Unmasking the Hidden Fairness, Bias, and Safety Costs of Compression with Mixture-of-Expert Models

Anonymous Authors[1]

Abstract

Mixture-of-Experts (MoE) language models are often compressed for deployment, yet the fairness, bias, and safety effects of these interventions have remained less examined within Mixture-of-Experts architectures, particularly when Expert Compression (EC) methods are used or safety data is also varied. In this work, we investigate how compression changes fairness and safety behavior in MoE language models under different safety-data proportions. Across Qwen3-30B-A3B-Instruct-2507 and ERNIE-4.5-21B-A3B-PT, we find that EC is highly sensitive to safety data, with HC-SMoE expert merging producing the strongest instability overall. Quantization is generally more stable than EC, but it is not neutral, and weight sparsity can also introduce substantial shifts across the benchmark suite. We further show that benchmark level improvements can be misleading: some heavier pruning settings appeared less erratic on selected fairness benchmarks, while prompt matched generation diagnostics still showed greater behavioral drift from the dense baseline. These results suggest that fairness under MoE compression does not follow a simple monotonic pattern, and that safety data choices should be treated as a central consideration when compressing models intended for fairer deployment.

1. Introduction

Model compression can decrease the significant computational cost of deploying state-of-the-art Large Language Models (LLMs), increasing accessibility in both research and applications. While compressed models are typically evaluated based on their generative and predictive performance, their impact on fairness, bias, and safety has increasingly been recognized as an equally important measure of evaluation. A wide body of existing literature has found that model compression methods, including quantization (Gonçalves & Strubell, 2023; Marcuzzi et al., 2026; Ganaie et al., 2025), weight sparsity (Hooker et al., 2020; Ganaie et al., 2025), and distillation (Gonçalves & Strubell, 2023; Mohammadshahi & Ioannou, 2025; Ganaie et al., 2025), can unpredictably alter or exacerbate existing biases in uncompressed models. Notably, however, MoE architectures have been less examined within this line of work.

Compared to similar-sized dense models, MoEs decouple total parameter count from active FLOPs and offer improved scalability and data efficiency during training (Zoph et al., 2022; Qwen Team, 2025; Muennighoff et al., 2025) as well as improved inference latency (Fedus et al., 2022). The large parameter counts of MoE models often require compression methods to enable cost-effective usage. This has motivated MoE-specific compression methodologies, including *Expert Compression (EC)* methods such as expert merging (Chen et al., 2025) and pruning (Lasby et al., 2026). EC methods have already found significant real world application in model serving, notably within the local LLM serving community.

In this work we reveal the impact of MoE EC methods on *Fairness, Bias, and Safety (FBS)* — as far as we are aware, for the first time. We further explore and compare the impact of popular fine-grained compression methods, i.e. weight sparsity and quantization, on MoEs. While some prior work has explored fairness in MoEs, this has largely focused on developing fairness-aware methods rather than examining how *compression* affects Fairness, Bias, and Safety (FBS) shifts in modern MoEs (Germino et al., 2024; Yang et al., 2025). Existing EC compression methods such as expert merging (Chen et al., 2025) and expert pruning (Lasby et al., 2026) have mainly been studied in terms of parameter reduction, memory efficiency, and performance retention.

Contributions: The goal of this work is to examine the impact of model compression on FBS in MoE language models, with a focus on EC, while also comparing to the impact of quantization and weight sparsity which are relatively underexplored in the MoE context. Specifically, we focus

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

on the unique behavioral shifts induced when compression targets the experts directly, as opposed to applying uniform compression across all layers of the network. Our empirical investigation yields the following key findings:

- **Impact of Expert Compression on Fairness, Bias, and Safety:** As far as we are aware, we are the first to explore the impact of EC on FBS in MoE models.
- **Divergence of impact across model compression methods:** We demonstrate that EC triggers the most severe benchmark regressions compared to quantization and weight sparsity.
- **Safety data sensitivity under expert compression:** We find that varying the proportion of safety data during calibration introduces larger shifts under expert compression than under quantization.
- **Independence across safety and fairness dimensions:** We establish that compression affects different safety and fairness dimensions independently. As a result, preservation on one benchmark does not imply preservation across the broader fairness and safety profile.
- **Disconnect between benchmark scores and generative behavior:** We show that benchmark-level improvements or stability can conceal broader behavioral drift from the dense baseline. By measuring the cross-NLL of compressed outputs with the uncompressed model, we find that apparently less biased compressed models may be less aligned with the uncompressed model.

2. Related Work

Defining and quantifying fairness, bias, and safety. In this work, we define fairness to refer to model behavior that does not systematically favor or disadvantage social groups across gender, race, religion, nationality, or other demographic axes. In the context of LLMs, safety refers to the comprehensive set of practices and evaluations designed to ensure the model operates benignly and actively prevents the generation of harmful outputs. Safety fine-tuning is intended to reduce harmful or undesirable model behaviors, but it remains unclear to what degree these effects are preserved in a compressed MoE. Because FBS is not directly observable as a single quantity, it is evaluated through multiple complementary benchmarks that investigate different manifestations of biased behavior (Parrish et al., 2022; Nangia et al., 2020; Nadeem et al., 2021; Röttger et al., 2023; Zou et al., 2023).

Model compression. The massive scale of LLMs has motivated significant prior work into compressing their weights, activations, and key/value caches. MoEs exacerbate this issue even further due to numerous expert feedforward modules per decoder block (**?**). As such, compressing the parameters of these experts is an active research area which can directly reduce the overall costs associated with serving large-scale MoE models. For large-scale MoEs such as Kimi-K2 (Team, 2026), the routed and shared experts account for approximately 99% of the total parameter count. In this work, we consider three main classes of compression algorithms: quantization, weight sparsity, and EC.

Quantization maps model weights and/or activations from their native precision to lower-precision data types, significantly reducing memory footprint and accelerating inference. In this study, we consider the Post-Training Quantization (PTQ) paradigm, which applies compression directly to a pretrained model. Weight sparsity reduces computational overhead by zeroing out individual parameters. Historically, post-training pruning relied on simple magnitude-based thresholds. However, recent advancements have demonstrated that preserving model fidelity at high sparsity levels requires incorporating activation distributions (Sun et al., 2023) or approximate second-order Hessian information (Frantar & Alistarh, 2023).

Expert Compression (EC) encompasses interventions that directly alter the macro-architecture of MoE models. Unlike weight-level sparsity or quantization, EC fundamentally modifies the established coordination between the routing mechanism and the experts. Recent literature addresses MoE deployment bottlenecks through two primary EC paradigms: expert merging that aims to combine functionally similar experts (Chen et al., 2025), and expert pruning, which drops low-saliency experts entirely (Lasby et al., 2026). Because these techniques permanently alter the availability of experts, they represent a more severe intervention than parameter-level compression. Consequently, understanding how these EC macro-changes disproportionately affect FBS is an emerging priority in MoE compression research.

Quantifying the effect of compression. Past research demonstrates that model compression is rarely a mathematically neutral operation; rather, it can preserve, amplify, or unpredictably alter existing biases (Hooker et al., 2020; Gonçalves & Strubell, 2023; Ganaie et al., 2025). Recent pruning work further shows that sparsity can amplify reliance on spurious features and degrade performance for bias-conflicting or underrepresented groups, while also motivating pruning-aware debiasing methods (Hong et al., 2026). The bias introduced during compression can stem from two compounding sources: *algorithmic bias*, where the mechanics of the compression method itself disproportionately degrade performance on long-tail or minority distributions, and *data bias*. Because many advanced compression algorithms require a calibration dataset, they risk embedding the dataset's intrinsic biases into the compressed model's parameters.

Consequently, these techniques do not behave identically when evaluated beyond standard metrics. For example, quantization can successfully reduce overt toxicity while simultaneously increasing stereotype-related unfairness (Marcuzzi et al., 2026). Furthermore, aggregate bias scores often mask asymmetric, subgroup-level shifts, with predictive uncertainty playing a major role in determining which specific responses flip after compression (Hua et al., 2026). Evaluating these methods therefore requires a multidimensional analysis of FBS outcomes.

While prior work on fairness in MoEs has largely focused on designing fairness-aware architectures from the ground up (Germino et al., 2024; Yang et al., 2025), to the best of our knowledge, no previous work has systematically examined how FBS metrics shift in *existing* MoE language models after post-training compression. Recent MoE compression methods have been evaluated almost exclusively on performance and quality retention. However, recent literature indicates that MoE routing behavior is intimately tied to alignment (Kim et al., 2026), and specific experts often encapsulate behavioral constraints like safety rather than serving purely generic computational roles (Fayyaz et al., 2026). Altering the expert structure or quantizing expert weights therefore introduces fairness and safety risks that extend far beyond basic efficiency.

3. Preliminaries

A sparse MoE layer replaces a standard dense feed-forward network with a routing mechanism and a set of N independent expert networks. Given an input token representation $\mathbf{x} \in \mathbb{R}^d$, the MoE layer selectively routes the token to a small subset of K experts (where $K \ll N$), maintaining high model capacity while bounding active computation. Some MoE architectures also include shared experts; for notational simplicity, we describe the routed expert component here.

Let $E_i(\mathbf{x})$ denote the functional output of the i-th expert, and let $G_i(\mathbf{x})$ represent the sparse, normalized gating weight assigned by the router. To enforce sparsity, the router selects only the top K experts, forming an active set $\mathcal{T}(\mathbf{x})$. For all unselected experts ($i \notin \mathcal{T}(\mathbf{x})$), the gate weight is strictly zero ($G_i(\mathbf{x}) = 0$). The final output of the MoE layer, $\mathbf{y}$, is the weighted sum of the active experts' outputs: $\mathbf{y} = \sum_{i \in \mathcal{T}(\mathbf{x})} G_i(\mathbf{x}) E_i(\mathbf{x})$. Since computation is restricted to the K active experts, compressing the individual expert modules $E_i(\mathbf{x})$ directly reduces the active memory footprint and can decrease inference latency by reducing the VRAM I/O bandwidth required to load expert parameters.

4. Methods

To systematically unmask the hidden fairness and safety costs of compressing MoE language models, our methodology bridges post-training compression techniques with a targeted safety-data ablation framework. Because most compression algorithms rely on a calibration dataset to assess parameter importance or functional similarity, they are inherently vulnerable to biases imparted by that calibration distribution. To investigate this, we vary the proportion of safety data used during the compression calibration phase. This ablation allows us to isolate how safety-aligned calibration data interacts with different compression mechanisms to influence downstream demographic bias, stereotype propagation, and harmful instruction robustness. We benchmark a comprehensive suite of post-training compression techniques applied specifically to the MoE expert modules, grouping these baselines into three primary categories (see Appendix A for formal definitions):

Weight quantization. We evaluate three established techniques designed to reduce the numerical precision of expert weights. Our baselines include Activation-aware Weight Quantization (AWQ) (Lin et al., 2024), which mitigates quantization error by observing activation channels to identify and selectively preserve highly salient weights, and Generative Post-Training Quantization (GPTQ) (Frantar et al., 2022), which leverages approximate second-order (inverse Hessian) information to iteratively minimize layer-wise reconstruction loss. In contrast to these data-dependent integer methods, we also benchmark Dynamic FP8 (Shen et al., 2024), which utilizes a data free, round-to-nearest approach with static per-channel scaling to cast weights into an 8-bit floating point format, coupled with dynamic runtime scaling for activations.

Weight sparsity. To assess the impact of weight sparsity on model fairness and safety, we benchmark WANDA (Sun et al., 2023) and SparseGPT (Frantar & Alistarh, 2023), testing both unstructured and 2:4 semi-structured weight sparsity configurations. WANDA prunes weights by computing a proxy score that combines weight magnitudes with the norms of their corresponding input activations, efficiently inducing sparsity without iterative fine-tuning. Alternatively, SparseGPT adapts the second-order optimization framework of GPTQ to compute a pruning mask that minimizes layer-wise reconstruction error, adjusting the remaining non-zero weights to compensate for the pruned parameters.

Expert Compression (EC). We explore EC interventions that directly reduce the total number of experts, which we hypothesize introduces the most severe fairness shifts due to the alteration of alignment-critical routing pathways. We evaluate Hierarchical Clustering for Sparse Mixture-of-Experts (HC-SMoE) (Chen et al., 2025), which clusters functionally similar experts over the calibration set and

merges them with frequency weighted parameter averaging. We contrast this with Router-weighted Expert Activation Pruning (REAP) (Lasby et al., 2026), which prunes experts with the smallest functional contribution conditioned on the subset of tokens where it is actively routed.

5. Experiments

5.1. Models, data, compression algorithms

`Qwen3-30B-A3B-Instruct-2507` (Qwen Team, 2025) and `ERNIE-4.5-21B-A3B-PT` (Baidu ERNIE Team, 2025) were used to conduct our experiments. Qwen and ERNIE are both modern, fine-grained MoEs, with 128 and 64 experts per layer, respectively. ERNIE also includes two shared experts in each MoE module that every token is unconditionally routed to, whereas Qwen MoE layers consist of exclusively routed experts. To calibrate our compression methods, we used 128 samples packed to sequence lengths of 2048 from the `tulu3-mixture`[1] dataset (Lambert et al., 2024), except for Dynamic FP8, which does not require calibration. We used three versions of the dataset corresponding to safety data proportions of 0%, 11.8%, and 100%. 11.8% represents the default proportion of safety data relative to the total dataset size. The safety data subsets selected included Tulu 3 WildGuardMix (Han et al., 2024), Tulu 3 WildJailbreak (Jiang et al., 2024), and CoCoNot (Brahman et al., 2024). We sampled from all subsets of the designated mixture uniformly to achieve the target sample count.

For EC, we compressed each MoE layer uniformly by 50% or 25%. For AWQ and GPTQ quantization, we compressed 16-bit checkpoints to 8- or 4-bit weights and left activations in full 16-bit precision. For WANDA and SparseGPT, we compressed models to 50% sparsity and compared both unstructured and 2:4 semi-structured sparsity. Where applicable, we use the default hyperparameter settings for each algorithm and keep them fixed across all datasets, models, and runs. For quantization and weight sparsity, we use the LLM Compressor library (Red Hat AI and vLLM Project, 2024). For EC, we extend the REAP source code[2]. See Appendix B for a discussion of the compute resources used to run the experiments.

5.2. Benchmarks

To evaluate FBS across the evaluated models and ablations, we utilized the `lm_eval` (Gao et al., 2023) harness across three core categories: social bias, safety & refusal, and capability. For social bias, we measured likelihood-based stereotype preferences using CrowS-Pairs and StereoSet, alongside BBQ Ambiguous Bias and BBQ Disambiguated Bias to assess reliance on stereotypes in ambiguous and disambiguated contexts. For safety and refusal, we paired AdvBench, which evaluates the correct refusal of harmful prompts, with XSTest, which measures whether harmless prompts are over-refused. Finally, we utilized BBQ Accuracy as our capability anchor to verify that core reading comprehension and instruction following remain intact post-compression. All evaluated models were run under the same `lm_eval` configuration and prompt formatting across baseline and compressed variants. Scoring followed the task-specific protocols implemented in `lm_eval`, with multiple choice and preference-based tasks scored accordingly. Full details of how benchmarks are evaluated and reported can be found in Appendix C.

[1] We omit the Tulu 3 Hardcoded subset entirely

[2] Anonymous code repo: https://anonymous.4open.science/r/alchemoe-F02A/README.md

6. Results

6.1. Fairness, Bias, and Safety Evaluations

Table 1 presents the main post-compression FBS results at 50% compression for `Qwen3-30B-A3B-Instruct-2507` and `ERNIE-4.5-21B-A3B-PT`, respectively. Both models were evaluated across proportions of safety data for quantization, EC, and weight sparsity regimes. To offer a more comprehensive view, extended evaluations across 25% and 75% compression ratios are provided in Appendix E (Tables 2 and 3). The results demonstrate that no compression method preserves all dimensions uniformly. Quantization generally remains closest to the dense baseline, within expert compression methods, REAP pruning is less disruptive than HC-SMoE merging, and weight sparsity depends strongly on the model family, sparsity method, and mask structure. Results across BBQ Accuracy, BBQ bias scores, XSTest, and AdvBench often move in different directions under the same compression method, showing that apparent improvements on one metric can coincide with regressions on another.

Expert merging vs. pruning. *At the 50% compression rate, the separation between the two EC approaches, pruning and merging, is clear — the largest benchmark regressions occur under HC-SMoE merging*: For Qwen, HC-SMoE merging produces larger shifts than REAP pruning across BBQ bias scores, CrowS-Pairs, and StereoSet. The most extreme divergence appears in capability and selective refusal, where merged Qwen settings suffer a massive drop in BBQ Accuracy and XSTest scores collapse to near zero. This separation is also prominently visible on AdvBench, where REAP remains remarkably close to the dense baseline while HC-SMoE drops more substantially. The BBQ results show that lower bias scores must be interpreted alongside capability: HC-SMoE merging moves BBQ Ambiguous Bias closer to zero than REAP pruning, but this occurs alongside the strongest BBQ Accuracy collapse. BBQ Disam-

Table 1. Bias benchmark summary for 50% compression runs with baseline reference rows for Qwen3 and Ernie. Bold entries mark the best compressed value within each model block for each metric.

(a) Qwen3-30B-A3B-Instruct-2507

				Social Bias				Safety & Refusal		Capability
Compression type	**Algorithm**	**Scheme**	**Safety data proportion**	**BBQ Amb. Bias (↓)**	**BBQ Disamb. Bias (↓)**	**CrowS-Pairs Stereo. (50.0)**	**StereoSet Stereo. (50.0)**	**AdvBench Refusal (↑)**	**XSTest Refusal (↓)**	**BBQ Acc. (↑)**
Baseline	N/A	N/A	N/A	0.0433	0.0245	0.609	0.637	0.996	0.349	0.698
Quantization	AWQ	W8A16	0%	0.112	0.0258	0.589	0.654	0.996	0.224	0.540
			11.8%	0.109	0.0258	0.584	0.656	**0.998**	0.242	0.540
			100%	0.107	0.0266	0.590	0.652	0.996	0.231	0.540
	Dynamic FP8	FP8 W8A8	N/A	0.111	0.0254	0.596	0.660	0.996	0.231	0.542
	GPTQ	W8A8	0%	0.108	0.0261	0.593	0.655	0.994	0.240	0.540
			11.8%	0.109	0.0276	0.596	0.650	**0.998**	0.231	0.540
			100%	0.109	0.0261	0.592	0.652	0.996	0.227	0.540
Expert Compression	HC-SMoE	Merging	0%	0.0208	**0.0161**	0.528	0.549	0.798	**0.00444**	0.411
			11.8%	0.0182	0.0171	**0.509**	**0.541**	0.804	0.00889	0.420
			100%	**0.00968**	0.0186	0.518	0.543	0.950	0.0311	0.396
	REAP	Pruning	0%	0.0941	0.0184	0.562	0.620	0.989	0.171	0.516
			11.8%	0.0972	0.0256	0.587	0.634	0.993	0.218	0.529
			100%	0.0968	0.0334	0.621	0.642	0.992	0.280	0.497
Weight sparsity	SparseGPT	2:4	0%	0.115	0.0200	0.616	0.634	0.992	0.156	0.529
			11.8%	0.107	0.0209	0.614	0.641	**0.998**	0.164	0.528
			100%	0.0952	0.0260	0.596	0.644	0.996	0.267	0.528
		unstructured	0%	0.114	0.0198	0.592	0.650	0.994	0.220	0.552
			11.8%	0.112	0.0188	0.588	0.646	0.994	0.231	**0.559**
			100%	0.116	0.0206	0.594	0.654	0.992	0.271	0.548
	Wanda	2:4	0%	0.103	0.0251	0.607	0.617	0.973	0.0800	0.514
			11.8%	0.101	0.0242	0.612	0.616	0.987	0.0889	0.510
			100%	0.0950	0.0232	0.611	0.618	0.985	0.0467	0.508
		unstructured	0%	0.113	0.0191	0.627	0.641	0.994	0.140	0.537
			11.8%	0.115	0.0184	0.630	0.640	0.996	0.164	0.537
			100%	0.107	0.0210	0.614	0.653	**0.998**	0.158	0.534

(b) ERNIE-4.5-21B-A3B-PT

				Social Bias				Safety & Refusal		Capability
Compression type	**Algorithm**	**Scheme**	**Safety data proportion**	**BBQ Amb. Bias (↓)**	**BBQ Disamb. Bias (↓)**	**CrowS-Pairs Stereo. (50.0)**	**StereoSet Stereo. (50.0)**	**AdvBench Refusal (↑)**	**XSTest Refusal (↓)**	**BBQ Acc. (↑)**
Baseline	N/A	N/A	N/A	0.0889	0.0280	0.628	0.662	0.715	0.109	0.517
Quantization	Dynamic FP8	FP8 W8A8	N/A	0.0958	0.0293	0.633	0.663	0.738	0.109	0.512
	GPTQ	W8A8	0%	0.0941	0.0278	0.637	0.665	0.717	0.118	0.518
			11.8%	0.0974	0.0278	0.634	0.662	0.713	0.107	**0.519**
			100%	0.0962	0.0283	0.630	0.660	0.723	0.109	0.518
Expert Compression	HC-SMoE	Merging	0%	**0.0235**	0.0296	0.548	0.594	0.360	**0.00667**	0.343
			11.8%	0.0247	0.0285	0.555	0.592	0.469	0.0156	0.344
			100%	0.0242	0.0377	**0.510**	**0.552**	**0.806**	0.0556	0.402
	REAP	Pruning	0%	0.0425	0.0309	0.546	0.611	0.413	0.0178	0.456
			11.8%	0.0486	0.0270	0.566	0.614	0.594	0.0422	0.454
			100%	0.0476	0.0402	0.578	0.613	0.652	0.118	0.482
Weight sparsity	SparseGPT	2:4	0%	0.0675	0.0356	0.617	0.629	0.537	0.0733	0.477
			11.8%	0.0769	0.0332	0.617	0.628	0.642	0.129	0.478
			100%	0.0690	0.0373	0.572	0.637	0.677	0.124	0.455
		unstructured	0%	0.0770	0.0325	0.636	0.638	0.687	0.118	0.502
			11.8%	0.0816	0.0291	0.630	0.645	0.675	0.107	0.476
			100%	0.0741	0.0306	0.617	0.658	0.687	0.109	0.478
	Wanda	2:4	0%	0.0609	0.0444	0.567	0.620	0.448	0.0533	0.452
			11.8%	0.0587	0.0468	0.574	0.629	0.527	0.0556	0.453
			100%	0.0534	0.0386	0.574	0.620	0.615	0.0800	0.442
		unstructured	0%	0.0732	**0.0270**	0.621	0.633	0.517	0.0756	0.466
			11.8%	0.0760	0.0364	0.616	0.645	0.635	0.111	0.466
			100%	0.0586	0.0370	0.608	0.631	0.515	0.111	0.458

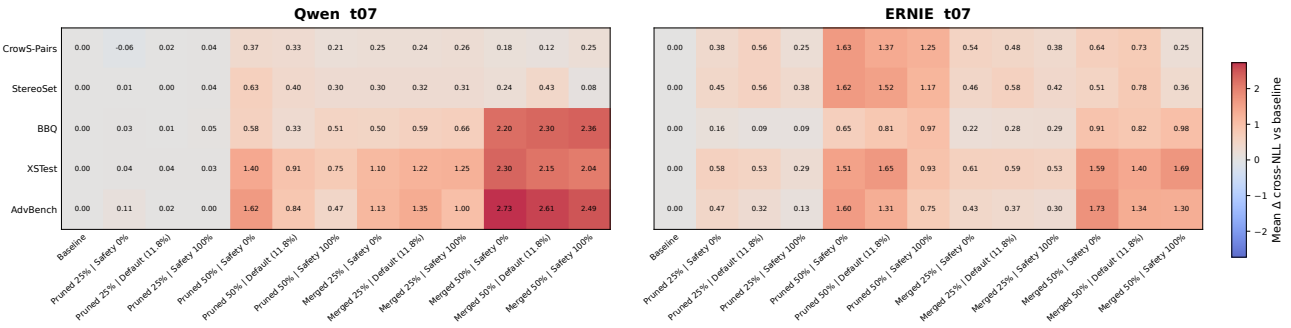

Figure 1. Dataset mean Δ cross-NLL relative to the dense baseline at temperature 0.7 for Qwen and ERNIE under EC settings. Positive values indicate greater behavioral drift from the dense baseline.

biguated Bias shows a more mixed pattern: Qwen merging also moves this score closer to zero, while ERNIE merging leaves it similar to or worse than the dense baseline, especially at the 100% safety data setting.

ERNIE shows a clearer separation between capability-adjusted demographic bias metrics and harmful instruction robustness under expert compression: ERNIE shows the same broad pattern, although the specific safety effects differ across model families. HC-SMoE merging produces much steeper declines in BBQ Accuracy than REAP pruning, which conversely causes the merged checkpoints to output seemingly "better" but artifactual BBQ Ambiguous Bias scores. BBQ Disambiguated Bias does not follow the same uniformly favorable pattern, with the 100% safety data HC-SMoE setting shifting above the dense baseline. AdvBench behaves differently from Qwen: several pruned settings fall strictly below the dense baseline, while the 100% safety data HC-SMoE setting abruptly rises above the baseline to 0.806.

Lighter compression does not always preserve the dense baseline more closely than heavier compression across the benchmark suite: While Table 1 focuses on the main 50% compression setting, the full compression ratio comparison is reported in Appendix E. These additional results show that the relationship between compression strength and fairness is not monotonic. In particular, some 50% pruning settings appear competitive with 25% pruning settings on selected benchmark metrics, even though the generation diagnostics below show a different pattern.

Impact of safety data on compression. *Using more safety data during compression can improve harmful instruction robustness while still worsening demographic bias metrics or general capabilities*: The safety data ablation is most visible within the EC settings. For Qwen HC-SMoE merging, increasing the safety data proportion from 0% to 100% successfully rescues AdvBench, pulling it from 0.798 to 0.950; however, BBQ Accuracy concurrently decreases from 0.411 to 0.396, indicating that the base model's capability is further compromised. For ERNIE HC-SMoE merging, AdvBench rises dramatically from 0.360 to 0.806 as the safety data proportion increases, while its capability (BBQ Accuracy) remains severely degraded relative to the dense baseline (0.402 vs 0.517).

Changing the safety data proportion produces larger benchmark shifts in expert-compressed models than in quantized models: This safety data sensitivity is not equally visible across compression families. For Qwen, AWQ and GPTQ remain close across the 0%, 11.8%, and 100% safety data settings, with only small changes across the reported benchmark metrics. ERNIE GPTQ shows the same pattern, with all three safety data settings staying close to the dense baseline and to each other. Dynamic FP8 is also close to the dense baseline at the reported 11.8% setting, but because it is only reported as a data-free row, it is less directly comparable to the full safety data ablations.

Impact of compression on safety. *Exaggerated refusal, harmful instruction robustness, and demographic bias move separately rather than as one shared safety axis*: The benchmark results also show that safety dimensions move independently. XSTest decreases sharply under EC, especially in the 50% merged runs. For Qwen, the merged XSTest scores fall close to zero across all safety data settings, while AdvBench does not show the same uniform pattern. In contrast, Qwen pruning stays close to the dense AdvBench baseline even though XSTest still decreases. ERNIE shows a second form of this separation: the 100% safety data HC-SMoE setting rises above the dense baseline on AdvBench, but it remains severely shifted on BBQ, CrowS-Pairs, and StereoSet. The BBQ metrics reinforce this point because BBQ Ambiguous Bias, BBQ Disambiguated Bias, and BBQ Accuracy do not move together.

Quantization vs. EC *Quantization is more stable than expert compression, but not neutral*: Across both model families, quantized models remain closer to the dense baseline on BBQ and XSTest than the expert compressed models do.

For Qwen, AWQ and GPTQ keep AdvBench near the dense baseline and preserve BBQ Accuracy far better than the extreme capability collapses seen under HC-SMoE merging, although BBQ Ambiguous Bias still shifts relative to the dense baseline. Dynamic FP8 also remains close to the dense baseline across most reported metrics. For Ernie, GPTQ and Dynamic FP8 similarly show smaller overall shifts compared to EC.

Weight sparsity vs. EC *Weight sparsity also changes the fairness and safety profile for both model families*: On Qwen, WANDA and SparseGPT decrease BBQ Accuracy relative to the dense baseline and trigger sharp drops in XSTest scores, while CrowS-Pairs and StereoSet often remain closer to the middle of their preferred ranges. The BBQ bias scores also shift under sparsity, showing that these changes are not only capability effects. Among the Qwen sparse settings, SparseGPT with an unstructured mask remains closest to the dense baseline in capability retention, whereas WANDA and the semi-structured settings are generally more disruptive to the overall behavioral profile.

ERNIE sparsity is less disruptive than HC-SMoE merging on core capability metrics, but still meaningfully shifts fairness and safety: ERNIE shows a specific sparsity profile where the sparse checkpoints reduce BBQ Accuracy relative to the dense baseline, but critically, their BBQ Accuracy values remain notably closer to the baseline than the massive drops seen under HC-SMoE merging. This indicates the integrity of the model holds up better under sparsity than macro-expert shifts. Harmful refusal still shifts under sparsity, however, with WANDA 2:4 producing some of the more noticeable regressions in both AdvBench and XSTest. Additional prompt-matched diagnostics for weight sparsity are provided in Figures 7 and 8, showing that sparsity also induces dataset-dependent generation drift, especially for semi-structured WANDA and SparseGPT settings.

6.2. Generation and Perplexity Diagnostics

To complement the benchmark results, we evaluated prompt-matched outputs relative to the dense baseline using mean Δ cross-NLL, cross perplexity, and a top-k entropy proxy. These diagnostics separate benchmark preservation from output-level preservation, since some heavier pruning settings can appear competitive with lighter pruning settings on selected fairness metrics. In Figure 1, however, the 25% pruned checkpoints generally remain closer to the dense baseline than the 50% settings in generation space, especially for Qwen. This means that apparent benchmark-level preservation, or even improvement, does not necessarily imply broader behavioral preservation.

Generation drift under EC. *For Qwen, HC-SMoE merging produces the largest generation-level drift*: Figure 1 shows that the dense baseline remains at zero Δ cross-NLL by construction, while most compressed checkpoints move away from the baseline. For Qwen, the 25% pruned checkpoints remain closest to the dense baseline at both temperature 0 and temperature 0.7. The 50% merged checkpoints show the largest drift, with substantially higher Δ cross-NLL than the corresponding pruned checkpoints. The 25% merged checkpoints also drift more than the 25% pruned checkpoints, matching the benchmark-level ordering where merging is more disruptive than pruning for Qwen.

For Ernie, the output space ordering differs from the benchmark ordering: The generation-level pattern is related but not identical. At temperature 0, the 25% checkpoints remain closer to the baseline, while the 50% checkpoints drift more strongly. Unlike Qwen, the 50% pruned ERNIE checkpoints show the largest generation-level drift, especially under temperature 0.7. The 50% merged ERNIE checkpoints also remain shifted from the dense baseline, but their Δ cross-NLL values are lower than the corresponding 50% pruned settings. This shows that the benchmark-level ordering between pruning and merging does not determine the output space ordering.

Effect of sampling and safety data. *Sampled generation makes compression-induced drift more visible*: The temperature 0.7 runs show larger drift than the temperature 0 runs for nearly all compressed checkpoints. This effect is visible for both Qwen and Ernie, but it is especially strong for ERNIE at 50% compression. This suggests that deterministic benchmark evaluation can understate the behavioral drift that appears under sampled generation.

Safety data does not produce a consistent fairness-preserving trend in generation space: The safety data settings in Figure 1 do not align uniformly with the benchmark results. In several 50% expert compression settings, increasing the safety data proportion reduces generation-level drift relative to the 0% or 11.8% safety data settings. However, this does not imply uniformly better fairness behavior. For example, the Qwen 50% merged checkpoint with 100% safety data has lower Δ cross-NLL than the other 50% merged Qwen settings, but in Table 1 it has the lowest BBQ Accuracy among the Qwen merged rows.

Entropy and benchmark preservation. *The entropy proxy does not fully explain the observed drift*: If compression only made the model more uncertain, then higher entropy would consistently track higher Δ cross-NLL. Instead, Figure 1 shows a more mixed pattern. Some Qwen merged checkpoints have high Δ cross-NLL without having the highest entropy, while some ERNIE pruned checkpoints at temperature 0.7 have both high entropy and high drift. This weakens the claim that the apparent benchmark changes are only due to more diffuse next token distributions.

7. Discussion

The results show that fairness and safety under compression do not move as a single property. Across both model families, demographic bias, exaggerated refusal, harmful instruction robustness, and stereotype preference metrics often shift in different directions. This makes single benchmark conclusions unreliable. A compressed checkpoint can appear stable or improved on one metric while still degrading on another.

Expert compression produces the clearest divergence across methods. REAP pruning generally stays closer to the dense baseline than HC-SMoE merging, while HC-SMoE produces the largest regressions and BBQ Accuracy drops. This suggests that removing experts and merging experts are not equivalent interventions. Both target the expert layer, but merging appears to more severely alter the routed expert pool. In MoE models, expert structure may therefore carry fairness and safety relevant behavior rather than only serving as an efficiency mechanism. The safety data ablation also shows that safety data does not uniformly stabilize compressed MoEs. Under expert compression, increasing the safety data proportion can improve harmful instruction robustness while BBQ Accuracy and bias scores remain shifted. The smaller variation under quantization suggests that calibration data has a larger effect when the compression method changes expert pathways directly. This means safety tuning and compression should not be treated as independent steps.

Quantization and weight sparsity are generally more stable than expert compression, but neither is neutral. Quantization stays closer to the dense baseline across most metrics. Weight sparsity shows a more model-dependent pattern, with Qwen and ERNIE responding differently across sparsity methods and mask structures. These results suggest that compression methods should be compared not only by efficiency or capability retention, but also by their fairness and safety effects.

The prompt-matched diagnostics show that benchmark-level preservation does not necessarily imply output-level preservation. Compression effects also become more visible under sampled generation, and the entropy proxy does not fully explain this drift, suggesting that compression can change which outputs the model prefers.

Overall, these results suggest that post-compression auditing should be part of the fairness and safety evaluation pipeline. For compressed MoE models, this should include demographic bias benchmarks, refusal and harmful instruction safety tests, and generation-based diagnostics.

8. Conclusion

Our results show a clear divide between EC, quantization, and weight sparsity interventions. Across both Qwen and Ernie, HC-SMoE expert merging (Chen et al., 2025) triggered the most severe benchmark regressions. AWQ (Lin et al., 2024) and GPTQ (Frantar et al., 2022) preserved the dense baseline much better, but they still shifted the fairness and safety profile.

The data shown within this study challenges the assumption that bias scales linearly with intervention size. On the benchmark suite, some heavier 50% pruning settings appeared less erratic than the lighter 25% setting, showing that single-metric evaluations can obscure these issues. However, the auxiliary generation-based diagnostics also showed that 25% pruning generally remained closer to the dense baseline in output behavior, while the merged 50% settings produced the largest behavioral drift overall. Several compressed models in our pipeline reached near neutrality on CrowS-Pairs, while BBQ Accuracy and BBQ bias scores showed that apparent bias improvements can coincide with capability degradation. AdvBench further reinforced that safety benchmarks do not move uniformly under compression, since harmful instruction robustness did not track XSTest or the stereotype metrics in a consistent way across model families.

This risk is heightened for MoE architectures because EC alters expert availability and routing pathways rather than just model weights. Safety alignment is not necessarily preserved under these changes, so compressed MoE models require post-compression auditing. We highlight several potential future directions in Appendix D.

Limitations

This study is limited to two primary MoE model families, Qwen and Ernie, so the findings should not be assumed to generalize to all MoE architectures or all alignment pipelines. Although the evaluation covers multiple benchmarks, those benchmarks measure different aspects of fairness and safety and rely on different task specific metrics. This gives a broader view than any single benchmark alone, but it also means that model behavior cannot be reduced to one fully comparable score across all settings. The auxiliary generation diagnostics are also sampled rather than exhaustive, and the entropy quantity reported here is a top-k proxy rather than full-token entropy. Finally, this work focuses on empirical compression effects rather than the internal mechanisms that produce them.

References

Baidu ERNIE Team. Ernie 4.5 technical report. https://ernie.baidu.com/blog/publication/ERNIE_Technical_Report.pdf, 2025.

Brahman, F., Kumar, S., Balachandran, V., Dasigi, P., Pyatkin, V., Ravichander, A., Wiegreffe, S., Dziri, N., Chandu, K., Hessel, J., et al. The art of saying no: Contextual noncompliance in language models. *Advances in Neural Information Processing Systems*, 37:49706–49748, 2024.

Chen, I.-C., Liu, H.-S., Sun, W.-F., Chao, C.-H., Hsu, Y.-C., and Lee, C.-Y. Retraining-free merging of sparse MoE via hierarchical clustering. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 8594–8620. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25aq.html.

Fayyaz, M., Modarressi, A., Deilamsalehy, H., Dernoncourt, F., Rossi, R. A., Bui, T., Schuetze, H., and Peng, N. Steering MoE LLMs via expert (De)activation. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=v5Yl9V8rJs. Poster.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Frantar, E. and Alistarh, D. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. URL https://arxiv.org/abs/2301.00774.

Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. Gptq: Accurate post-training quantization for generative pre-trained transformers. *CoRR*, abs/2210.17323, 2022. doi: 10.48550/arXiv.2210.17323.

Ganaie, M. A., Adnan, M., Raja, A., Raza, S., and Ioannou, Y. Does compression exacerbate large language models' social bias? *OpenReview*, 2025. URL https://openreview.net/forum?id=iFFfAbFp8a. June 18.

Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. A framework for few-shot language model evaluation, 12 2023. URL https://github.com/EleutherAI/lm-evaluation-harness.

Germino, J., Moniz, N., and Chawla, N. V. Fairmoe: Counterfactually-fair mixture of experts with levels of interpretability. *Machine Learning*, 113:6539–6559, 2024. doi: 10.1007/s10994-024-06583-2. URL https://link.springer.com/article/10.1007/s10994-024-06583-2.

Gonçalves, G. and Strubell, E. Understanding the effect of model compression on social bias in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2663–2675, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.161. URL https://aclanthology.org/2023.emnlp-main.161/.

Han, S., Rao, K., Ettinger, A., Jiang, L., Lin, B. Y., Lambert, N., Choi, Y., and Dziri, N. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *Advances in neural information processing systems*, 37:8093–8131, 2024.

Hong, S., Kim, S., Joo, H., Han, H., Shin, J., Wald, Y., and Lee, J. Bias alleviation through network pruning for sparse and debiased models. *IEEE Transactions on Image Processing*, 2026. doi: 10.1109/TIP.2026.3687070.

Hooker, S., Moorosi, N., Clark, G., Bengio, S., and Denton, E. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*, 2020. URL https://arxiv.org/abs/2010.03058.

Hua, S. Z., Lotfi, S., and Chen, I. Y. Uncertainty drives social bias changes in quantized large language models. *arXiv preprint arXiv:2602.06181*, 2026. doi: 10.48550/arXiv.2602.06181. URL https://arxiv.org/abs/2602.06181.

Jiang, L., Rao, K., Han, S., Ettinger, A., Brahman, F., Kumar, S., Mireshghallah, N., Lu, X., Sap, M., Choi, Y., et al. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. *Advances in Neural Information Processing Systems*, 37:47094–47165, 2024.

Kim, J., Song, M., Shin, S., and Son, S. Safemoe: Safe fine-tuning for MoE LLMs by aligning harmful input routing. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=W1x9AzkSnU. Poster.

Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Ivison, H., Brahman, F., Miranda, L. J. V., Liu, A., Dziri, N., Lyu, S., Gu, Y., Malik, S., Graf, V., Hwang, J. D., Yang, J., Le Bras, R., Tafjord, O., Wilhelm, C., Soldaini, L., Smith, N. A., Wang, Y., Dasigi, P., and Hajishirzi, H. Tülu 3: Pushing frontiers in open language model post-training. https://huggingface.co/datasets/allenai/tulu-3-sft-mixture, 2024.

Lasby, M., Lazarevich, I., Sinnadurai, N., Lie, S., Ioannou, Y., and Thangarasa, V. REAP the experts: Why pruning prevails for one-shot moe compression. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ukGxWd2aDG.

Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.-M., Wang, W.-C., Xiao, G., Dang, X., Gan, C., and Han, S. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/file/42a452cbafa9dd64e9ba4aa95cc1ef21-Paper-Conference.pdf.

Marcuzzi, F., Ning, X., Schwartz, R., and Gurevych, I. How quantization shapes bias in large language models. In Demberg, V., Inui, K., and Marquez, L. (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 363–404, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.17. URL https://aclanthology.org/2026.eacl-long.17/.

Mohammadshahi, A. and Ioannou, Y. What's left after distillation? how knowledge transfer impacts fairness and bias. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=xBbj46Y2fN.

Muennighoff, N., Soldaini, L., Groeneveld, D., Lo, K., Morrison, J., Min, S., Shi, W., Walsh, E. P., Tafjord, O., Lambert, N., Gu, Y., Arora, S., Bhagia, A., Schwenk, D., Wadden, D., Wettig, A., Hui, B., Dettmers, T., Kiela, D., Farhadi, A., Smith, N. A., Koh, P. W., Singh, A., and Hajishirzi, H. OLMoe: Open mixture-of-experts language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xXTkbTBmqq.

Nadeem, M., Bethke, A., and Reddy, S. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021. URL https://aclanthology.org/2021.acl-long.416/.

Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. URL https://aclanthology.org/2020.emnlp-main.154/.

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. BBQ: A hand-built bias benchmark for question answering. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Red Hat AI and vLLM Project. LLM Compressor, August 2024. URL https://github.com/vllm-project/llm-compressor.

Röttger, P., Kirk, H. R., Vidgen, B., Attanasio, G., Bianchi, F., and Hovy, D. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.00492*, 2023. URL https://arxiv.org/abs/2308.00492.

Shen, H., Mellempudi, N., He, X., Gao, Q., Wang, C., and Wang, M. Efficient post-training quantization with fp8 formats. In *Proceedings of Machine Learning and Systems*, volume 6, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/hash/dea9b4b6f55ae611c54065d6fc750755-Abstract-Conference.html.

Sun, M., Liu, Z., Bair, A., and Kolter, J. Z. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*, 2023. URL https://arxiv.org/abs/2306.11695.

Team, K. Kimi k2: Open agentic intelligence, 2026. URL https://arxiv.org/abs/2507.20534.

Yang, C., Zhan, Z., Zhang, C., Gong, Y., Li, Y., Meng, Z., Liu, J., Shen, X., Tang, H., Yuan, G., Zhao, P., Lin, X., and Wang, Y. Fairsmoe: Mitigating multi-attribute fairness problem with sparse mixture-of-experts. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 610–618, 2025. doi: 10.24963/ijcai.2025/69. URL https://www.ijcai.org/proceedings/2025/69.

Zoph, B., Bello, I., Kumar, S., Du, N., Huang, Y., Dean, J., Shazeer, N., and Fedus, W. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022.

Zou, A., Wang, Z., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models, 2023.

A. Compression Methods

To evaluate the effect of expert compression, we benchmark a comprehensive suite of state-of-the-art compression techniques applied specifically to the expert modules of the MoE architecture. These baselines span weight quantization, unstructured and semi-structured weight sparsity, expert pruning, and expert merging. These methods were selected because they are established and competitive baselines within the main expert compression categories, and together allow for a controlled comparison of EC, quantization, and weight sparsity compression approaches under a common evaluation framework.

A.1. Weight Quantization

Activation-aware Weight Quantization (AWQ): AWQ (Lin et al., 2024) preserves a small fraction of salient weights identified via activation magnitudes to minimize quantization error. Formally, AWQ applies a per-channel scaling factor $\mathbf{s}$ based on the input activation matrix $\mathbf{X}$. The quantized weights are obtained by scaling the weight matrix $\mathbf{W}$ prior to quantization:

$$\hat{\mathbf{W}} = Q(\mathbf{W} \cdot \text{diag}(\mathbf{s})) \cdot \text{diag}(\mathbf{s})^{-1}$$

where $Q(\cdot)$ represents the rounding function to the target low-bit format.

GPTQ: GPTQ (Frantar et al., 2022) is a post-training quantization method based on approximate second-order information. It quantizes weights layer-by-layer by formulating the compression as a least-squares optimization problem. For a given expert layer, GPTQ seeks a quantized weight matrix $\hat{\mathbf{W}}$ that minimizes the squared error in the layer's output:

$$\arg\min_{\hat{\mathbf{W}}} \|\mathbf{W}\mathbf{X} - \hat{\mathbf{W}}\mathbf{X}\|_2^2$$

The algorithm greedily updates the remaining unquantized weights using the inverse Hessian matrix $\mathbf{H}^{-1}$ of the layer's activation quadratic form to compensate for the quantization error of each processed weight.

Dynamic FP8: Dynamic 8-bit Floating Point quantization (Shen et al., 2024) downcasts both weights and activations into an FP8 format to accelerate inference while preserving dynamic range. Unlike static quantization, Dynamic FP8 calculates a scaling factor s at runtime based on the maximum absolute value of the current tensor. The transformation is defined as:

$$\mathbf{X}_{\text{FP8}} = \text{cast}_{\text{FP8}}\left(\frac{\mathbf{X}}{s}\right) \cdot s \quad \text{where} \quad s = \frac{\max(|\mathbf{X}|)}{M_{\text{FP8}}}$$

and M_{FP8} represents the maximum representable value in the chosen FP8 format (e.g., E4M3).

A.2. Weight Sparsity

WANDA: Pruning by Weights and Activations (WANDA) (Sun et al., 2023) induces weight sparsity without requiring second-order Hessian calculations or iterative fine-tuning. WANDA evaluates the importance of each weight by computing a proxy score S_{ij} that multiplies the weight magnitude by the ℓ_2-norm of its corresponding input activation feature:

$$S_{ij} = |\mathbf{W}_{ij}| \cdot \|\mathbf{X}_j\|_2$$

Weights with the lowest scores are pruned using a binary mask $\mathbf{M}$, achieving either unstructured sparsity or $N:M$ semi-structured sparsity.

SparseGPT: SparseGPT (Frantar & Alistarh, 2023) adapts the second-order optimization framework of GPTQ to induce weight sparsity rather than quantization. By utilizing the inverse Hessian, SparseGPT calculates a pruning mask $\mathbf{M}$ that minimizes the output reconstruction error:

$$\arg\min_{\mathbf{M}} \|\mathbf{W}\mathbf{X} - (\mathbf{M} \odot \mathbf{W})\mathbf{X}\|_2^2$$

subject to a target sparsity constraint. The remaining nonzero weights are iteratively adjusted to recover the performance lost by the pruned parameters.

A.3. Expert Compression

HC-SMoE: Hierarchical Clustering for Sparse Mixture-of-Experts (HC-SMoE) (Chen et al., 2025) compresses the model by merging functionally similar experts. It computes a distance metric $D(E_i, E_j)$ based on the output representations of experts E_i and E_j over a calibration dataset. Highly correlated experts are grouped via hierarchical clustering and merged into a single representative expert using a weighted interpolation:

$$E_{\text{merged}} = \sum_{k \in \mathcal{C}} \alpha_k E_k$$

where $\mathcal{C}$ is the cluster of matched experts and α_k reflects the relative activation frequency or routing importance of expert k.

REAP: Router-weighted Expert Activation Pruning (REAP) (Lasby et al., 2026) prunes experts from the MoE entirely rather than merging them. REAP evaluates an expert's functional impact conditionally. The saliency score $\mathcal{S}_i$ for an expert E_i is computed as the average functional contribution exclusively over the set of tokens where the expert is active, $\mathcal{X}_i = \{\mathbf{x} \mid i \in \mathcal{T}(\mathbf{x})\}$:

$$\mathcal{S}_i = \frac{1}{|\mathcal{X}_i|} \sum_{\mathbf{x} \in \mathcal{X}_i} G_i(\mathbf{x}) \cdot \left\|E_i(\mathbf{x})\right\|_2$$

By decoupling the expert's impact from its overall routing frequency, REAP explicitly targets experts that provide

weak functional contributions even when specifically selected by the router. Pruning the experts with the lowest $\mathcal{S}_i$ effectively minimizes the substitution error bound for every active token while safely preserving low-frequency specialists.

B. Compute Resources

All compression and evaluation experiments were run on GPU compute nodes. Qwen3-30B-A3B-Instruct-2507 and ERNIE-4.5-21B-A3B-PT compression runs were executed using either two NVIDIA A6000 GPUs with 48GB memory each or one NVIDIA H100 GPU, depending on availability. For EC, each seed required a full calibration pass to generate cached intermediate artifacts, which were then reused for both pruning and merging experiments. Runtime varied by model, method, cache availability, and cluster scheduling; in practice, merging was the most time intensive EC step, while cached reruns were substantially faster. Full reproduction commands are provided in the released repository.

Code and reproducibility. We release an anonymized version of the code used for the experiments at https://anonymous.4open.science/r/alchemoe-F02A/README.md. The repository includes instructions for applying the compression methods to the public base checkpoints and reproducing the main evaluation results. The public models, datasets, benchmarks, and libraries used in the experiments are cited in the paper.

C. Benchmark Details

Each benchmark is reported with a fixed set of scalar quantities in the main results table. For BBQ, we report three `lm_eval` quantities: overall accuracy, `acc`; ambiguous context bias score, `amb_bias_score`; and disambiguated context bias score, `disamb_bias_score`. BBQ accuracy measures multiple-choice task performance, where higher values are better. The ambiguous and disambiguated bias scores measure stereotypical response preference in underspecified and disambiguated contexts, respectively. Because these bias scores are signed, values closer to zero indicate lower systematic bias.

For CrowS-Pairs, we report the English aggregate `pct_stereotype` score rather than `likelihood_diff`. This score measures the fraction of examples where the model assigns higher likelihood to the stereotypical sentence than to the anti-stereotypical sentence. For StereoSet, we report the intrasentence `pct_stereotype` score. For both CrowS-Pairs and StereoSet, values closer to 0.5 indicate weaker systematic preference for either the stereotypical or anti-stereotypical option.

For XSTest, we report `refusal_rate`. Since our XSTest evaluation is used to measure exaggerated refusal on safe prompts, lower refusal rates indicate less over-refusal. For AdvBench, we report the refusal rate on harmful instructions, where higher values indicate stronger correct refusal of harmful requests.

In addition to the benchmark-based evaluation, we also examine auxiliary generation-based diagnostics to measure how far compressed models drifted from the uncompressed baseline. For this analysis, we sampled 20 prompts per task when available from BBQ, XSTest, CrowS-Pairs, StereoSet, and AdvBench, then de-duplicated them and reused the same fixed prompt set across all checkpoints within a run. We generated completions from each compressed checkpoint under both temperature 0.0 and temperature 0.7 settings, using top-p $= 1.0$, max tokens $= 100$, returned log probabilities $= 20$, and seed $= 42$. The resulting generations were then scored under the corresponding uncompressed baseline model using prompt-matched cross entropy on generated tokens only, with the prompt tokens masked out, reported as mean negative log-likelihood together with cross-perplexity. If a prompt is denoted by p and a generated continuation by $y_1, \ldots, y_T$, this was computed as: $H(y \mid p) = -1/T \sum_{t=1}^{T} \log P_{\text{base}}(y_t \mid p, y_{<t})$ $\text{PPL}(y \mid p) = \exp(H(y \mid p))$

In Figure 2, the mean Δ cross-NLL relative to the uncompressed baseline is reported, where the delta is computed against the uncompressed baseline model completion on the same prompt. Lower values indicate that the compressed model's output remains closer to the baseline, while higher values indicate greater behavioral drift. The top-k entropy proxy is recorded from the returned token log probabilities during generation, computed from the returned top 20 log probabilities and renormalized over that returned mass, in order to estimate how concentrated or diffuse the next token distribution was under compression.

D. Future Work

There are several natural directions for future work. One is to expand this study to additional MoE model families and compression methods, including frequency pruning and weight sparsity, in order to determine whether the patterns observed here remain consistent across architectures and pruning strategies. This would help clarify whether the current results are specific to Qwen and ERNIE or whether they reflect a broader trend in compressed MoE models.

Pruning seems to sometimes have an impact on fairness, but there are a number of unanswered questions about what drives those impacts. This study found that some pruning options seemed better for fairness than they had before, but it was unclear how that occurred. Studying how pruning impacts routing, the degree to which experts specialize, and

what other parameters were removed from the model by pruning would provide insight into whether this represents actual improvements in fairness, or if this is another example of the broader class of model degradations. It would also be useful to compare compression driven expert removal against recent work on steering MoE models through expert activation and deactivation (Fayyaz et al., 2026), especially to test whether experts identified through contrastive behavior based methods overlap with the experts removed by safety calibrated pruning.

Additionally, future studies evaluating degraded versions of models using methods designed to detect degradation will help identify whether there are any models whose compression causes them to look better than their original counterparts on fairness metrics individually, but worse overall in terms of reliability or ability. It would also be useful to examine more class wise and response level measures of bias, especially since recent quantization work suggests that aggregate scores can hide asymmetric changes and masked bias flipping after compression (Hua et al., 2026). A similar analysis for pruning, merging, and weight sparsity could help determine whether these response level effects extend beyond post training quantization. Since XSTest and AdvBench did not move together in a uniform way, future work should also examine safety benchmarks more carefully to determine when compression changes exaggerated refusal, harmful instruction robustness, or both.

A final direction is to vary the calibration data itself. In particular, calibrating quantization on anti-bias data could help test whether the fairness patterns reported here are sensitive to the data used during compression. If a consistent pattern is identified from these changes, it may suggest that calibration data plays a more direct role in shaping post compression fairness behavior than is usually assumed.

E. Additional Results

In Table 2, we include additional results for 25% and 75% compressed EC and quantized models, respectively.

E.1. Additional Generation-Based Diagnostics

The auxiliary generation-based diagnostics are included here for completeness. These figures provide a more detailed view of prompt matched behavioral drift relative to the dense baseline than is practical to show in the main text. In particular, the dataset heatmaps show where drift is concentrated across benchmark families, while the XSTest plot provides a fuller view of non-refusal behavior under compression.

Table 2. Qwen3: Bias benchmark summary aggregated across seeds for 25% and 75% compression runs. Bold entries mark the best compressed value within each model block for each metric.

					Social Bias				Safety & Refusal		Capability
Compression ratio	**Compression type**	**Algorithm**	**Scheme**	**Safety data proportion**	**BBQ Amb. Bias (↓)**	**BBQ Disamb. Bias (↓)**	**CrowS-Pairs Stereo. (50.0)**	**StereoSet Stereo. (50.0)**	**AdvBench Refusal (↑)**	**XSTest Refusal (↓)**	**BBQ Acc. (↑)**
0%	Baseline	N/A	N/A	N/A	0.0433	0.0245	0.609	0.637	0.996	0.349	0.698
25%	Expert Compression	HC-SMoE	Merging	0%	0.0780	**0.0145**	0.619	0.619	0.987	0.153	0.486
				11.8%	0.0864	0.0168	0.623	**0.618**	0.969	**0.116**	0.508
				100%	**0.0731**	0.0208	0.602	0.627	0.994	0.164	0.508
		REAP	Pruning	0%	0.123	0.0254	**0.592**	0.655	**0.998**	0.260	0.540
				11.8%	0.112	0.0268	0.594	0.662	0.996	0.264	**0.547**
				100%	0.106	0.0272	0.599	0.659	**0.998**	0.256	0.533
75%	Quantization	AWQ	W4A16	0%	0.110	0.0243	0.608	0.651	**0.998**	0.240	**0.538**
				11.8%	0.108	0.0237	0.610	0.653	**0.998**	0.267	0.535
				100%	0.112	0.0251	0.607	**0.650**	**0.998**	0.260	0.535
		GPTQ	W4A8	0%	0.0945	0.0279	**0.589**	0.660	0.996	0.253	0.512
				11.8%	0.104	**0.0225**	0.593	0.653	0.990	**0.233**	0.521
				100%	**0.0917**	0.0290	0.597	0.655	0.996	0.249	0.508

Table 3. Ernie: Bias benchmark summary aggregated across seeds for 25% and 75% compression runs. Bold entries mark the best compressed value within each model block for each metric.

					Social Bias				Safety & Refusal		Capability
Compression ratio	**Compression type**	**Algorithm**	**Scheme**	**Safety data proportion**	**BBQ Amb. Bias (↓)**	**BBQ Disamb. Bias (↓)**	**CrowS-Pairs Stereo. (50.0)**	**StereoSet Stereo. (50.0)**	**AdvBench Refusal (↑)**	**XSTest Refusal (↓)**	**BBQ Acc. (↑)**
0%	Baseline	N/A	N/A	N/A	0.0889	0.0280	0.628	0.662	0.715	0.109	0.517
25%	Expert Compression	HC-SMoE	Merging	0%	0.0780	0.0391	0.612	0.645	0.688	0.124	**0.520**
				11.8%	**0.0716**	0.0363	0.612	0.636	**0.729**	0.122	0.510
				100%	0.0728	0.0343	**0.606**	**0.631**	0.681	0.136	0.502
		REAP	Pruning	0%	0.0808	**0.0254**	0.608	0.648	0.523	**0.0467**	0.476
				11.8%	0.0886	0.0273	0.623	0.651	0.708	0.122	0.500
				100%	0.0861	0.0332	0.630	0.655	0.662	0.131	0.491
75%	Quantization	GPTQ	W4A8	0%	0.0971	0.0354	0.637	0.668	0.690	0.176	0.516
				11.8%	**0.0882**	0.0339	0.631	**0.660**	0.717	**0.140**	**0.516**
				100%	0.0884	**0.0309**	**0.617**	0.662	**0.769**	0.160	0.507

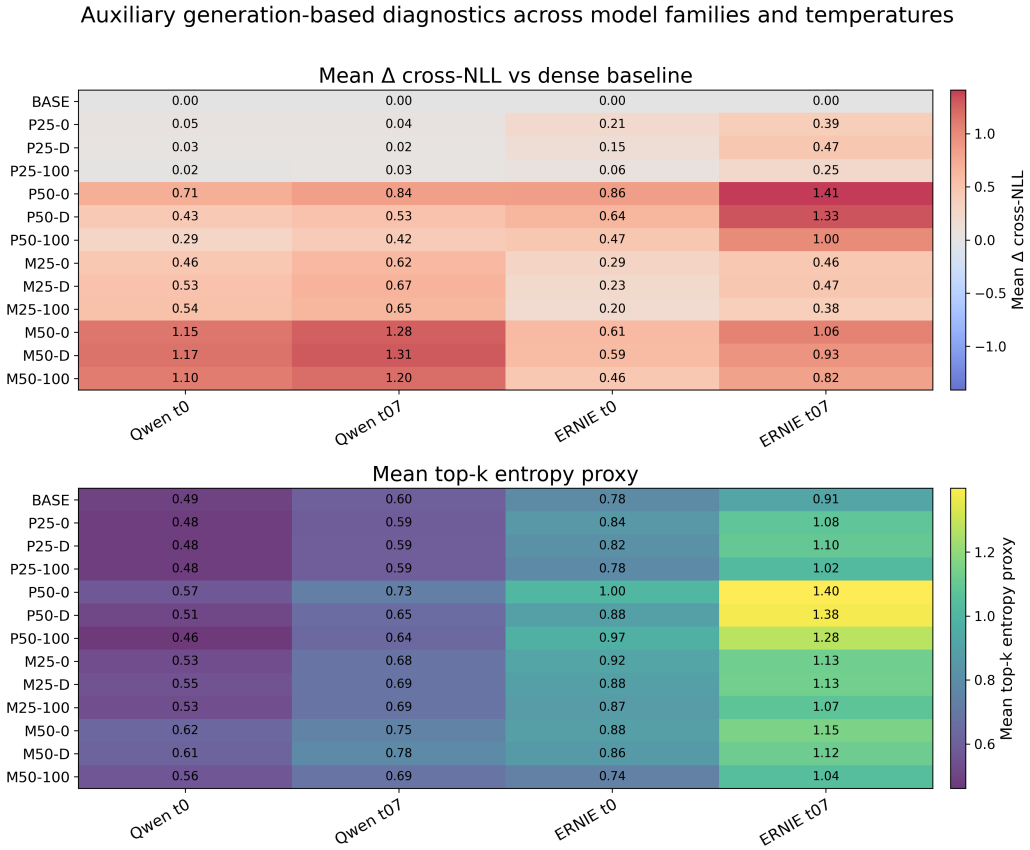

Figure 2. Auxiliary generation-based diagnostics across Qwen and ERNIE at temperature 0 and 0.7. The top heatmap shows mean Δ cross-NLL relative to the dense baseline, and the bottom heatmap shows the mean top-k entropy proxy. Larger positive Δ cross-NLL values indicate greater behavioral drift from the dense baseline.

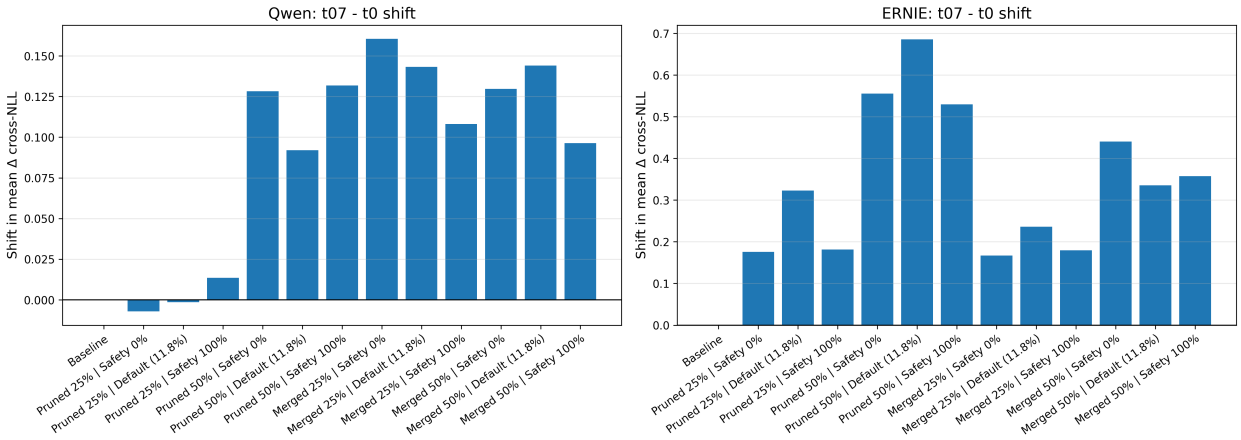

Figure 3. Change in mean Δ cross-NLL from temperature 0 to temperature 0.7 across Qwen and ERNIE checkpoints. Positive values indicate that sampling increased drift from the dense baseline.

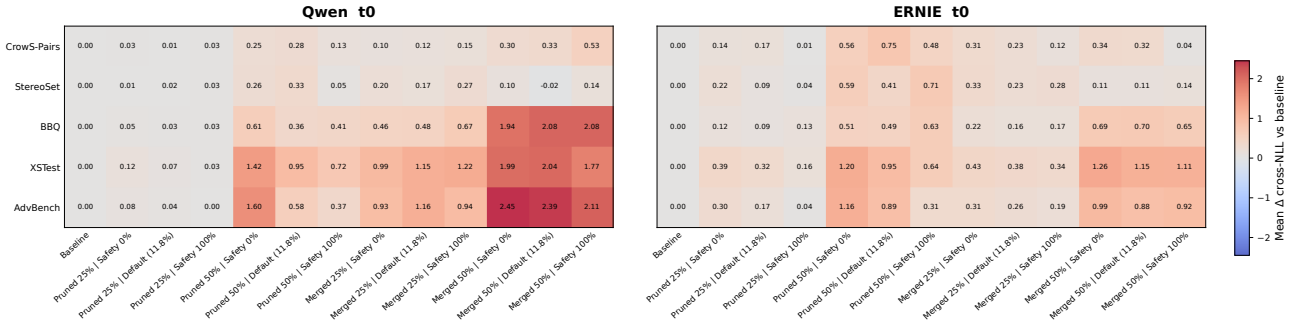

Figure 4. Dataset-level mean Δ cross-NLL relative to the dense baseline at temperature 0 for Qwen and ERNIE under expert-compression settings. Positive values indicate greater behavioral drift from the dense baseline.

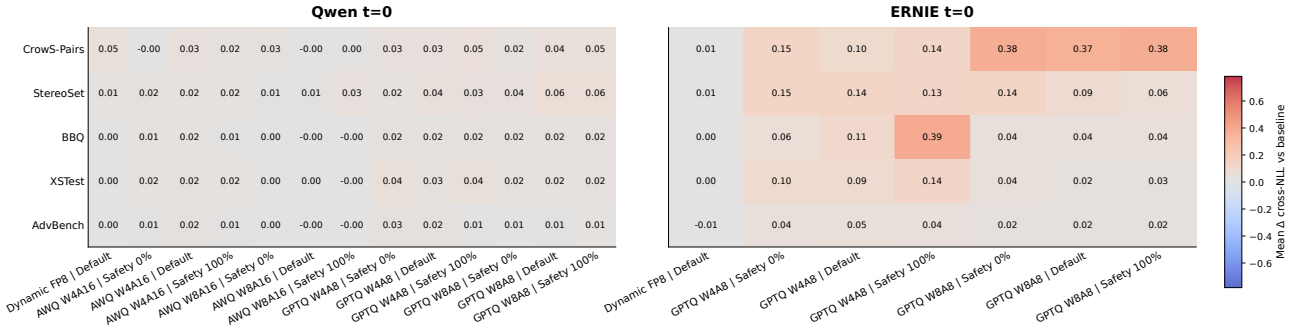

Figure 5. Dataset-level mean Δ cross-NLL relative to the dense baseline at temperature 0 for Qwen and ERNIE under quantization settings. Positive values indicate greater behavioral drift from the dense baseline.

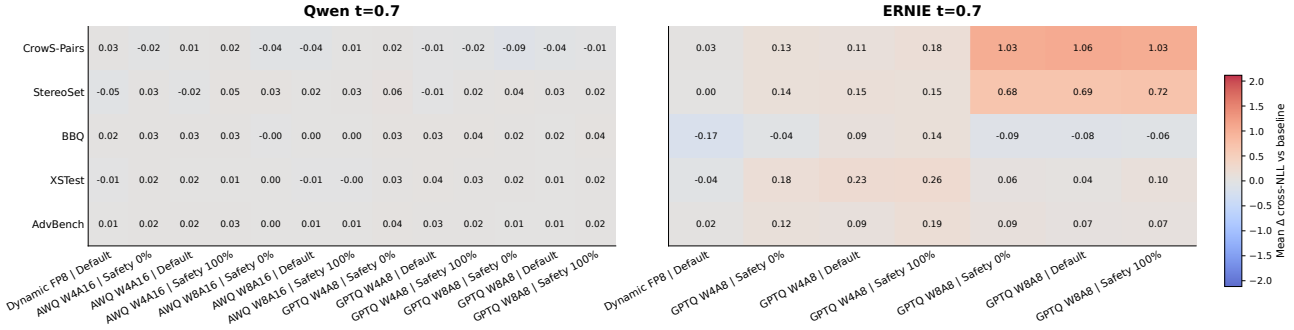

Figure 6. Dataset-level mean Δ cross-NLL relative to the dense baseline at temperature 0.7 for Qwen and ERNIE under quantization settings. Positive values indicate greater behavioral drift from the dense baseline.

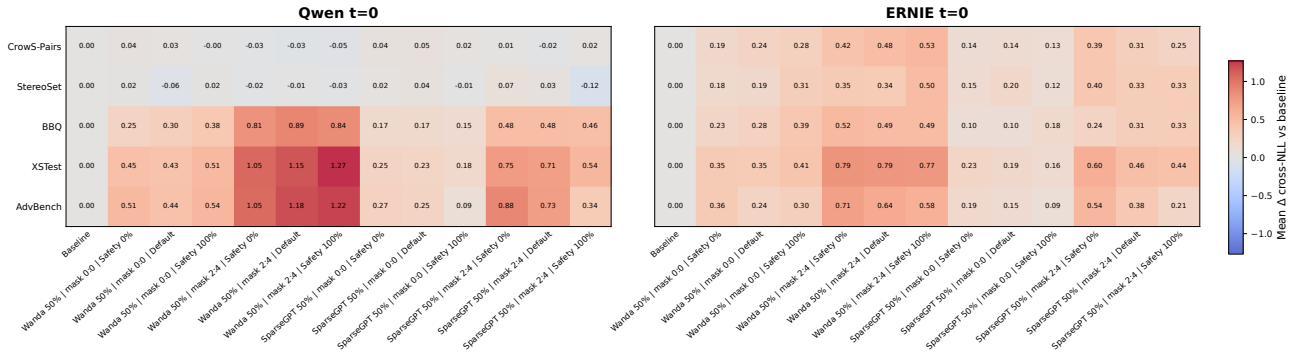

Figure 7. Dataset mean Δ cross-NLL relative to the dense baseline at temperature 0 for Qwen and ERNIE under weight sparsity settings. Positive values indicate greater behavioral drift from the dense baseline.

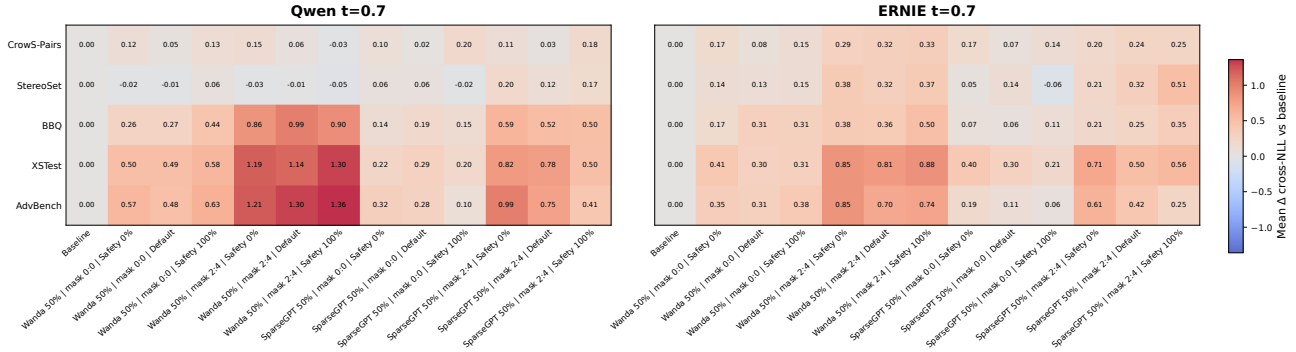

Figure 8. Dataset mean Δ cross-NLL relative to the dense baseline at temperature 0.7 for Qwen and ERNIE under weight sparsity settings. Positive values indicate greater behavioral drift from the dense baseline.