthetic format demonstrations; only temperature is varied; and adherence is one format constraint rather than a benchmark such as IFEval (Zhou et al., 2023), which Kurt (2026) and Huang et al. (2026) use, so we establish that the decoupling occurs, not how far it generalizes across constraint types. Strict compliance is definition-sensitive (Appendix A.2); the MMLU extractor falls back to the last standalone A–D letter for 1.0% of Llama 3 responses but 21.4–26.0% of Phi-3's, though the INT4 accuracy difference holds within both subsets (Appendix A.6); and we have no logit- or activation-level evidence for the mechanism in § 4.2.

Implications. Evaluation of a quantized instruction-tuned model should report accuracy, strict compliance and response-length drift together, since accuracy alone is insufficient when the first two move in opposite directions; deterministic decoding does not protect compliance, which is already degraded at $T=0$; and where output format is part of a machine interface, structural enforcement (stop sequences, token caps, constrained decoding) is the only measure that does not depend on the model's own instruction following surviving compression.

5 Conclusions

Across 12,600 controlled evaluations, we find that INT4 quantization can degrade instruction following while leaving task accuracy untouched or improving it: Phi-3's strict contract compliance falls by 48.5% relative ($0.107 \rightarrow 0.055$) while its accuracy rises ($0.344 \rightarrow 0.373$), and Llama 3 degrades on neither. The effect is invisible to standard evaluation twice over, since accuracy does not reveal it and lenient answer extraction is saturated above 0.99. It is also structured rather than uniform: Phi-3's compliance falls across all seven prompt families, with the largest absolute loss under concise instruction and the largest relative losses in the three- and five-shot settings (-76% and -61%), which are among those that improve Llama 3's accuracy, so robustness under compression is a joint property of model, precision, and the metric optimized. Our distributional hypothesis predicts greater robustness under exemplar-based prompting, which this breadth does not support. We therefore leave it open for logit- and activation-level testing alongside same-lineage comparisons and benchmarks such as IFEval. For practitioners: a model that still answers correctly after quantization has not thereby been shown to still follow instructions.

AI Use Statement

Generative AI tools were used to assist with language editing, organization, and refinement of author-written text. The authors reviewed and verified all AI-assisted content and remain responsible for the accuracy, interpretation, and integrity of the manuscript. Generative AI tools were not used to generate experimental data, conduct statistical analyses, or determine or interpret the reported results.

Ethics Statement

This study evaluates publicly released, open-weight instruction-tuned language models on publicly available benchmark datasets (GSM8K and MMLU). The study involves no human participants, private data, or personally identifiable information.

Reproducibility Statement

Section 2 describes the model identifiers, quantization scheme, inference settings, prompts and evaluation protocol; Appendices A.1 and A.4–A.7 give the prompt templates and the statistical and extraction procedures. The harness, per-generation records and analysis code are in an anonymous repository: https://anonymous.4open.science/r/artifact-7421.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Patrick Blumenberg, Thomas Graave, and Tim Fingscheidt. Improving block-wise LLM quantization by 4-bit block-wise optimal float (BOF4): Analysis and variations. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. arXiv:2505.06653.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems*, 35: 30318–30332, 2022.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Zhuocheng Gong, Jiahao Liu, Jingang Wang, Xunliang Cai, Dongyan Zhao, and Rui Yan. What makes quantization for large language model hard? an empirical study from the lens of perturbation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18082–18089, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Jahid Hasan. Optimizing large language models through quantization: A comparative analysis of ptq and qat techniques. *arXiv preprint arXiv:2411.06084*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Junyuan Hong, Jinhao Duan, Chenhui Zhang, Zhangheng Li, Chulin Xie, Kelsey Lieberman, James Diffenderfer, Brian R. Bartoldson, Ajay Kumar Jaiswal, Kaidi Xu, Bhavya Kailkhura, Dan Hendrycks, Dawn Song, Zhangyang Wang, and Bo Li. Decoding compressed trust: Scrutinizing the trustworthiness of efficient LLMs under compression. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 18611–18633, 2024.

Zhehao Huang, Yuhang Liu, Baijiong Lin, Yixin Lou, Zhengbao He, Hanling Tian, Tao Li, and Xiaolin Huang. RAIN-merging: A gradient-free method to enhance instruction following in large reasoning models with preserved thinking format. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Erik Johannes Husom, Arda Goknil, Merve Astekin, Lwin Khin Shar, Andre Kåsen, Sagar Sen, Benedikt Andreas Mithassel, and Ahmet Soylu. Sustainable llm inference for edge ai: Evaluating quantized llms for energy efficiency, output accuracy, and inference latency. *ACM Transactions on Internet of Things*, 2025.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Uygar Kurt. Which quantization should I use? a unified evaluation of llama.cpp quantization on Llama-3.1-8b-instruct, 2026. URL https://arxiv.org/abs/2601.14277.

Eldar Kurtic, Alexandre Marques, Shubhra Pandit, Mark Kurtz, and Dan Alistarh. "give me bf16 or give me death"? accuracy-performance trade-offs in llm quantization. *arXiv preprint arXiv:2411.02355*, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Jemin Lee, Sihyeong Park, Jinse Kwon, Jihun Oh, and Yongin Kwon. 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)*, 2025. arXiv:2409.11055.

Zhen Li, Yupeng Su, Songmiao Wang, Runming Yang, Congkai Xie, Aofan Liu, Ming Li, Jiannong Cao, Yuan Xie, Ngai Wong, and Hongxia Yang. Quantization meets reasoning: Exploring and mitigating degradation of low-bit LLMs in mathematical reasoning. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. arXiv:2505.11574.

Yesheng Liang, Haisheng Chen, Zihan Zhang, Song Han, and Zhijian Liu. ParoQuant: Pairwise rotation quantization for efficient reasoning LLM inference. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. arXiv:2511.10645.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6:87–100, 2024.

Peiyu Liu, Zikang Liu, Ze-Feng Gao, Dawei Gao, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. Do emergent abilities exist in quantized large language models: An empirical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 5174–5190, Torino, Italia, 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.461.

Yutong Liu, Cairong Zhao, and Guosheng Hu. A comprehensive evaluation on quantization techniques for large language models. *arXiv preprint arXiv:2507.17417*, 2025.

Meta AI. Meta Llama 3 model card, 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md. Llama 3 8B Instruct, April 2024 release.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Mariam Rakka, Marios Fournarakis, Olga Krestinskaya, Jinane Bazzi, Khaled N Salama, Fadi Kurdahi, Ahmed M Eltawil, and Mohammed E Fouda. Mixed-precision quantization for language models: Techniques and prospects. *arXiv preprint arXiv:2510.16805*, 2025.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RIu5lyNXjT.

Ying Sheng, Lianmin Zheng, Binhang Yuan, Zhuohan Li, Max Ryabinin, Beidi Chen, Percy Liang, Christopher Ré, Ion Stoica, and Ce Zhang. Flexgen: High-throughput generative inference of large language models with a single gpu. In *International Conference on Machine Learning*, pp. 31094–31116. PMLR, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Xiaoxia Wu, Zhewei Yao, and Yuxiong He. Zeroquant-fp: A leap forward in llms post-training w4a8 quantization using floating-point formats. *arXiv preprint arXiv:2307.09782*, 2023.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International conference on machine learning*, pp. 38087–38099. PMLR, 2023.

Zhewei Yao, Xiaoxia Wu, Cheng Li, stephen youn, and Yuxiong He. From comprehensive study to low-rank compensation: Exploring post-training quantization in LLMs, 2023. URL https://openreview.net/forum?id=qO9VagA7kF.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

A Appendix

A.1 System Prompt and Prompt Templates

This appendix reproduces verbatim the system prompt and all seven user-prompt templates used in the experiments. Because both evaluated models are small and context-limited (4,096–8,192 tokens, with `num_ctx`=2048 at inference), the templates are deliberately concise; the few-shot demonstrations use short synthetic exemplars rather than full benchmark items so that five-shot prompts fit comfortably within the context budget.

A.1.1 System Prompt (output contract)

The following system prompt was applied identically to every one of the 84 configurations, and defines the output contract whose violation we measure in § 3:

```
Your task is to answer the question. CRITICAL: You must
provide ONLY the final answer in the format of Answer:
<answer>. Do not provide any explanations. Do not provide
any extra text.
```

A.1.2 User-Prompt Templates

Below, {`question`} denotes the benchmark item. For MMLU, the four answer options (A–D) are appended to the question text before substitution. Where a template differs between GSM8K and MMLU, both variants are shown.

Zero-shot.

```
{question}
Answer:
```

Zero-shot chain-of-thought.

```
Let's solve this step by step:
{question}
Think through it, then provide ONLY the final answer after
"Answer:"
Answer:
```

Concise instruction.

```
Question: {question}
Can you provide the final answer to the question? Only
provide the answer.
Answer:
```

Detailed instruction.

```
Follow these steps:
1. Read the question carefully
2. Identify what is being asked
3. Work through the necessary reasoning
4. Provide only the final answer
Question: {question}
Answer:
```

One-shot (GSM8K).

```
Question: If John has 5 apples and buys 3 more, how many
apples does he have?
Answer: 8
Question: {question}
Answer:
```

One-shot (MMLU).

```
Question: What is the capital of France?
A. London
B. Berlin
C. Paris
D. Madrid
Answer: C
Question: {question}
Answer:
```

Three-shot (GSM8K).

```
Question: If John has 5 apples and buys 3 more, how many
apples does he have?
Answer: 8
Question: A baker made 24 cookies and sold 15. How many
are left?
Answer: 9
Question: Tom has $20 and buys a book for $12. How much
money is left?
Answer: 8
Question: {question}
Answer:
```

Three-shot (MMLU).

```
Question: What is the capital of France?
A. London
B. Berlin
C. Paris
D. Madrid
Answer: C
Question: What is 2 + 2?
A. 3
B. 4
C. 5
D. 6
Answer: B
Question: Which planet is closest to the Sun?
A. Venus
B. Earth
C. Mercury
D. Mars
```

```
Answer: C
Question: {question}
Answer:
```

Five-shot (GSM8K).

```
Question: If John has 5 apples and buys 3 more, how many
apples does he have?
Answer: 8
Question: A baker made 24 cookies and sold 15. How many
are left?
Answer: 9
Question: Tom has $20 and buys a book for $12. How much
money is left?
Answer: 8
Question: Sarah has 48 stickers and gives half to her
friend. How many does she have left?
Answer: 24
Question: A parking lot has 36 cars, 12 leave, then 8
arrive. How many cars now?
Answer: 32
Question: {question}
Answer:
```

Five-shot (MMLU).

```
Question: What is the capital of France?
A. London
B. Berlin
C. Paris
D. Madrid
Answer: C
Question: What is 2 + 2?
A. 3
B. 4
C. 5
D. 6
Answer: B
Question: Which planet is closest to the Sun?
A. Venus
B. Earth
C. Mercury
D. Mars
Answer: C
Question: What is the chemical symbol for water?
A. H2O
B. CO2
C. O2
D. N2
Answer: A
Question: Who wrote "Romeo and Juliet"?
A. Charles Dickens
B. William Shakespeare
C. Jane Austen
D. Mark Twain
Answer: B
Question: {question}
Answer:
```

A.2 Compliance Metric Definition

We distinguish two output-level measures that are easily conflated.

Parse rate is the fraction of responses from which an answer could be extracted by the evaluation harness. The extractor is deliberately lenient: it scans for lines containing any of the markers `answer`, `final answer`, `result`, or `therefore`, and falls back to numeral extraction and English number-word conversion. Parse rate therefore measures whether an answer is *recoverable*, not whether the model obeyed the contract, which is why it exceeds 0.99 in every configuration and carries little discriminative information.

Contract compliance is the fraction of responses that satisfy the system prompt as written, namely that the response consists of exactly one non-empty line and that this line begins with the marker `Answer:`. Formally, for response text t, let $L(t)$ be the list of non-empty lines of t after stripping surrounding whitespace; the response is compliant iff $|L(t)| = 1$ and $L(t)_1$ matches the case-insensitive regular expression `^\s*Answer\s*:`. The test enforces the single-line format and therefore excludes appended rationale, which is the violation the contract is written against; it does not separately verify that nothing follows the answer on that one line.

Because compliance is definition-sensitive, we report the two looser variants as a robustness check in Table 5: *marker-anywhere* (the regex matches anywhere in t) and *marker-at-line-start* (the regex matches at the start of any line), neither of which enforces the single-line requirement. The qualitative conclusion, that Phi-3 compliance degrades sharply under INT4 while Llama 3 compliance does not, holds under the strict definition; the looser definitions, which tolerate appended explanations, show the effect attenuated or absent because they do not measure the clause that Phi-3 actually violates.

Table 5: Sensitivity of measured compliance to the metric definition. Parse rate is near-saturated and non-discriminative. Only the strict definition, which enforces the single-line output requirement and so excludes appended rationale, isolates the INT4 effect on Phi-3.

Model	Precision	Parse rate	Marker anywhere	Marker line-start	Strict contract
Llama 3	FP16	0.9997	0.9292	0.9292	0.9292
Llama 3	INT4	1.0000	0.9476	0.9476	0.9476
Phi-3	FP16	0.9927	0.6203	0.4365	0.1073
Phi-3	INT4	0.9927	0.6222	0.4390	0.0552

A.3 Runtime and Temperature Table

All 84 configurations complete in 59,512 seconds (16.5 hours) on the hardware described in § 2, averaging 4.7 seconds per query; we note this only to confirm that the full design is practical to run locally, not as a finding in itself.

Table 6: Mean accuracy by decoding temperature under INT4. Adherence indicators, previously reported alongside these values, are given in Table 3.

Model	T=0.0	T=0.3	T=0.7
Llama 3 INT4	0.200	0.200	0.193
Phi-3 INT4	0.410	0.384	0.326

A.4 Relative-Change Estimator

Averaging the three temperature-specific relative changes and taking the relative change of the temperature-averaged accuracies give the same sign for all 14 prompt-level effects but different magnitudes. The largest divergence is Phi-3 one-shot, -42.8% under the reported estimator and -37.6% under the alternative, because the T=0.7 cell alone contributes -74.1%; Phi-3 concise instruction moves from -18.7% to -11.1%, and the remaining twelve rows differ by less than one point.

A.5 Accuracy by Compliance Status

Table 7 reports accuracy separately for compliant and non-compliant generations. Phi-3's INT4 accuracy gain appears in both strata, so it is not an artifact of the shift in compliance status. In a logistic model of correctness on precision, log response length and dataset, fitted to Phi-3's 6,300 generations, the estimated INT4 coefficient stays at approximately $+0.120$ across alternative variance specifications, while its nominal p-value varies substantially with the assumed dependence structure, from 0.010 clustering by item to 0.354 clustering by prompt, temperature and precision. We therefore treat this model as exploratory, do not use its p-value for inference, and rest the argument on the stratified counts.

Table 7: Accuracy by strict-compliance status, pooled over prompts, temperatures and datasets (n=3,150 per row).

Model	Precision	Compliant	Accuracy	Non-compliant	Accuracy
Llama 3	FP16	2927	0.187	223	0.359
Llama 3	INT4	2985	0.189	165	0.364
Phi-3	FP16	338	0.343	2812	0.345
Phi-3	INT4	174	0.448	2976	0.369

A.6 MMLU Extraction Stages

Table 8 decomposes the 1,050 MMLU responses in each model–precision cell by the stage at which the extractor succeeds. The stages are mutually exclusive and exhaustive: an explicit answer pattern, the last standalone A–D letter, a match between the response and an option's text, a fuzzy option match, and no extraction at all. The fuzzy stage fires twice in 4,200 MMLU generations. The *non-pattern subset* used for the sensitivity check below is every response without an explicit answer pattern, regardless of which later stage resolved it, including those never resolved, which are scored as incorrect. Accuracy is lower in that subset, and the INT4 difference holds within both: for Phi-3, $0.373 \rightarrow 0.384$ among pattern-matched responses and $0.277 \rightarrow 0.293$ among the rest.

Table 8: Share of MMLU responses resolved at each extraction stage (n=1,050 per row).

Model	Precision	Answer pattern	Standalone letter	Option text	Fuzzy	None
Llama 3	FP16	0.980	0.010	0.010	0.000	0.001
Llama 3	INT4	0.979	0.010	0.010	0.001	0.000
Phi-3	FP16	0.770	0.214	0.011	0.000	0.005
Phi-3	INT4	0.714	0.260	0.022	0.001	0.003

A.7 Per-Cell Counts and Prompt-Level Tests

Every model–precision–prompt cell contains exactly 450 generations (150 items $\times$ 3 temperatures). Table 9 gives the strict-compliance count, 95% Wilson interval and paired McNemar test for each of the 14 prompt-level comparisons. Fourteen prompt-level McNemar comparisons were prespecified as a secondary exploratory family; Holm correction is applied across all 14, with zero-discordance comparisons assigned p=1. They are not part of the six primary comparisons of § 3.3. Each test uses the exact binomial form when fewer than 25 pairs are discordant and the continuity-corrected form otherwise.

A.8 Outstanding Validity Checks

Two checks remain outside the scope of this version.

- **Seed replication.** Each configuration is a single run; at $T>0$ we have no estimate of variance across stochastic seeds. At least three seeds per configuration would be needed to separate sampling variance from the precision effect.

Table 9: Strict compliance per prompt family, with counts out of 450, 95% Wilson intervals, and paired McNemar tests (Holm-adjusted across all 14 comparisons). Gain and loss are the discordant pairs: generations that become compliant and that stop being compliant under INT4. Two Llama 3 cells are compliant in all 450 generations at both precisions, leaving no discordant pairs; they are assigned p=1 and retained in the family.

Model	Prompt	FP16	INT4	Gain	Loss	Holm p
Llama 3	Zero-shot	443 [0.968, 0.992]	444 [0.971, 0.994]	4	3	1.00
Llama 3	Zero-shot CoT	450 [0.992, 1.000]	450 [0.992, 1.000]	0	0	1.00
Llama 3	Concise instruction	359 [0.758, 0.832]	375 [0.796, 0.865]	27	11	0.13
Llama 3	Detailed instruction	450 [0.992, 1.000]	450 [0.992, 1.000]	0	0	1.00
Llama 3	One-shot	337 [0.707, 0.787]	369 [0.782, 0.853]	41	9	1.3×10^{-4}
Llama 3	Three-shot	440 [0.960, 0.988]	447 [0.981, 0.998]	7	0	0.13
Llama 3	Five-shot	448 [0.984, 0.999]	450 [0.992, 1.000]	2	0	1.00
Phi-3	Zero-shot	12 [0.015, 0.046]	10 [0.012, 0.040]	4	6	1.00
Phi-3	Zero-shot CoT	11 [0.014, 0.043]	6 [0.006, 0.029]	4	9	1.00
Phi-3	Concise instruction	133 [0.255, 0.339]	79 [0.143, 0.213]	17	71	2.3×10^{-7}
Phi-3	Detailed instruction	59 [0.103, 0.165]	42 [0.070, 0.124]	13	30	0.13
Phi-3	One-shot	12 [0.015, 0.046]	2 [0.001, 0.016]	1	11	0.06
Phi-3	Three-shot	55 [0.095, 0.156]	13 [0.017, 0.049]	6	48	3.1×10^{-7}
Phi-3	Five-shot	56 [0.097, 0.158]	22 [0.033, 0.073]	4	38	4.3×10^{-6}

- **Prompt delivery and harness validation.** The fully rendered prompt string after each model's chat template is applied, an ablation that places the contract in the user turn, and accuracy under standard prompting without the contract, against published values.

Scope notes moved from § 4.3 for space. The findings may not generalize to open-ended or non-reasoning tasks. Our primary metric tests one format constraint, whereas instruction-level IFEval scoring covers a broad family of verifiable constraints. We report parse rate only as a saturated auxiliary measure, to document that the standard extraction-based view of adherence is blind to this effect (Table 3), and response length as a corroborating proxy; the looser compliance variants in Table 5 attenuate or hide the effect precisely because they tolerate appended rationale. Top-p and top-k sampling are not studied, and 2- and 3-bit, other mixed-precision, and activation-quantized regimes may behave differently.