Interaction Effects in Hybrid Compression of Small Language Models

Iheb Bouriel[1] , **Qassim Nasir**[2] , **Manar Abu Talib**[2]
[1]Technical University of Munich, Germany
[2]University of Sharjah, UAE
iheb.bouriel@tum.de, nasir@sharjah.ac.ae, mtalib@sharjah.ac.ae

Abstract

We investigate the interaction between quantization and pruning in the compression of open-weight small language models (SLMs), as these techniques are frequently combined in practice without a clear understanding of whether their effects are antagonistic, synergistic or additive. Prior work has reported ordering effects and studied pruning or quantization in isolation, but rarely quantifies their interaction across models and evaluation metrics. We introduce an interaction coefficient that isolates non-additive effects and apply it to Falcon3-1B-Base and LLaMA-3.2-1B, with a limited study on Qwen2.5-1.5B. Our experiments cover four hybrid application orders, multiple pruning sparsities and 4/8-bit quantization. We find non-monotonic interactions: memory savings are dominated by 4-bit quantization, while pruning alone provides limited benefits. Hybrids exhibit the greatest effects at global unstructured pruning sparsity $s \in \{0.2, 0.3\}$ while getting close to additivity at $s = 0.5$. At low sparsity, truthfulness improves, while reasoning losses diminish at higher sparsity. Future work will extend interaction analysis to hardware-aware sparse kernels and larger models.

1 Introduction

Open-weight SLMs in the 1–1.5B parameter range are increasingly suitable for on-device assistants and retrieval-augmented systems that must operate within single 8–12 GB accelerators [Zhou *et al.*, 2025; Shinde, 2024; Girija *et al.*, 2025]. However, even these models can exceed memory and latency budgets when combined with retrieval, safety and multi-tenant serving layers [Zhou *et al.*, 2025; Kim *et al.*, 2025]. Magnitude pruning and post-training quantization are standard tools for reducing model footprint and inference cost [Frantar and Alistarh, 2023; Elias and Dan, 2023]. Recent work has explored hybrid pipelines that combine pruning and quantization, motivated by hardware sparsity support and potentially complementary effects [Kang *et al.*, 2025; Zhou *et al.*, 2025; Shinde, 2024]. These studies often report that hybrids outperform single-technique baselines in efficiency, yet they generally present aggregate results without distinguishing whether the techniques reinforce or interfere with each other [Kang *et al.*, 2025; Zhou *et al.*, 2025].

A central open question is whether pruning and quantization interact additively, synergistically or antagonistically across models, tasks and deployment metrics. Most prior work implicitly assumes near-additivity, where the combined effect reflects the sum of individual interventions. However, SLMs have limited redundancy and structured perturbations can redistribute capacity in unpredictable ways [Frantar and Alistarh, 2023; Zhou *et al.*, 2025; Ma *et al.*, 2023]. Consequently, stacking these techniques may produce synergy or antagonism that cannot be predicted from standalone performance curves [Zhou *et al.*, 2025]. We fill this gap by adapting a 2×2 interaction term from factorial design into an interaction coefficient computed from four standard runs—baseline, pruned, quantized and hybrid [Zhou *et al.*, 2025]. We apply this approach to 1–1.5B models across pruning levels, 4/8-bit quantization and multiple hybrid orderings under single-GPU constraints. We investigate the following research questions:

RQ1 How do pruning-quantization interactions affect the nine harness benchmarks (reasoning, knowledge, commonsense and inference, and truthfulness; Section 3.2)?

RQ2 How do these interactions influence model size, peak VRAM, latency and throughput across architectures (Falcon vs. LLaMA) and sparsity levels?

The rest of the paper is organized as follows: Section 2 reviews prior work on pruning, quantization and hybrids; Section 3 details our experimental setup and interaction coefficient; Section 4 reports results, including a Qwen2.5-1.5B supplement; Section 5 discusses implications; Section 6 concludes.

2 Related Work

We survey prior work on pruning and quantization for small and large language models, then discuss hybrid pipelines and how interactions have been analyzed. This context clarifies which findings we build upon and where SLM settings remain underexplored.

2.1 Compression of Small and Large Language Models

Pruning and quantization have long been used to compress deep networks [Han *et al.*, 2016] and have recently been

adapted to modern LLMs. GPTQ and its variants [Elias and Dan, 2023; Lin *et al.*, 2025; Dettmers *et al.*, 2023] demonstrate that post-training weight quantization to 4-8 bits can preserve performance for 7B+ LLMs across many tasks. Pruning methods tailored to transformers, such as SparseGPT [Frantar and Alistarh, 2023] and LLM-Pruner [Ma *et al.*, 2023], exploit magnitude and Hessian approximations to remove large fractions of weights with minimal loss. Most of this work targets large models where redundancy is abundant and results are sometimes extrapolated to smaller scales. However, recent studies [Zhou *et al.*, 2025; Shinde, 2024; Ma *et al.*, 2023] emphasize that small language models are more fragile: moderate pruning can induce sharp accuracy drops, while quantization tends to be more forgiving.

2.2 Hybrid and Joint Compression Pipelines

Hybrid compression, combining sparsity and low precision, has attracted significant attention. HWPQ [Kang *et al.*, 2025] proposes a Hessian-free scheme that jointly optimizes pruning and quantization to maximize hardware efficiency. Other work explores mixed-precision policies and adaptive sparsity [Shinde, 2024; Kim *et al.*, 2025], allocating bits or sparsity levels per layer to match sensitivity profiles and hardware constraints. While many studies report that hybrid schemes outperform isolated compression, they typically present aggregate results without explicit interaction decomposition [Kang *et al.*, 2025; Zhou *et al.*, 2025; Kim *et al.*, 2025]. Efficiency metrics are rarely analyzed through a Pareto lens.

2.3 Interaction Analysis and Metricization

Our work is closest in spirit to recent analyses that treat hybrid compression as a factorial intervention and examine whether its outcome deviates from the sum of individual effects [Qu *et al.*, 2025; Zhou *et al.*, 2025]. However, these studies either focus on individual metrics or LLMs or treat interaction analysis as an auxiliary diagnostic rather than a primary object of study. To our knowledge, none introduce a simple reusable statistic that quantifies interaction simultaneously for accuracy and deployment costs across architectures and sparsity levels in the SLM regime. We formalize such a statistic and show that its sign structure yields clear, interpretable patterns across tasks and models. We make four main **contributions**:
(1) We adapt a standard interaction term from factorial design into an interaction coefficient I for both accuracy and efficiency metrics, quantifying whether pruning and quantization combine additively, synergistically ($I > 0$) or antagonistically ($I < 0$). (2) Through a detailed study on Falcon3-1B-Base and LLaMA-3.2-1B, we reveal that interaction effects are *non-monotonic in sparsity* and show consistent patterns across both architectures: interactions are strongest at moderate sparsity ($s = 0.2, 0.3$) and move toward zero (near-additive behavior) at extreme sparsity ($s = 0.5$). This non-monotonicity contradicts the assumption that compression effects stack monotonically and demonstrates that optimal hybrid strategies must account for sparsity-dependent interaction behavior. (3) We jointly analyze accuracy against model size, peak VRAM and throughput, demonstrating that 4-bit post-training quantization is the main driver of memory savings, while magnitude pruning alone is usually dominated. (4) We release an open-source toolkit that automates compression, metric collection, interaction computation and visualization, enabling reproducible experiments and serving as a testbed for future hybrid designs.[1]

Preliminary experiments on Qwen2.5-1.5B, limited by GPU time, corroborate the qualitative findings but reveal even sharper antagonism on reasoning and knowledge tasks (Section 4.5). Compared to prior work that reports aggregate hybrid outcomes without isolating interaction structure [Kang *et al.*, 2025; Zhou *et al.*, 2025; Kim *et al.*, 2025], our study provides a reusable interaction statistic, spans multiple sparsity levels and hybrid orderings and links interaction signs to both accuracy and deployment metrics in the SLM regime.

3 Methodology

We outline the experimental setup, models, benchmarks and compression pipelines, then define the interaction coefficient used throughout our study.

3.1 Experimental Setup

All experiments run on a single NVIDIA RTX 3080Ti GPU (12 GB) under a consistent hardware and software stack. We evaluate with `lm-evaluation-harness` [Gao *et al.*, 2023] using identical decoding settings across all models and pipelines; task metrics use bootstrap resampling with `bootstrap_iters=400` as implemented in the harness. In stored eval logs, bootstrap standard errors for each task are consistently small (typically under 3 percentage points on our subsamples); the main tables report point estimates for readability. Batch size is one and latency is averaged over five trials per benchmark per configuration. We initially tested LLaMA-3.2-1B with three seeds and observed identical scores, so Falcon3-1B-Base hybrid configurations use a single seed due to time and GPU resource constraints. Compression and evaluation use a Python pipeline that, given a HuggingFace identifier, applies quantization and global unstructured pruning, enforces a VRAM budget, saves compressed checkpoints with metadata and evaluates all configurations for interaction analysis and plotting.

3.2 Models and Benchmarks

We study small open-weight language models: **Falcon3-1B-Base** [Technology Innovation Institute (TII), 2025], **LLaMA-3.2-1B** [Meta AI, 2024] and **Qwen2.5-1.5B** [Yang and others, 2024] (evaluated on a subset as a supplement in Section 4.5). Models load via `transformers` with `trust_remote_code=True` when needed. **Rationale.** We include one checkpoint per mainstream open family (TII Falcon, Meta LLaMA, Alibaba Qwen) in the 1–1.5B range typical for single GPU deployment, so interaction sign patterns are tested across architecture lineages rather than a single model series. Benchmarks span **reasoning** (GSM8K [Cobbe *et al.*, 2021]), **knowledge** (MMLU [Hendrycks *et al.*, 2021], ARC-Challenge [Clark *et al.*, 2018], WINOGRANDE [Sakaguchi *et*

[1]Code and configurations: https://github.com/ihebbrl/slm-hybrid-compression.

al., 2021]), **commonsense and inference** (HellaSwag [Zellers *et al.*, 2019], PIQA [Bisk *et al.*, 2020], BoolQ [Clark *et al.*, 2019], XQuAD English subset [Artetxe *et al.*, 2020]) and **truthfulness** (TruthfulQA$_{\text{MC2}}$ [Lin *et al.*, 2022]). Together they cover nine standard harness tasks spanning math, multi-task knowledge, commonsense QA, reading comprehension and truthfulness-oriented multiple choice.

3.3 Compression Pipelines

We examine four hybrids and three base pipelines under various sparsity levels:

- **Baseline (no compression):** Full-precision (FP16) model, as provided by the original checkpoints.
- **Pruned(s):** Global unstructured magnitude pruning with sparsity $s \in \{0.2, 0.3, 0.5\}$ over all linear layers, implemented with `torch.nn.utils.prune.L1Unstructured` followed by mask removal.
- **Quantized4 / Quantized8:** Post-training weight quantization to 4-bit or 8-bit using `bitsandbytes` [Dettmers *et al.*, 2023]. Activations remain in FP16.
- **Q4P(s) and PQ4(s):** 4-bit hybrids with sparsity s. Q4P first quantizes to 4-bit and then prunes; PQ4 reverses the order.
- **Q8P(s) and PQ8(s):** 8-bit hybrids with sparsity s, defined analogously.

Unless otherwise stated, all main-text tables and figures use the primary sparsity $s = 0.2$ (20% global pruning), which delivered the best TruthfulQA improvements while keeping other benchmarks within a measurable range. Section 4.3 analyzes $s \in \{0.3, 0.5\}$. Crucially, pruning is unstructured and implemented via masking and subsequent mask removal only; we do not exploit hardware-supported sparse kernels or compressed storage. As a result, pruning alone leaves model size and peak VRAM essentially unchanged on our hardware, an intentional design choice that isolates the logical effect of sparsity from hardware-specific sparse implementations.

3.4 Interaction Coefficient

Let M be any scalar metric of interest (e.g. accuracy on GSM8K, average latency, throughput, peak VRAM). We denote the metric under four configurations as

$$M_{\text{B}}, M_{\text{P}}, M_{\text{Q}}, M_{\text{H}}$$

corresponding to Baseline, Pruned(s), Quantized and Hybrid (e.g. Q4P(s)). We define the *interaction coefficient*

$$I = M_{\text{H}} - M_{\text{P}} - M_{\text{Q}} + M_{\text{B}} \tag{1}$$

If pruning and quantization effects were strictly additive, we would expect $M_{\text{H}} \approx M_{\text{P}} + M_{\text{Q}} - M_{\text{B}}$, implying $I \approx 0$. Deviations indicate non-additivity: $I > 0$ indicates **synergy** (hybrid outperforms additive expectation), $I < 0$ indicates **antagonism** (hybrid underperforms additive expectation) and $I \approx 0$ indicates approximately additive behavior. We compute I for every combination of model, benchmark, sparsity level s and hybrid ordering (Q4P, PQ4, Q8P, PQ8) and for both accuracy and efficiency metrics, allowing us to disentangle interaction effects from simple "more sparsity is worse" trends.

3.5 Evaluation Metrics

For each model and pipeline we report **task metrics** (GSM8K, MMLU, TruthfulQA$_{\text{MC2}}$, HellaSwag, ARC-Challenge, WinoGrande, PIQA, BoolQ, XQuAD English) and **efficiency metrics** (model size on disk in MB, average per-sample latency in seconds, throughput in samples/s and peak VRAM in MB). All numbers are obtained from a unified logging pipeline and aggregated into per-model CSV and JSON files for analysis and plotting.

4 Results

We begin with task accuracy, then examine interaction patterns and their dependence on sparsity and finally assess efficiency metrics to understand deployment trade-offs.

4.1 Accuracy Across Pipelines

We first examine how compression choices affect benchmark accuracy. We compare baseline, pruned, quantized (4/8-bit) and hybrid pipelines at sparsity $s = 0.2$ for both Falcon3-1B-Base and LLaMA-3.2-1B, focusing on GSM8K, TruthfulQA$_{\text{MC2}}$ and MMLU. **Table focus.** We foreground these three metrics because they illustrate the main interaction regimes at $s = 0.2$: strong antagonism on GSM8K, moderate antagonism on MMLU and synergy on TruthfulQA$_{\text{MC2}}$; every pipeline is evaluated on all nine tasks, with full-suite interactions in Figures 3–4. Table 1 summarizes the scores drawn from the aggregated CSVs.

For both architectures, 4-bit post-training quantization preserves most of the baseline accuracy. On Falcon, MMLU drops slightly from 43.8% to 42.2% and GSM8K from 32.3% to 28.7%; on LLaMA, MMLU decreases from 39.5% to 36.8% while TruthfulQA slightly increases. 8-bit quantization is essentially lossless and sometimes slightly improves accuracy.

Pruning consistently harms accuracy: Falcon's GSM8K falls from 32.3% to 20.0% and LLaMA's MMLU from 39.5% to 33.2%. Importantly, pruning does not reduce VRAM or model size in our implementation (Section 4.4), making it rarely attractive in isolation.

At moderate sparsity, hybrid pipelines (Q4P, PQ4) substantially increase TruthfulQA accuracy for both models, from roughly 42-38% to 49% on Falcon and 48-49% on LLaMA, while operating at less than half of the baseline VRAM. However, they severely reduce GSM8K (to 0% exact match) and MMLU (Falcon: 22.8%; LLaMA: 24.8-26.6%), indicating that the same hybrid that regularizes truthfulness-style evaluation markedly damages reasoning and knowledge capabilities at this sparsity level.

When we include 8-bit hybrids (Q8P, PQ8), they interpolate between the 4-bit hybrids and the standalone 8-bit quantized models: TruthfulQA gains remain substantial but somewhat smaller than with Q4P/PQ4, while GSM8K and MMLU still exhibit large drops relative to the additive expectation. This indicates that the phenomenon is not tied to extreme 4-bit quantization, but to the combination of low-precision weights with global pruning. Section 4.3 analyzes how these interaction patterns vary with sparsity. Across the full benchmark suite for Falcon3-1B-Base and LLaMA-3.2-1B (HellaSwag, ARC-Challenge, WinoGrande, XQuAD, BoolQ and PIQA),

Pipeline	Falcon3-1B-Base			LLaMA-3.2-1B		
	GSM8K	$\text{TruthfulQA}_{\text{MC2}}$	MMLU	GSM8K	$\text{TruthfulQA}_{\text{MC2}}$	MMLU
Baseline	32.3	42.4	43.8	5.0	37.6	39.5
Pruned	20.0	42.7	40.5	3.3	35.7	33.2
Quantized4	28.7	41.2	42.2	5.0	38.1	36.8
Quantized8	32.3	42.4	43.2	7.3	37.0	39.0
Q4P	0.0	48.7	22.8	0.0	48.6	24.8
PQ4	0.0	48.9	22.8	0.0	49.2	26.6
Q8P	0.0	49.5	25.3	0.0	47.4	24.4
PQ8	0.0	49.9	24.4	0.0	48.7	24.7

Table 1: Main accuracy metrics (%) for Falcon3-1B-Base and LLaMA-3.2-1B across compression pipelines (primary sparsity $s = 0.2$). Numbers are rounded for readability.

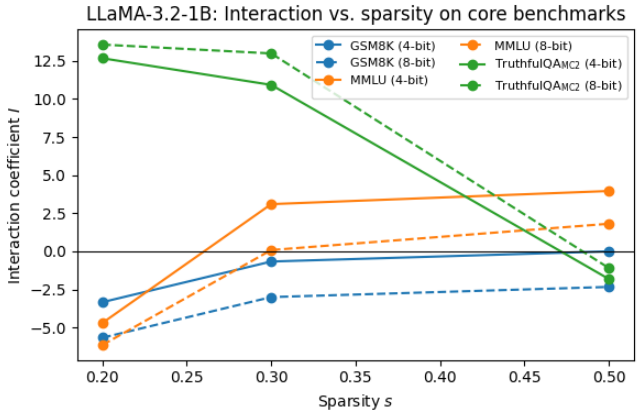

Figure 1: Interaction coefficients I for LLaMA-3.2-1B as a function of sparsity s on three core benchmarks. Solid lines show 4-bit hybrids (Q4P/PQ4 averaged) and dashed lines show 8-bit hybrids (Q8P/PQ8 averaged). TruthfulQA exhibits strong positive interactions at low sparsity ($s = 0.2$) that weaken at $s = 0.5$, while GSM8K and MMLU show negative interactions at low sparsity that move toward zero (near-additive) at high sparsity, illustrating the non-monotonic, task-dependent nature of hybrid compression interactions.

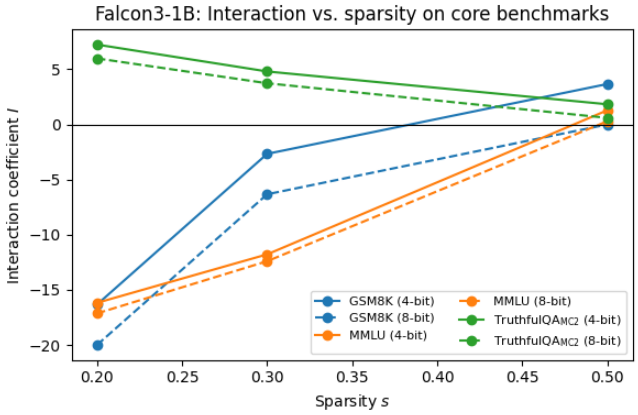

Figure 2: Interaction coefficients I for Falcon3-1B-Base as a function of sparsity s on three core benchmarks. Solid lines show 4-bit hybrids (Q4P/PQ4 averaged) and dashed lines show 8-bit hybrids (Q8P/PQ8 averaged). Falcon exhibits strong negative interactions across all tasks at low sparsity ($s = 0.2$), with GSM8K showing particularly severe antagonism ($I \approx -20$). At high sparsity ($s = 0.5$), interactions move toward zero or become mildly positive, indicating that Falcon's compression sensitivity is highest at moderate pruning levels rather than extreme ones.

we observe that 4-bit and 8-bit hybrids induce large additional losses on HellaSwag and substantial drops on ARC-Challenge and WinoGrande relative to the additive expectation, while XQuAD, BoolQ and PIQA degrade more moderately. These trends align closely with the interaction coefficients discussed next and indicate that antagonism under hybrids is not limited to GSM8K and MMLU.

4.2 Interaction Effects

We next interpret the interaction coefficients to see how pruning and quantization combine. We begin at $s = 0.2$, where effects are largest and then track how signs and magnitudes change with sparsity for both architectures. Table 2 lists I for GSM8K, $\text{TruthfulQA}_{\text{MC2}}$ and MMLU computed from Eq. (1) and the aggregated CSVs, while Figures 1 and 2 illustrate the sparsity trends. Figures 3 and 4 provide heatmap visualizations of interaction patterns across all benchmarks for both models.

At $s = 0.2$, I on TruthfulQA is strongly positive for both architectures: approximately $+7$ for Falcon and between $+12$ and $+14$ for LLaMA, depending on the hybrid. Hybrid pipelines thus yield higher TruthfulQA accuracy than would be predicted by simply adding the separate effects of pruning and quantization, behaving as a powerful regularizer for truthfulness-style evaluation. However, as shown in Figures 1 and 2, this positive synergy weakens at extreme sparsity ($s = 0.5$).

In contrast, at $s = 0.2$, GSM8K and MMLU exhibit large negative I values, especially on Falcon (around -16 and -15). Extended metrics on LLaMA show similarly negative interactions on HellaSwag (around -20), ARC-Challenge and WinoGrande. Combining pruning and low-bit quantization therefore removes more effective capacity than the sum of their independent effects at moderate sparsity. As sparsity increases, these negative interactions diminish or flip to mildly positive values at extreme sparsity ($s = 0.5$), as illustrated in Figures 1 and 2.

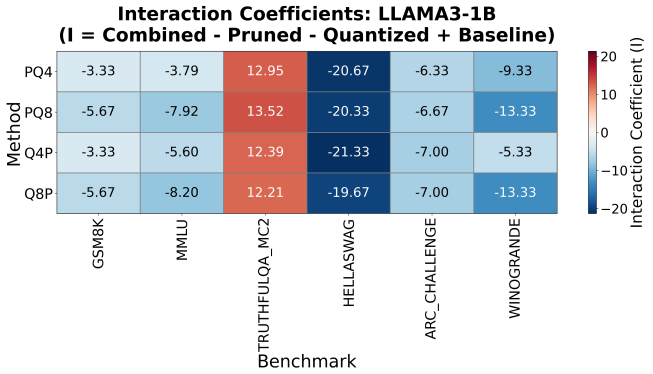

Figure 3: Interaction heatmap for LLaMA-3.2-1B at $s = 0.2$. TruthfulQA exhibits strong positive interactions, while reasoning/commonsense benchmarks are negative.

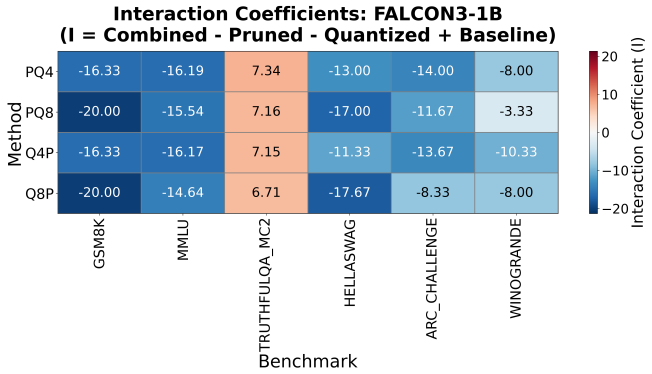

Figure 4: Interaction heatmap for Falcon3-1B-Base ($s = 0.2$) across Q4P/PQ4 and Q8P/PQ8. TruthfulQA is the only benchmark with consistently positive I.

Model	Metric	Q4P	PQ4	Q8P	PQ8
Falcon3-1B-Base					
	GSM8K	−16.2	−16.2	−20.0	−20.0
	$\text{TruthfulQA}_{\text{MC2}}$	+7.1	+7.3	+6.3	+5.7
	MMLU	−14.6	−15.5	−17.1	−17.1
LLaMA-3.2-1B					
	GSM8K	−3.8	−3.8	−5.7	−5.7
	$\text{TruthfulQA}_{\text{MC2}}$	+12.4	+13.5	+13.3	+13.9
	MMLU	−5.6	−3.8	−7.0	−5.3

Table 2: Interaction coefficients I for hybrid pipelines at $s = 0.2$. Positive values indicate synergy, negative antagonism.

Comparing Q4P and PQ4 ordering has relatively small influence compared to task and architecture: I differs by at most 1-2 points within each metric. This suggests that, at least for global magnitude pruning plus post-training quantization, the core behavior of hybrids is largely invariant to whether pruning happens before or after quantization. The decisive factor is whether the techniques are combined, not in which order.

4.3 Effect of Sparsity and 8-bit Hybrids

To determine whether interaction behavior is specific to a single sparsity level or 4-bit precision, we sweep sparsity over $s \in \{0.2, 0.3, 0.5\}$ and include 8-bit hybrids (Q8P, PQ8). Across Falcon3-1B-Base and LLaMA-3.2-1B, we observe three main trends. First, increasing s generally reduces accuracy across benchmarks and pipelines, but not monotonically: moving from no pruning to $s = 0.2$ produces the largest drops on GSM8K and MMLU, while further increasing sparsity to $s = 0.5$ often recovers some reasoning performance even as TruthfulQA declines relative to its $s = 0.2$ peak. Second, Figures 1 and 2 reveal consistent non-monotonic interaction patterns across both architectures: interactions are strongest at moderate sparsity ($s = 0.2, 0.3$) and approach zero (near-additive behavior) at extreme sparsity ($s = 0.5$). At $s = 0.2$, TruthfulQA exhibits strong positive I (Falcon: $\approx +7$; LLaMA: $\approx +13$), while GSM8K, MMLU, ARC, HellaSwag and WinoGrande show strong negative interactions (Falcon GSM8K: ≈ -20; LLaMA MMLU: ≈ -4 to -6). As s increases to 0.3, magnitudes generally grow while signs remain consistent. At $s = 0.5$, TruthfulQA's positive synergy weakens on both models and many reasoning and knowledge tasks move back toward zero or flip to mild positive values, indicating near-additive behavior rather than continued amplification. Both architectures show a peak in non-monotonic interaction at intermediate pruning levels, challenging the assumption that compression effects stack monotonically. Third, 8-bit hybrids (Q8P, PQ8) produce minimal practical differences compared to 4-bit hybrids: their accuracies closely match those of Q4P/PQ4 at the same s and their interaction coefficients maintain the same sign. Moving from 4-bit to 8-bit does not eliminate antagonism on reasoning and knowledge tasks, nor does it reduce the strong synergy observed on TruthfulQA.

4.4 Efficiency Metrics and Pareto Analysis

We next examine deployment metrics at sparsity $s = 0.2$, focusing on model size, latency, throughput and peak VRAM. This view clarifies which pipelines actually reduce resource use and where trade-offs arise. Table 3 reports the logged values for Falcon and LLaMA.

For both models, 4-bit quantization dominates memory savings, reducing model size and VRAM by approximately 2-3×. Falcon's size drops from 3339 MB to 1640 MB and VRAM from 3198 MB to 1627 MB; LLaMA's from 2472 MB to 1012 MB and 2373 MB to 1025 MB, respectively. Pruning alone leaves these quantities unchanged.

Latency and throughput exhibit clear trade-offs. Due to backend differences (vLLM for full-precision baselines vs. HF+BitsAndBytes for quantized models) and dequantization overhead, quantization increases latency and decreases throughput: Falcon's throughput falls from ∼46 to ∼18 samples/s and LLaMA's from ∼52 to ∼21. Hybrid pipelines closely mirror the 4-bit profile, with negligible additional cost from pruning.

From an accuracy-VRAM perspective, 4-bit and hybrid configurations constitute the Pareto frontier: they dominate baseline and pruned-only models in memory while maintaining acceptable accuracy for some tasks, as shown in Figure 5. From an accuracy-throughput perspective, full-precision baselines remain competitive because quantization trades memory

Pipeline	Falcon3-1B-Base				LLaMA-3.2-1B			
	Size (MB)	Latency (s)	TP	VRAM (MB)	Size (MB)	Latency (s)	TP	VRAM (MB)
Baseline	3339	0.0219	45.6	3198	2472	0.0194	51.7	2373
Pruned	3339	0.0217	46.0	3198	2472	0.0196	50.9	2372
Quantized4	1640	0.0542	18.5	1627	1012	0.0479	20.9	1025
Quantized8	2206	0.1096	9.1	2122	1499	0.0928	10.8	1445
Q4P	1640	0.0559	17.9	1627	1012	0.0498	20.1	1025
PQ4	1640	0.0541	18.5	1627	1012	0.0509	19.7	1025
Q8P	2206	0.0727	13.8	2120	1499	0.0923	10.8	2120
PQ8	2206	0.0729	13.7	2120	1499	0.0909	11.0	2120

Table 3: Efficiency metrics for Falcon3-1B-Base and LLaMA-3.2-1B at sparsity $s = 0.2$. TP = throughput (samples/s).

for speed on this hardware, illustrated in Figure 6. These observations highlight that the most suitable pipeline depends strongly on which deployment axis is prioritized.

4.5 Preliminary Qwen Results

Due to GPU time constraints, Qwen2.5-1.5B is evaluated only under a reduced set of pipelines and benchmarks. Qualitatively, it follows the same structure as Falcon and LLaMA: 4-bit quantization preserves most accuracy and substantially reduces VRAM, while hybrids substantially increase TruthfulQA scores and markedly degrade GSM8K and MMLU. The negative I values on Qwen's GSM8K and MMLU are even larger in magnitude than those of Falcon, suggesting greater fragility to hybrid compression. A full exploration of Qwen-family models is left to future work.

5 Discussion

In this section, we first interpret interaction effects and why hybrids help or hurt specific tasks, then draw deployment takeaways and limitations. Together these points answer how pruning–quantization interactions shape accuracy (RQ1) and deployment metrics (RQ2).

5.1 Interpretation of Interaction Effects

Our experiments demonstrate that hybrid compression cannot be inferred by simply interpolating between pruning and quantization curves. The interaction coefficient I reveals three key insights. First, TruthfulQA exhibits consistently positive I across Falcon and LLaMA, with hybrids delivering up to 13.5 absolute points over the additive baseline at moderate sparsity ($s = 0.2$) while halving VRAM. However, this positive synergy weakens at extreme sparsity ($s = 0.5$), as shown in Figures 1 and 2. Hybrid compression therefore provides a practical option for truthfulness-oriented deployments at moderate pruning levels, such as moderation, safety filtering or fact-checking, but pushing sparsity too high diminishes these benefits. Second, most other benchmarks (GSM8K, MMLU, ARC, HellaSwag, WinoGrande) show large negative I at moderate sparsity ($s = 0.2, 0.3$), indicating that hybrid pipelines underperform the naive expectation based on standalone techniques. This antagonism is strong enough to offset the gains of quantization and pruning individually. Critically, these negative interactions diminish or flip to mild positive values at extreme sparsity ($s = 0.5$), revealing that hybrid compression sensitivity peaks at intermediate pruning levels rather than increasing monotonically. Third, interaction behavior shows consistent non-monotonic patterns across both Falcon and LLaMA: interactions are strongest at moderate sparsity ($s = 0.2, 0.3$) and move toward zero (near-additive behavior) at extreme sparsity ($s = 0.5$). This pattern holds for both positive interactions (TruthfulQA) and negative interactions (reasoning tasks), suggesting a fundamental property of hybrid compression rather than an architecture-specific phenomenon. Because interaction effects peak at intermediate pruning levels instead of steadily increasing, practitioners should measure interaction coefficients across sparsity levels to select an operating point suited to their task mix.

5.2 Implications for Deployment

Applications that focus on truthfulness can benefit from Q4P and PQ4 hybrids at moderate sparsity ($s = 0.2, 0.3$), maximizing positive interaction coefficients on TruthfulQA within 12 GB VRAM, though these settings degrade reasoning and knowledge tasks and synergy diminishes at $s = 0.5$ as interactions approach additivity. For general-purpose systems, standalone 4-bit quantization at moderate sparsity provides superior trade-offs, achieving roughly 50% memory reduction with minimal accuracy loss while avoiding the severe negative interactions that hybrids exhibit on GSM8K and MMLU. When hybrid compression is required, increasing sparsity to $s = 0.5$ reduces antagonism but absolute accuracy remains limited. The non-monotonic interaction pattern indicates that compression strategies must be sparsity-specific: hybrids show peak interaction magnitudes at $s = 0.2, 0.3$ and converge toward zero at $s = 0.5$, necessitating systematic profiling of interaction coefficients across sparsity levels for optimal pipeline selection.

5.3 Limitations and Future Work

This study analyzes post-training hybrid compression under single-GPU constraints across two 1B-scale architectures, three sparsity levels and standard 4/8-bit quantization, using unstructured pruning without sparse kernels to isolate interaction effects on accuracy.

Hardware sparsity and transferability. Measured model size and peak VRAM for pruning-only pipelines here do not reflect savings from hardware-supported unstructured or structured sparsity (e.g. 2:4 patterns). The interaction coefficient

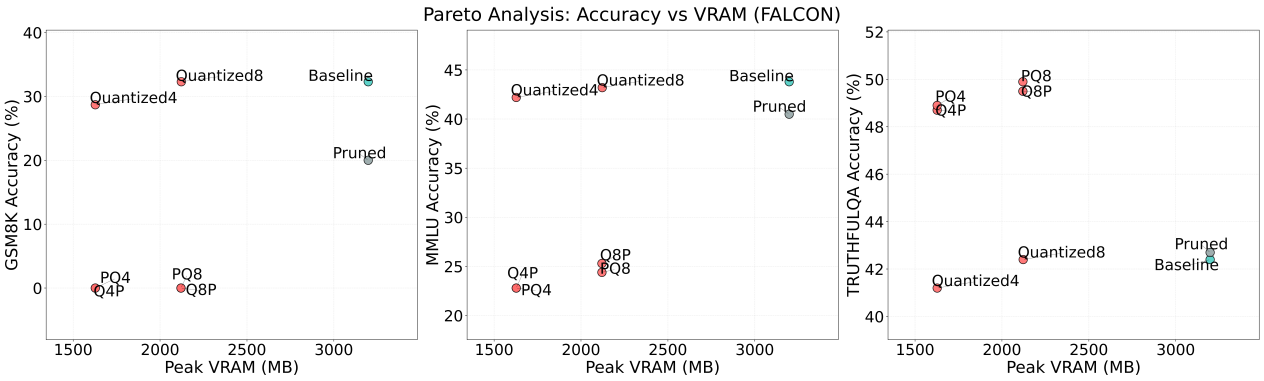

Figure 5: Accuracy vs. peak VRAM for Falcon3-1B-Base ($s = 0.2$). Quantized and hybrid points define the VRAM Pareto frontier.

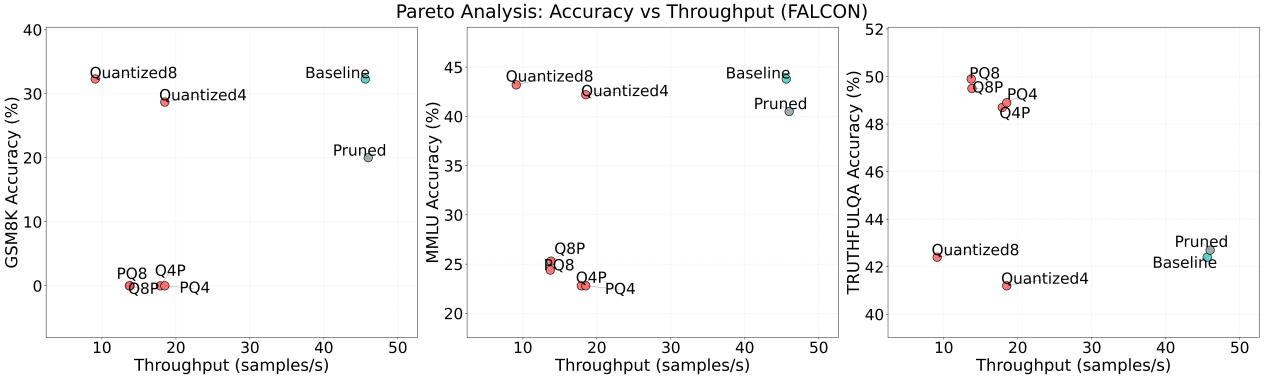

Figure 6: Accuracy vs. throughput for Falcon3-1B-Base ($s = 0.2$). Full-precision baselines retain the best speed, while hybrids trade accuracy for memory headroom.

remains defined from task metrics under fixed pipelines; we expect *accuracy-side* signs and relative ordering of interactions to remain a useful guide when the same logical sparsity is executed with sparse kernels, subject to kernel and format details, while *efficiency* Pareto curves should be re-measured when pruning reduces memory or latency. Intermediate sparsities, mixed-precision or structured sparsity and hardware-aware kernels (e.g. 2:4) could further probe the interaction landscape and reveal how acceleration changes the observed patterns. All runs omit retraining, distillation or hybrid-aware fine-tuning to expose intrinsic interactions; exploring light fine-tuning (e.g. LoRA) may mitigate antagonism or amplify synergy and could be guided by I profiles. Efficiency results are measured at batch size one on a single GPU; extending to larger batches, multi-GPU settings or additional hardware would broaden deployment insights, while the accuracy-side interaction trends are expected to persist. Evaluation is limited to task accuracy and common efficiency metrics; extending I to calibration, robustness or long-form generation would test whether the same non-monotonic behavior holds across broader metrics. Adding more architectures beyond Falcon, LLaMA and the preliminary Qwen2.5-1.5B check (e.g. Mistral, Phi, Gemma) would further validate the generality of the interaction patterns. Future directions include using I as an objective or constraint in automated compression policy search, layer-wise interaction profiling to localize sources of synergy or antagonism and hybrid-aware fine-tuning to shift the sparsity level that balances truthfulness gains against reasoning losses.

6 Conclusion

This paper introduces an interaction coefficient I for analyzing hybrid pruning-quantization pipelines and applies it to a systematic study of Falcon3-1B-Base and LLaMA-3.2-1B under realistic single-GPU constraints. We find that interaction effects are *non-monotonic in sparsity* and show consistent patterns across both architectures: interactions are strongest at moderate sparsity ($s = 0.2, 0.3$) and move toward zero (near-additive behavior) at extreme sparsity ($s = 0.5$). This non-monotonicity contradicts the common assumption that compression effects stack additively or increase monotonically with sparsity and reveals that optimal hybrid compression recipes must account for sparsity-dependent interaction behavior. Quantization emerges as the primary lever for memory savings, while pruning alone is usually dominated and hybrids must be applied selectively based on interaction-aware profiling. By making interaction effects explicit and reproducible, our work bridges a key gap in the LLM compression literature and provides a practical framework for sparsity-aware, deployment-specific compression decisions.

References

[Artetxe *et al.*, 2020] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, 2020.

[Bisk *et al.*, 2020] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:7432–7439, 2020.

[Clark *et al.*, 2018] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint*, 2018.

[Clark *et al.*, 2019] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2924–2936, 2019.

[Cobbe *et al.*, 2021] 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.

[Dettmers *et al.*, 2023] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2023.

[Elias and Dan, 2023] Frantar Elias and Alistarh Dan. Gptq: Accurate post-training quantization for generative pretrained transformers. In *International Conference on Learning Representations*, 2023.

[Frantar and Alistarh, 2023] Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning*, page 414, 2023.

[Gao *et al.*, 2023] Leo Gao, Jonathan Tow, Stella Biderman, et al. A framework for few-shot language model evaluation, 2023.

[Girija *et al.*, 2025] Sanjay Surendranath Girija, Lakshit Arora, Shashank Kapoor, Dipen Pradhan, Aman Raj, and Ankit Shetgaonkar. Optimizing llms for resource-constrained environments: A survey of model compression techniques. In *IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1657–1664, 2025.

[Han *et al.*, 2016] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *4th International Conference on Learning Representations, ICLR 2016*, San Juan, Puerto Rico, 2016.

[Hendrycks *et al.*, 2021] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[Kang *et al.*, 2025] Yuhan Kang, Zhongdi Luo, Mei Wen, Yang Shi, Jun He, Jianchao Yang, Zeyu Xue, Jing Feng, and Xinwang Liu. Hwpq: Hessian-free weight pruning-quantization for llm compression and acceleration. *arXiv preprint*, 2025.

[Kim *et al.*, 2025] Gun Il Kim, Sunga Hwang, and Beakcheol Jang. Efficient compressing and tuning methods for large language models: A systematic literature review. *ACM Computing Surveys*, 57(10):1–39, 2025.

[Lin *et al.*, 2022] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3214–3252, 2022.

[Lin *et al.*, 2025] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Guangxuan Xiao, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *GetMobile: Mobile Computing and Communications*, 28(4):12–17, 2025.

[Ma *et al.*, 2023] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: on the structural pruning of large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.

[Meta AI, 2024] Meta AI. LLaMA 3.2-1B model card, 2024. Model card.

[Qu *et al.*, 2025] Xiaoyi Qu, David Aponte, Colby Banbury, Daniel P. Robinson, Tianyu Ding, Kazuhito Koishida, Ilya Zharkov, and Tianyi Chen. Automatic joint structured pruning and quantization for efficient neural network training and compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15234–15244, 2025.

[Sakaguchi *et al.*, 2021] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

[Shinde, 2024] Tushar Shinde. Adaptive quantization and pruning of deep neural networks via layer importance estimation. In *Workshop on Machine Learning and Compression, NeurIPS 2024*, 2024.

[Technology Innovation Institute (TII), 2025] Technology Innovation Institute (TII). Falcon3-1B-Base model card, 2025. Model card.

[Yang and others, 2024] Xia Yang et al. Qwen2.5-1.5B model card, 2024. Model card.

[Zellers *et al.*, 2019] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the*

57th Annual Meeting of the Association for Computational Linguistics, pages 4791–4800, 2019.

[Zhou *et al.*, 2025] Zihan Zhou, Simon Kurz, and Zhixue Zhao. Revisiting pruning vs. quantization for small language models. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 12055–12070, 2025.