Not All Linguistic Capabilities Degrade Equally Under Quantization

Arunabh Majumdar
Independent Researcher
arunabh.majumdar@gmail.com

Abstract

Low-precision quantization enables efficient LLM deployment, but its effect on distinct linguistic capabilities is poorly characterised: prior evaluations report aggregate metrics (perplexity, task accuracy), obscuring differential degradation. We evaluate four instruction-tuned models (Qwen 2.5-7B, Mistral 7B, Llama 3.1-8B, Gemma-2-9B) across four NLG tasks (summarisation, dialogue, data-to-text, simplification) and five quantization conditions (4-bit AWQ/GPTQ/GGUF, 2-bit GGUF, 16-bit baseline), measuring lexical accuracy (ROUGE), syntactic well-formedness (grammar error rate), factual consistency (AlignScore), and discourse/instruction-following (GPT-4o-mini judge). Across 92 conditions (≈500 samples each): (1) Friedman tests reject uniform degradation in **84%** of model-task-bit-width combinations ($p < 0.05$); (2) factual consistency is more stable than grammar at 4-bit across all architectures (AlignScore mean change: -1.8%, $d = -0.355$; grammar degrades more in 18/27 cells, Bonferroni-corrected); (3) grammar direction is *architecture-dependent*: Qwen degrades ($+19.0\%$) while Llama improves (-12.8%); (4) grammar improvement on dialogue (3/4 models) is not a length artefact: per-sample length–grammar correlations are near zero, and Mistral falsifies the length hypothesis (output $+34.4\%$ longer, grammar rate -37.9%). Compression does not yield a single robustness profile; it induces capability-specific changes depending on architecture, task, and quantization method.

1 Introduction

Quantization — reducing the numerical precision of model weights from 16-bit to 8-bit, 4-bit, or lower — has become the primary mechanism for deploying large language models under memory and throughput constraints (Frantar et al., 2023; Lin et al., 2024; Dettmers et al., 2022). Production deployments routinely use 4-bit quantized models without systematic understanding of which linguistic capabilities are preserved and which are not.

The dominant evaluation paradigm answers this question at the level of aggregate benchmarks: perplexity, question-answering accuracy, or overall summarisation quality scores. These measures conflate distinct linguistic capabilities into a single number. A model whose factual accuracy holds steady while syntactic well-formedness collapses will appear to "pass" a ROUGE-based evaluation, yet fail qualitatively in ways that matter for downstream use.

We ask a more fine-grained question: **does quantization degrade different linguistic capabilities at different rates, and is this pattern consistent across model architectures and NLG tasks?** The null hypothesis — that all capabilities degrade uniformly as bit-width decreases — is the implicit assumption behind aggregate evaluations. We test it directly.

Our study covers four models from distinct architectural families, four NLG tasks that isolate different primary capability dimensions, and five quantization conditions spanning the range from 8-bit to 2-bit precision, using three quantization methods (AWQ, GPTQ, GGUF). Across 92 experimental conditions (≈ 500 samples each), our results indicate:

1. Friedman tests reject uniform degradation in **84%** of conditions (27/32, $p < 0.05$); post-hoc Wilcoxon confirms grammar degrades significantly more than factual consistency in 18/27 cells after Bonferroni correction.

2. Factual consistency (AlignScore) is substantially more stable than grammar at 4-bit across all four architectures ($d = -0.355$, mean $= -1.8\%$; no model crosses the 10% degradation threshold).

3. Grammar direction is *architecture-dependent*:

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 310–320
October 17-21, 2026 ©2026 Association for Computational Linguistics

Qwen degrades ($+19.0\%$, d up to $+0.320$) while Llama improves (-12.8%), with the largest differential on Llama dialogue: grammar $d = -1.142$ vs AlignScore $d = -0.131$ at 4-bit AWQ.

4. Grammar improvement on dialogue (3/4 models) is not a length artefact: per-sample length–grammar correlations are near zero; Mistral output gets longer ($+34.4\%$) while grammar rate *decreases* (-37.9%).

Within the four models and settings evaluated here, compression does not produce a single robustness profile; it induces capability-specific changes whose magnitude and direction depend on the interaction between architecture, task, and quantization method.

2 Related Work

Early quantization papers (Frantar et al., 2023; Lin et al., 2024; Dettmers et al., 2022; Gerganov et al., 2023) reported quality via perplexity and zero-shot benchmarks — single aggregate numbers that do not decompose change into distinct linguistic capability dimensions. Later work broadened task coverage but still reports *per-task* rather than *per-capability* accuracy: Xu et al. (2024) examined safety and bias axes; Jin et al. (2024) evaluated knowledge and alignment across three model scales; Li et al. (2024) benchmark eleven model families across basic NLP, emergent-ability, trustworthiness, dialogue, and long-context task *categories*; Lee et al. (2025) evaluate accuracy against task difficulty and model scale over 13 datasets, using an LLM judge for reasoning, coding, and STEM tasks. In all four, a task (e.g. dialogue, summarisation) receives one aggregate score rather than being decomposed into the linguistic capabilities that compose it, so a model that maintains fluency while losing factual grounding, or vice versa, is not distinguishable from uniform degradation. Fu et al. (2025) studied three truthfulness sub-dimensions, the closest prior work to a within-task capability decomposition, but confined to one capability domain.

Three papers conduct single-capability or single-axis deep dives that complement our multi-dimensional approach. Li et al. (2025) find mathematical reasoning disproportionately sensitive to quantization; Mekala et al. (2025) show quantization effects on long-context tasks. Neither compares the target capability against others, making it impossible to assess whether the observed sensitivity is *selective* or simply reflects global degradation. Marchisio et al. (2024) is methodologically closest to our design — combining automatic metrics, LLM-as-judge, and human evaluation — but decomposes degradation along the axis of *language* rather than linguistic capability; their central finding, that automatic metrics can substantially underestimate quantization damage relative to human judgement (a 1.7% average automatic-metric drop vs. a 16.0% human-rated drop for Japanese), is why we treat our own automatic and LLM-judge scores as operationalisations rather than ground truth (§3) and revisit it directly in Limitations.

Our capability framework builds on Gatt and Krahmer (2018)'s survey of NLG quality dimensions and Celikyilmaz et al. (2020)'s evaluation taxonomy, using AlignScore (Zha et al., 2023) for factual consistency and LanguageTool grammar error rate (Naber, 2003) normalised per 100 tokens for syntactic well-formedness. No prior work applies a structured, multi-dimensional NLG capability framework to characterise *differential* degradation across bit-widths within a task, or identifies capability-specific precision thresholds, or reports per-capability effect sizes that allow the *magnitude* of differential degradation to be assessed.

3 Capability Framework

We evaluate six capability dimensions:

- **Lexical accuracy — ROUGE-1**: unigram overlap with the reference output, measuring whether the model reproduces key content words. Standard proxy for surface-level accuracy in summarisation and data-to-text tasks.
- **Lexical accuracy — ROUGE-L**: longest-common-subsequence overlap with the reference output, capturing sentence-level structural similarity beyond unigram overlap.
- **Syntactic well-formedness** (grammar error rate): grammatical error count per 100 tokens, measured with LanguageTool (Naber, 2003). Operationalises surface-level fluency and morpho-syntactic correctness.
- **Factual consistency** (AlignScore): NLI-based entailment score of the generated output with respect to the source document (Zha et al., 2023). Operationalises whether the model preserves source-grounded factual

claims. The relationship is *directional, not symmetric*: AlignScore scores whether each claim in the generated output is entailed by the source, not the converse, so it penalises fabricated or contradicted content but not omission of source content the output simply leaves out. This fits summarisation and data-to-text, where omission is an acceptable compression strategy but fabrication is not; it is a weaker fit for dialogue, where the "source" is conversation history rather than a single grounding document (see Limitations).

- **Discourse coherence** (GPT-4o-mini judge, 1–5): logical structure and ease of following, rated by GPT-4o-mini from incoherent (1) to fully coherent (5).

- **Instruction-following quality** (GPT-4o-mini judge, 1–5): degree to which the output follows the task instruction with clear reasoning, rated from non-compliant (1) to fully compliant (5).

Why these six dimensions. We select capabilities that (i) map onto an established axis in the NLG evaluation literature (Gatt and Krahmer, 2018; Celikyilmaz et al., 2020), (ii) admit a widely used automatic or judge-based proxy so all 92 conditions can be scored at scale, and (iii) span a range of representational depth: surface pattern matching (lexical overlap, local grammaticality), external grounding (factual consistency), and properties requiring long-range dependency tracking (discourse coherence, instruction-following). This depth gradient is the basis for the mechanistic discussion in §7. We do not claim these six dimensions exhaust linguistic competence — register, pragmatics, and cross-sentence entity tracking are not directly measured here. The selection prioritises established, automatically-scorable coverage over completeness.

Limitations of operationalisation. ROUGE measures lexical overlap, not semantic accuracy. Grammar error rate reflects LanguageTool's rule coverage, which differs by register and genre. AlignScore uses an NLI proxy trained on academic and news text; it may under-report factual errors in dialogue. LLM-as-judge scores are subject to model bias and may conflate fluency with quality. We report these metrics as operationalisations, not as direct capability measurements.

Model	Size	Family	Cond.
Qwen 2.5-7B Instruct	7B	Qwen	7
Mistral 7B Instruct v0.3	7B	Mistral	6
Llama 3.1-8B Instruct	8B	Llama	5
Gemma-2-9B Instruct	9B	Gemma	5

Table 1: Models and condition counts. See §4.2 for exclusions.

Task	Dataset	n	Primary dimension
Summarisation	CNN/DM (See et al., 2017)	500	Factual consistency
Dialogue	PersonaChat (Zhang et al., 2018)	500	Discourse coherence
Data-to-text	WebNLG 2020 (Gardent et al., 2017)	500	Lexical accuracy
Simplification	TurkCorpus (Xu et al., 2016)	359†	Syntactic fluency

Table 2: Tasks and datasets. †After deduplication.

4 Experimental Setup

4.1 Models

We evaluate four instruction-tuned models from distinct architectural families (Table 1):

4.2 Quantization Conditions

Each model is evaluated at 16-bit (baseline) and at reduced precision using three methods:

- **AWQ** (Lin et al., 2024): 4-bit and 8-bit. All AWQ models loaded via HuggingFace Transformers' native AWQ path (AwqMarlinLinear kernel), avoiding the AutoAWQ `awq_ext` extension which has an undefined symbol in the CUDA 12.8 environment used.

- **GPTQ** (Frantar et al., 2023): 4-bit and 8-bit (where Hub weights exist).

- **GGUF** (Gerganov et al., 2023): 4-bit and 2-bit (via llama.cpp-compatible loading through HuggingFace `llama-cpp-python`).

AWQ-8 is included only for Qwen (exploratory); it incurs a 7× throughput penalty via BitsAndBytes INT8 and was excluded from Mistral, Llama, and Gemma to manage GPU cost. GPTQ-8 is excluded for Llama and Gemma (no Hub weights available). The resulting matrix spans 23 model×condition pairs (per-model counts in Table 1) collapsed into 92 conditions across 4 tasks ($\approx$ 500 samples each).

4.3 Tasks and Datasets

Four NLG tasks are selected to isolate distinct primary capability dimensions (Table 2):

Dataset notes. DailyDialog was replaced by PersonaChat (multi-turn human dialogue) due to a loading-script incompatibility with `datasets`≥2.20; both target the same discourse coherence dimension. WebNLG and TurkCorpus were loaded via direct GCS/GitHub downloads, bypassing the `datasets`≥3.0 loading-script restriction.

4.4 Generation Protocol

All generations use greedy decoding (temperature = 0) with `torch.manual_seed(42)`. Only bit-width and quantization method vary across conditions; prompts, datasets, and generation parameters are held constant. Each generation file is produced once and reused across scoring passes. The Mistral TurkCorpus 16-bit condition produced anomalously low ROUGE-1 scores (0.28 vs 0.56–0.63 for compressed variants), consistent with an instruction-following formatting failure at full precision. This condition is retained in the dataset but flagged; its effect on pooled statistics is noted where relevant.

4.5 Statistical Analysis

Degradation metric. For each condition, we compute a normalised degradation score per capability c at bit-width b:

$$\delta(c, b) = \frac{\text{score}_{16}(c) - \text{score}_b(c)}{\text{score}_{16}(c)}$$

For grammar error rate (lower is better), the numerator is reversed so that positive δ always denotes quality loss.

H_0 test. For each model×task×bit-width combination, we apply a Friedman test (Friedman, 1937) across the $k = 6$ capability degradation scores to test whether their ranks are drawn from a uniform distribution (H_0: all capabilities degrade equally). The 32 tests are treated as exploratory per-cell hypothesis tests — each evaluates whether degradation is uniform within one specific (model, task, bit-width) combination — rather than as a single family-wise claim, and no additional correction is applied across cells. Family-wise correction is applied within each cell in the post-hoc Wilcoxon step (Section 5.2).

Effect sizes. Cohen's d is computed from raw score distributions (baseline vs compressed), pooled across model×task combinations at each bit-width. We report effect sizes at the raw-score level rather than the per-sample ratio level to avoid sensitivity to small baseline values in the denominator. Bootstrap 95% confidence intervals ($n = 1000$) are reported for pooled means.

Model	Task	4-bit		2-bit
		χ^2	sig	sig
Qwen 2.5-7B	Summarisation	87.28	***	***
	Dialogue	102.65	***	***
	WebNLG	171.98	***	***
	TurkCorpus	338.64	***	ns
Mistral 7B	Summarisation	38.97	***	ns
	Dialogue	57.07	***	***
	WebNLG	453.26	***	***
	TurkCorpus	416.77	***	***
Llama 3.1-8B	Summarisation	25.35	***	***
	Dialogue	298.61	***	***
	WebNLG	18.37	**	***
	TurkCorpus	30.67	***	***
Gemma-2-9B	Summarisation	7.37	ns	ns
	Dialogue	26.21	***	*
	WebNLG	21.15	***	***
	TurkCorpus	3.78	ns	***
Sig. (p<0.05)			**14/16**	**13/16**

Table 3: Friedman test (H_0: uniform degradation; $df = 5$, six capabilities). χ^2 reported for 4-bit (primary); 2-bit significance coded only. *** p<0.001, ** p<0.01, * p<0.05, ns.

Post-hoc tests. For cells where Friedman rejects H_0, we apply pairwise Wilcoxon signed-rank tests between grammar and AlignScore degradation arrays, with Bonferroni correction for three comparisons per cell.

Regression. We fit an OLS regression with HC3 heteroscedasticity-robust standard errors:

$$\delta \sim \log_2(\text{bw}) + \text{Arch} + \text{Task} + \text{Cap} + \log_2(\text{bw}) \times \text{Cap} + \varepsilon$$

The $\log_2(\text{bw}) \times$ Capability interaction is the primary test for whether capability-specific degradation rates depend on bit-width.

5 Results

5.1 Primary: Rejection of Uniform Degradation

Table 3 reports Friedman test outcomes for all 32 model×task× bit-width combinations. **27 of 32 (84%) reject H_0 at $p < 0.05$**, indicating that quantization does not uniformly degrade all capabilities in the large majority of experimental conditions.

Model	Summ	Dial	Web	Turk	Mean
Qwen 2.5-7B	$+0.2\%$	$\mathbf{-22.5\%}$	$+0.3\%$	$+1.4\%$	-5.2%
Mistral 7B	$+2.2\%$	-3.7%	$+1.3\%$	-0.5%	-0.2%
Llama 3.1-8B	-0.4%	-11.1%	$+0.0\%$	$+3.4\%$	-2.0%
Gemma-2-9B	$+0.5\%$	$+6.3\%$	$+0.4\%$	$+0.7\%$	$+2.0\%$
Mean	$+0.6\%$	-7.7%	$+0.5\%$	$+1.2\%$	$\mathbf{-1.4\%}$

Table 4: AlignScore change at 4-bit AWQ vs 16-bit. Positive = degraded (lower AlignScore); negative = improved. Overall mean $= -1.4\%$; largest improvement $= -22.5\%$ (Qwen dialogue); largest degradation $= +6.3\%$ (Gemma dialogue).

Model	Summ	Dial	Web	Turk	Mean
Qwen 2.5-7B	$+18.8\%$	$+58.5\%$	$+28.0\%$	$+52.8\%$	$+19.0\%$
Mistral 7B	$+40.5\%$	-37.9%	-12.5%	-9.3%	$\approx 0.0\%$
Llama 3.1-8B	-6.3%	-62.7%	-7.5%	-4.2%	-12.8%
Gemma-2-9B	$+0.4\%$	-28.6%	$+5.9\%$	-5.2%	-2.6%
Pooled					$\approx +0.9\%$

Table 5: Grammar error rate change at 4-bit AWQ vs 16-bit. Positive = more errors. Pooled mean ≈ 0 because Qwen degrades and Llama improves. red : degrades; yellow : flat; green : improves.

The five non-significant cases (Qwen TurkCorpus 2-bit; Gemma Summarisation 4-bit and 2-bit; Gemma TurkCorpus 4-bit; Mistral Summarisation 2-bit) are not counterexamples to the thesis: they represent conditions where overall degradation is small enough that differences across capability dimensions cannot be resolved at the available sample sizes ($n = 53$–441).

5.2 Post-Hoc: Grammar vs Factual Consistency

Among the 27 Friedman-significant cells, we apply pairwise Wilcoxon signed-rank tests comparing grammar and AlignScore degradation arrays (Bonferroni correction: $n_{\text{comparisons}} = 3$). **18 of 27 cells show significantly greater grammar degradation than factual degradation after correction** ($p_{\text{Bonf}} < 0.05$). The 9 non-significant cells are concentrated in Llama (where grammar *improves*) and Gemma (where grammar effects are near zero).

5.3 Factual Consistency: Substantially More Stable Than Grammar

Table 4 presents AlignScore changes at 4-bit AWQ relative to 16-bit baseline across all model×task combinations. The dominant pattern is near-zero or negative degradation (i.e., factual consistency is stable or improves under compression).

Pooled across all models and tasks at 4-bit, the AlignScore Cohen's $d = -0.355$ (95% CI: $[-4.2\%, +0.1\%]$ for mean degradation). The near-zero upper bound on the confidence interval is consistent with the absence of meaningful factual degradation at 4-bit, within the four models evaluated.

The largest single deviation is Qwen dialogue (-22.5%, an improvement): compressed variants score substantially higher on entailment than the 16-bit baseline on PersonaChat. Llama dialogue shows the same direction (-11.1%). Both are consistent with 4-bit AWQ producing more conservative, source-hugging responses on the dialogue task; absolute AlignScore values remain within 0.02–0.09 points on the $[0, 1]$ scale. The only cell with notable degradation is Gemma dialogue ($+6.3\%$), which is still small relative to grammar effects in the same conditions.

5.4 Grammar Error Rate: Architecture-Dependent

Table 5 presents grammar error rate changes at 4-bit AWQ. The pattern differs strikingly from factual consistency: the direction of change is architecture-dependent.

The pooled Cohen's $d = +0.049$ (95% CI: $[-6.4\%, +8.0\%]$) is near zero and spans zero. This pooled statistic is not informative on its own: Qwen 2.5-7B shows a mean grammar degradation of $+19.0\%$, while Llama 3.1-8B shows a mean improvement of -12.8%. These opposing effects cancel in the pool. The grammar result must be read per-architecture.

The dialogue task shows the most dramatic pattern: grammar *improves* for 3 of 4 models at 4-bit (Mistral: -37.9%, Llama: -62.7%, Gemma: -28.6%), while Qwen degrades ($+58.5\%$). Section 6.3 investigates this in depth.

5.5 Per-Model Effect Sizes: Grammar vs Factual Consistency

Table 6 reports Cohen's d at 4-bit AWQ for grammar and AlignScore per model×task, computed from raw score distributions. Positive d indicates degradation (compressed model worse); negative d indicates improvement. This table provides the magnitude information that pooled statistics obscure.

The headline case is **Llama 3.1-8B dialogue**: grammar $d = -1.142$ versus AlignScore $d = -0.131$ — a $>10\times$ effect-size gap on the same model and condition. Qwen shows the opposite

Model	Task	Gram d	Align d	R-1 d
Qwen 2.5-7B	Summarisation	$+0.173$	$+0.005$	$+0.486$
	Dialogue	$+0.230$	-0.283	$+0.622$
	WebNLG	$+0.224$	$+0.019$	$+0.438$
	TurkCorpus	$+0.320$	$+0.174$	$+2.730$
Mistral 7B	Summarisation	$+0.167$	$+0.061$	$+0.328$
	Dialogue	-0.118	-0.040	$+0.141$
	WebNLG	-0.116	$+0.109$	$+0.651$
	TurkCorpus	-0.080	-0.041	-1.540
Llama 3.1-8B	Summarisation	-0.064	-0.011	$+0.044$
	Dialogue	$\mathbf{-1.142}$	-0.131	-0.045
	WebNLG	-0.073	$+0.002$	$+0.005$
	TurkCorpus	-0.050	$+0.129$	-0.129
Gemma-2-9B	Summarisation	$+0.004$	$+0.016$	$+0.017$
	Dialogue	-0.041	$+0.096$	$+0.207$
	WebNLG	$+0.046$	$+0.047$	-0.097
	TurkCorpus	-0.027	$+0.071$	$+0.075$

Table 6: Cohen's d at 4-bit AWQ (positive = degradation). Llama dialogue grammar ($d = -1.142$, bold) vs AlignScore ($d = -0.131$): largest capability differential.

pattern: grammar d up to $+0.320$ while AlignScore stays near zero. Gemma grammar d is near zero throughout (-0.041 to $+0.046$).

5.6 Lexical Accuracy

ROUGE-1 shows small to moderate degradation, pooled Cohen's $d = +0.134$ (95% CI: $[-5.9\%, +9.0\%]$), ROUGE-L $d = +0.114$ (95% CI: $[-6.1\%, +8.6\%]$). These effect sizes are positive but their CIs span zero, reflecting high variance across model×task cells.

Qwen 2.5-7B is the strongest degrader lexically: ROUGE-1 crosses the 10% degradation threshold at 4-bit (both AWQ and GPTQ). The Qwen TurkCorpus ROUGE-1 $d = +2.730$ is the largest single Cohen's d in the entire study; manual inspection suggests this is driven by Qwen 16-bit producing unusually reference-faithful simplifications on TurkCorpus, while compressed variants generate shorter paraphrases that diverge from the reference vocabulary. This mirrors the Mistral TurkCorpus pattern in direction, though here it is the 16-bit model that over-performs (not under-performs) relative to compressed variants.

The Mistral TurkCorpus ROUGE-1 $d = -1.540$ is a negative outlier driven by the instruction-following anomaly at 16-bit precision (Section 6.3), not genuine compression improvement. Llama, Gemma, and Mistral do not exceed the 10% threshold at 4-bit for either ROUGE metric (Table 7).

Model	R-1	R-L	Grammar	AlignScore	Coherence	Reasoning
Qwen 2.5-7B	4-bit	4-bit	2-bit†	never	never	never
Mistral 7B	never	never	8-bit	never	never	never
Llama 3.1-8B	never	never	never‡	never	2-bit	2-bit
Gemma-2-9B	never	never	2-bit	never	never	never

Table 7: Precision thresholds: lowest bit-width exceeding 10% degradation. †Qwen grammar degrades; ‡Llama grammar improves (threshold never crossed). Coherence and Reasoning: GPT-4o-mini judge scores.

5.7 Capability Precision Thresholds

Table 7 shows the lowest bit-width at which each model's mean degradation first exceeds 10% relative to 16-bit. The threshold is "never" when the model does not cross the 10% mark at any bit-width tested (down to 2-bit).

The most practically significant finding in this table is **Mistral's grammar threshold at 8-bit**: Mistral 7B's grammar degrades meaningfully at $\geq$ 8-bit compression on summarisation ($+40.5\%$), even though 8-bit is widely considered "nearly lossless." This contradicts the assumption that 8-bit is a safe floor for all capabilities. For the LLM-judge dimensions, coherence ($d = +0.311$) and reasoning ($d = +0.343$) show moderate 4-bit effect sizes with zero-spanning CIs; only Llama 3.1-8B crosses the 10% threshold, and only at 2-bit — consistent with the automatic-metric pattern.

5.8 Regression: Global Model

The OLS regression ($R^2 = 0.098$, $F_p < 0.001$, $N = 432$) is globally significant, confirming that bit-width, capability type, architecture, and task are all relevant predictors of degradation. With six capability dimensions, two of five $\log_2(\text{bw}) \times$ Capability interaction coefficients reach significance: discourse coherence ($\beta = -0.062$, $p = 0.025$) and instruction-following quality ($\beta = -0.071$, $p = 0.019$). The automatic-metric interactions remain non-significant (grammar: $\beta = -0.086$, $p = 0.187$; ROUGE-1: $\beta = -0.005$, $p = 0.951$; ROUGE-L: $\beta = +0.008$, $p = 0.911$).

The null interaction for automatic metrics is not contradicted by the Friedman results: Friedman operates *within* each (model, task, bit-width) cell; OLS *pools* across cells. When Qwen grammar degrades strongly and Llama grammar improves strongly, the grammar×bit-width interaction averages toward zero — the cancellation is itself the finding. The LLM-judge interactions are significant precisely because coherence and reasoning degrade more uniformly in direction across models,

Model	ΔLen	ΔGram	r	p
Qwen 2.5-7B	$+48.1\%$	$+58.5\%$	-0.023	0.604
Mistral 7B	$+34.4\%$	-37.9%	$+0.024$	0.593
Llama 3.1-8B	-12.1%	-62.7%	$+0.121$	0.007
Gemma-2-9B	-30.5%	-28.6%	$+0.049$	0.278

Table 8: Per-sample correlation: length change vs grammar rate change (dialogue, 4-bit AWQ). Near zero for all models except Llama ($r = 0.121$).

Model	Task	AWQ4	GPTQ4	GGUF4
Llama 3.1-8B	Dialogue	-62.7%	-57.7%	-10.3%
Qwen 2.5-7B	TurkCorpus	$+52.9\%$	$+24.4\%$	$+55.5\%$
Mistral 7B	Summarisation	$+40.5\%$	$+33.9\%$	$+34.6\%$
Qwen 2.5-7B	Dialogue	$+58.5\%$	-26.5%	-10.9%

Table 9: Grammar change across methods. Top: large-effect cells replicate. Bottom: Qwen dialogue is the only large-effect reversal (AWQ-specific).

allowing the pooled model to recover them.

6 Analysis

6.1 Length Confound: Grammar Rate Already Normalised

The grammar improvement finding for Llama and Gemma on dialogue raises a potential confound: if compressed models produce shorter outputs, they may simply accumulate fewer raw LanguageTool errors, and a lower error count could appear as "improvement" even if error *density* is unchanged.

This confound is mitigated at the metric level because we use LanguageTool's grammar error rate normalised per 100 tokens — not raw error count. To verify this normalisation is sufficient, we compute the Pearson correlation between per-sample output length change and per-sample grammar rate change for each model on dialogue (Table 8).

Three of four models show near-zero correlations ($p > 0.05$). Mistral falsifies the length hypothesis directly: output length *increases* ($+34.4\%$) while grammar rate *decreases* (-37.9%), with a near-zero sample-level correlation. Llama shows a weak positive correlation ($r = 0.121$, $p = 0.007$), accounting for $<2\%$ of variance; Llama's grammar improves by -62.7% against a -12.1% length change — the magnitude disparity and non-covariance at the sample level rule out length as the driver.

6.2 Cross-Method Consistency of Grammar Direction

To assess whether grammar direction findings are AWQ-specific, we compare the sign of grammar rate change across AWQ4, GPTQ4, and GGUF4. Direction is consistent for 8/16 model×task cells; the key observation is that consistency depends on effect size, not method. Table 9 shows the four most informative cells; the full 16-cell table with consistency flags is in Appendix A.

All three methods agree in direction for the three large-effect cells: Llama dialogue (-62.7%, -57.7%, -10.3%), Qwen TurkCorpus ($+52.9\%$, $+24.4\%$, $+55.5\%$), and Mistral summarisation ($+40.5\%$, $+33.9\%$, $+34.6\%$). Near-zero cells (Gemma) do not replicate direction consistently — sign differences at that scale reflect noise. The sole large-effect exception is Qwen dialogue: AWQ4 shows degradation ($+58.5\%$) while GPTQ4 and GGUF4 show improvement; claims about Qwen dialogue grammar should be treated as potentially AWQ-specific.

6.3 Grammar Directionality: Architecture and Task Effects

The grammar result is the most architecturally differentiated finding. Qwen 2.5-7B consistently degrades across three of four tasks; Llama 3.1-8B consistently improves; Gemma-2-9B stays near zero. Two candidate explanations exist — baseline verbosity differences (compression shifts models toward shorter, less LanguageTool-flagged outputs) and architectural weight sensitivity (different attention mechanisms may place quality-critical weights in activation ranges that AWQ's per-channel scale handles differently) — but distinguishing them requires weight-level analysis outside our current scope.

The most consistent task-specific pattern is grammar improvement on dialogue (3/4 models at 4-bit AWQ). The length confound analysis (Table 8) confirms this is not a normalisation artefact. The more plausible explanation is a distributional style shift: 16-bit models on PersonaChat generate multi-turn, stylistically varied dialogue that LanguageTool flags heavily (contractions, ellipsis, discourse markers); 4-bit variants produce shorter, more formulaic responses. Whether this constitutes a genuine syntactic improvement or a metric interaction with generation style cannot be determined from automatic scores alone, and we flag it as a limitation.

Model-specific anomalies. Qwen 2.5-7B is the only model with consistent grammar degradation across all four tasks and the only model whose ROUGE threshold is crossed at 4-bit; Qwen's grouped-query attention and RoPE variant may place quality-critical weights in activation ranges that AWQ handles less effectively, though this remains a hypothesis. Mistral 7B at 16-bit achieves ROUGE-1 $= 0.2823$ on TurkCorpus versus 0.56–0.63 for compressed variants: this is an instruction-following failure at full precision (the 16-bit model paraphrases rather than simplifies), not a compression benefit. This condition is retained but excluded from threshold calculations.

7 Discussion

Sections 5–6 establish *that* capabilities degrade differentially and *that* the pattern is architecture- and task-dependent, but not *why*. We offer two candidate mechanistic accounts; both remain hypotheses, since testing them would require weight- or activation-level analysis outside this study's scope.

Why factual consistency is comparatively robust. AlignScore reduces to an entailment judgement between a generated span and a source document already present in the context window: a single anchored comparison. Grammar error rate, by contrast, integrates local dependency correctness over *every* token transition in the output, so a single corrupted activation anywhere in a long generation is enough to register as a new error, while the same perturbation is far less likely to flip an entailment judgement grounded in surrounding, unperturbed source tokens. This is consistent with the depth gradient underlying our capability selection (§3): factual consistency is grounded in present context, whereas grammatical well-formedness must be sustained across the full generation trajectory. It does not, however, explain every cell — Gemma dialogue shows the one meaningful AlignScore degradation ($+6.3\%$, Table 4) alongside a near-zero grammar effect, the reverse of the general pattern — so this account is necessarily partial.

Why grammar direction is architecture-dependent. Qwen 2.5-7B degrades on grammar in three of four tasks while Llama 3.1-8B improves on three of four (Table 5), and Mistral 7B is the only model whose grammar crosses the 10% threshold at 8-bit, not just 4-bit (Table 7) — contradicting the common assumption that 8-bit is a safe floor across the board. One candidate explanation is that different attention mechanisms and positional encodings (Qwen's grouped-query attention and RoPE variant vs. Llama's) place quality-critical weights in different activation ranges, which AWQ's per-channel scaling handles with different fidelity; a second, non-exclusive explanation is a style-shift artefact, where 16-bit baselines on dialogue generate longer, more stylistically varied text that LanguageTool flags more heavily, and compression shifts generation toward shorter, more templated output (§6.1 rules out raw length as the driver, but not a style change at fixed length). Distinguishing these requires either weight-level ablation or a human/style-controlled evaluation of the dialogue outputs directly, neither of which we perform here.

A shared account. Both observations point toward the same underlying claim: the more a capability's automatic proxy depends on precise activations sustained over a long generation trajectory, rather than a single grounded comparison against present context, the more exposed it is to compression noise, and the more its response to compression varies by architecture. This reframes the paper's central finding from a purely empirical one (profiles are non-uniform) to a testable structural claim about *why* they are non-uniform — one that a follow-up study with layer-wise ablation could directly evaluate.

8 Practical Recommendations

Table 10 summarises the key deployment implications within the scope of this study. No single bit-width recommendation generalises across all capabilities and architectures; factual consistency is the most compression-tolerant dimension, while grammar requires per-architecture decisions and 8-bit is not lossless for Mistral on summarisation.

9 Conclusion

We presented a systematic empirical evaluation of capability-level degradation under quantization across four instruction-tuned LLMs, four NLG tasks, and five compression conditions, rejecting the null hypothesis of uniform degradation in 84% of experimental conditions. Three findings stand out. First, **factual consistency is substantially more stable than grammar**: AlignScore shows near-zero degradation across all four models ($d = -0.355$, mean $= -1.8\%$), while

Capability	Bit-width	Notes
Factual consist.	4-bit, all models	Near-zero degradation; no model crosses 10%.
Lexical accuracy	4-bit (Llama, Mistral, Gemma); caution for Qwen	Qwen exceeds 10% at 4-bit; others below threshold.
Grammar (summ.)	**Caution at 8-bit for Mistral**	Mistral grammar $+40.5\%$ at 8-bit AWQ, replicated GPTQ ($+33.9\%$) and GGUF ($+34.6\%$). 8-bit is *not* a safe floor for grammar on summarisation.
Grammar (general)	*Architecture-dependent*	Qwen: degrades from 4-bit. Llama/Gemma: *improve* on dialogue (metric interaction with generation style unresolved).

Table 10: Deployment guidance (four 7–9B models, four tasks). Results may not generalise to other sizes or architectures.

grammar degradation is both larger in magnitude and architecture-dependent — Qwen 2.5-7B degrades ($+19.0\%$) while Llama 3.1-8B improves (-12.8%, $d = -1.142$ on dialogue vs. AlignScore $d = -0.131$ on the same condition). Second, **no capability precision threshold is shared across models**: Qwen ROUGE-1 degrades at 4-bit; Mistral grammar degrades at 8-bit; Llama and Gemma cross no threshold. Third, **the grammar improvement for Llama/Gemma on dialogue is not a length artefact** (per-sample correlations near zero; Mistral falsifies the hypothesis directly), but cannot be ruled out as a metric interaction with generation style.

Future work. A longer-term research direction is *capability-aware precision allocation*: using capability-specific precision thresholds from this study to assign bit-widths per-layer based on estimated capability demand during decoding.

Limitations

The grammar error rate metric interacts with output length and generation style, creating confounds in dialogue settings that length normalisation partially but not fully addresses. The AlignScore proxy uses the same NLI methodology as the original package via `cross-encoder/nli-deberta-v3-base`. LLM-as-judge scores are subject to GPT-4o-mini model bias and compress fine-grained variation onto a 1–5 integer scale. All models are in the 7–9B parameter range; findings may not generalise to smaller or larger models, or to model families not represented in this study.

This study relies entirely on automatic metrics and an LLM judge; no human evaluation was conducted. Marchisio et al. (2024) report that automatic metrics can substantially underestimate quantization damage relative to human judgement on realistic prompts (a 1.7% average automatic-metric drop vs. a 16.0% human-rated drop for Japanese). If a comparable gap holds here, the effect sizes we report — particularly for the judge-scored coherence and instruction-following dimensions, and for AlignScore's directional (source-to-output only) notion of consistency (§3) — should be read as a lower bound on perceptible degradation rather than its full extent. A targeted human evaluation on the largest-effect cells identified here (Table 6) is the most direct way to close this gap and is left to future work.

Data Availability

Analysis code, per-sample metric scores (92 conditions), and the quantization and generation pipeline are available at `https://github.com/motornomad/capability-degradation-quantization`. Raw model generation outputs (≈500 samples × 92 conditions) are withheld from this release due to size but are available on request.

References

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*.

Milton Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701.

Yao Fu, Xianxuan Long, Runchao Li, Haotian Yu, Mu Sheng, Xiaotian Han, Yu Yin, and Pan Li. 2025. Quantized but deceptive? A multi-dimensional truthfulness evaluation of quantized LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. ArXiv:2508.19432.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from DBPedia data. In *Proceedings of the 10th International Conference on Natural Language Generation (INLG)*.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Georgi Gerganov et al. 2023. llama.cpp: Inference of LLaMA model in pure C/C++.

Renren Jin, Jiangcun Du, Wuwei Huang, Wei Liu, Jian Luan, Bin Wang, and Deyi Xiong. 2024. A comprehensive evaluation of quantization strategies for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*. ArXiv:2402.16775.

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)*, pages 8113–8121. ArXiv:2409.11055.

Shiyao Li, Xuefei Ning, Luning Wang, Tengxuan Liu, Xu Shi, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. 2024. Evaluating quantized large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 28480–28524. ArXiv:2402.18158.

Zhen Li, Yupeng Su, Runming Yang, Congkai Xie, Zheng Wang, Zhongwei Xie, Ngai Wong, and Hongxia Yang. 2025. Quantization meets reasoning: Exploring LLM low-bit quantization degradation for mathematical reasoning. *arXiv preprint arXiv:2501.03035*.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. AWQ: Activation-aware weight quantization for llm compression and acceleration. In *Proceedings of Machine Learning and Systems (MLSys)*.

Kelly Marchisio, Saurabh Dash, Hongyu Chen, Dennis Aumiller, Ahmet Üstün, Sara Hooker, and Sebastian Ruder. 2024. How does quantization affect multilingual LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15928–15947.

Anmol Mekala, Anirudh Atmakuru, Yixiao Song, Marzena Karpinska, and Mohit Iyyer. 2025. Does quantization affect models' performance on long-context tasks? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. ArXiv:2505.20276.

Daniel Naber. 2003. LanguageTool: An open-source grammar checker. https://languagetool.org.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of ACL 2017*.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Zhichao Xu, Ashim Gupta, Tao Li, Oliver Bentham, and Vivek Srikumar. 2024. Beyond perplexity: Multi-dimensional safety evaluation of LLM compression. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. ArXiv:2407.04965.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of ACL 2018*.

A Cross-Method Grammar Consistency: Full 16-Cell Table

Table 11 reports grammar error rate change (%) at 4-bit for all 16 model×task combinations across AWQ4, GPTQ4, and GGUF4. "Consistent" means all three methods agree on direction. The 8 consistent cells are those where a genuine directional effect exists regardless of method; the 8 inconsistent cells are concentrated in Gemma (near-zero absolute changes) and Qwen/Mistral/Llama cells where the AWQ direction does not replicate.

Model	Task	AWQ4	GPTQ4	GGUF4	Consistent?
Qwen 2.5-7B	Summarisation	+18.8%	+1.3%	+23.6%	Yes
	Dialogue	+58.5%	−26.5%	−10.9%	**No**
	WebNLG	+28.1%	+2.0%	+0.2%	Yes
	TurkCorpus	+52.9%	+24.4%	+55.5%	Yes
Mistral 7B	Summarisation	+40.5%	+33.9%	+34.6%	Yes
	Dialogue	−37.9%	−34.0%	+2.4%	**No**
	WebNLG	−12.5%	−17.2%	−5.0%	Yes
	TurkCorpus	−9.3%	−1.3%	+6.3%	**No**
Llama 3.1-8B	Summarisation	−6.3%	+1.0%	+11.8%	**No**
	Dialogue	−62.7%	−57.7%	−10.3%	Yes
	WebNLG	−7.5%	−1.2%	−7.5%	Yes
	TurkCorpus	−4.2%	−8.4%	−1.2%	Yes
Gemma-2-9B	Summarisation	+0.4%	−0.0%	+7.8%	**No**
	Dialogue	−28.6%	+29.7%	−29.4%	**No**
	WebNLG	+5.9%	−1.6%	+3.2%	**No**
	TurkCorpus	−5.2%	−15.2%	+1.1%	**No**
Direction consistent					**8/16**

Table 11: Grammar error rate change across three 4-bit quantization methods for all 16 model×task cells. "Consistent" = all three methods agree on direction (+ or −). Bold "No" = direction reverses across methods. All Gemma cells are inconsistent, reflecting near-zero absolute changes where sign classification reflects noise.