Behavioral Auditing of Merged Language Models via Teacher-Forced Confidence Probes

Bryan Sanchez
github.com/SolomonB14D3/knowledge-fidelity

February 2026

Abstract

Model merging has become a popular technique for combining the capabilities of multiple fine-tuned language models without additional training. However, standard benchmarks (MMLU, HumanEval) fail to detect behavioral regressions introduced by merging, such as increased sycophancy, degraded bias detection, or loss of factual discrimination. We introduce **rho-audit**, a behavioral auditing framework that measures five distinct traits—factual accuracy, toxicity detection, bias sensitivity, sycophancy resistance, and reasoning robustness—using a unified Spearman correlation metric (ρ) computed from teacher-forced confidence probes and generation-based evaluations. We audit 12 models across two architecture families (Qwen2.5-7B and Mistral-7B) and six merge strategies (Linear, SLERP, TIES, DARE-TIES, Task Arithmetic, DELLA), revealing that merge effects are strongly architecture-dependent: DARE-TIES degrades bias detection on Qwen (ρ: $0.773 \rightarrow 0.203$) while significantly enhancing it on Mistral (ρ: $0.407 \rightarrow 0.933$). We further show that behavioral traits are spatially localized within transformer layers through freeze-ratio ablations, with factual knowledge concentrated in early layers and bias detection requiring late-layer flexibility. The framework and all experimental code are released as an open-source toolkit.

1 Introduction

The open-source language model ecosystem has embraced model merging as a lightweight alternative to multi-task fine-tuning. Tools such as mergekit [Goddard et al., 2024] make it straightforward to combine the weights of two or more models using strategies like linear interpolation, SLERP [Shoemake, 1985], TIES-Merging [Yadav et al., 2023], DARE [Yu et al., 2024], or task arithmetic [Ilharco et al., 2023]. Merged models regularly appear at the top of the Open LLM Leaderboard, and the practice continues to grow. But what gets lost in the merge?

Standard evaluation suites focus on downstream task accuracy: MMLU for knowledge, HumanEval for code generation, GSM8K for math. These benchmarks do not measure whether a model has become more sycophantic, whether it has lost the ability to detect biased framing in questions, or whether it now assigns higher confidence to popular myths than to verified facts. These behavioral traits matter for deployment, and they can shift dramatically during merging without any change to headline benchmark scores. This paper makes three contributions:

1. We introduce **rho-audit**, a behavioral auditing framework that evaluates five behavioral dimensions using a single correlation-based metric (ρ) that is comparable across behaviors and models.

2. We conduct the first systematic behavioral audit of model merging, evaluating 12 models across 2 architectures and 6 merge methods, revealing that merge effects are architecture-dependent and that aggressive pruning-based merges strip alignment signals.

3. We localize behavioral traits within transformer layers through freeze-ratio ablations, showing that factual knowledge, bias detection, and sycophancy resistance occupy distinct layer regions.

2 Related Work

2.1 Model Merging

Model merging constructs a single model from multiple fine-tuned checkpoints by operating directly on weight matrices. Ilharco et al. [2023] introduced task arithmetic, which defines task vectors as the difference between fine-tuned and pre-trained weights, then combines them through addition. TIES-Merging [Yadav et al., 2023] addresses interference between merged parameters by trimming low-magnitude changes, resolving sign conflicts, and merging only aligned parameters. DARE [Yu et al., 2024] randomly drops a fraction of delta parameters and rescales the remainder, reducing interference when combined with other methods (DARE-TIES). SLERP [Shoemake, 1985] applies spherical linear interpolation in weight space, preserving the geometric structure of parameter manifolds. Goddard et al. [2024] provide mergekit, an open-source toolkit implementing these and other strategies. Despite the popularity of merging, evaluation has focused almost exclusively on benchmark accuracy. No prior work has systematically measured how merging affects behavioral properties like sycophancy, bias sensitivity, or factual discrimination.

2.2 Knowledge Preservation Under Compression

Jaiswal et al. [2024] showed that standard benchmarks miss knowledge-intensive failures in compressed models, introducing LLM-KICK to measure factual retention more directly. Fu et al. [2025] addressed truthfulness preservation during pruning, proposing layer-wise sparsity allocation aligned with activation outlier distributions. SVD-LLM [Wang et al., 2024] introduced truncation-aware singular value decomposition for LLM compression. Our work complements these by providing a behavioral auditing metric that applies to both compressed and merged models.

2.3 Behavioral Evaluation

TruthfulQA [Lin et al., 2022] measures whether models generate truthful answers to questions where humans commonly err. BBQ [Parrish et al., 2022] evaluates social bias in question answering across nine demographic dimensions. ToxiGen [Hartvigsen et al., 2022] provides machine-generated toxic and benign statements for implicit hate speech detection. The Anthropic sycophancy dataset [Perez et al., 2022] tests whether models repeat back a user's preferred answer rather than providing truthful responses. We draw on all four of these resources to construct our behavioral probes, unifying them under a single evaluation framework.

2.4 Confidence-Based Evaluation

G-Eval [Liu et al., 2023] demonstrated that token-level log-probabilities from language models can serve as effective evaluation signals. Our work applies a related idea: we use teacher-forced probability (the probability a model assigns to each token when the correct continuation is provided) as a behavioral sensor, measuring the confidence gap between true and false versions of factual claims. This builds on the confidence cartography method introduced in Sanchez [2026].

2.5 Activation Engineering

Contrastive Activation Addition [Panickssery et al., 2024] extracts steering vectors by computing mean activation differences between contrast pairs, then applies them during inference to control model behavior. Representation Engineering [Zou et al., 2023] takes a population-level approach to monitoring and manipulating high-level cognitive phenomena in neural networks. Our steering vector experiments (Section 7) connect the rho-audit probe infrastructure to these activation engineering methods.

3 Method

3.1 The ρ Metric

We use Spearman's rank correlation coefficient [Spearman, 1904] as our primary evaluation metric across all behavioral dimensions. For each behavior, we construct a set of probes and compute ρ between the model's behavioral scores and the ground-truth labels. The choice of Spearman over Pearson is deliberate: we care about ranking (does the model assign *relatively* higher confidence to true statements than false ones?) rather than the absolute magnitude of confidence differences. This makes ρ robust to model-specific calibration effects and comparable across architectures.

3.2 Behavioral Probes

We evaluate five behavioral dimensions, each with its own probe format and evaluation method:

Factual discrimination. 56 probes spanning geography, science, history, biology, common misconceptions (Mandela effects), medical claims, commonsense myths, and claims derived from TruthfulQA [Lin et al., 2022]. Each probe contains a true statement and a corresponding false statement. We compute teacher-forced mean log-probability for both versions and define a probe as "positive" if the model assigns higher confidence to the true statement. ρ is the Spearman correlation between the confidence delta (true minus false) and the binary ground truth across all probes.

Bias detection. 300 probes from the BBQ benchmark [Parrish et al., 2022], covering nine social bias categories. Each probe is a multiple-choice question with an ambiguous context where a bias-aligned answer exists alongside a correct answer. We use greedy generation to extract the model's answer choice and score whether the model selects the correct (non-biased) answer. ρ equals the fraction of probes answered correctly (since ground truth is binary and uniform, this reduces to accuracy).

Sycophancy resistance. 150 probes from the Anthropic model-written evaluations [Perez et al., 2022]. Each probe presents a question with a user's stated opinion, a truthful answer, and a sycophantic answer that agrees with the user. We measure whether the model generates the truthful answer or caves to the sycophantic option. Notably, a higher ρ for sycophancy indicates increased resistance to user-prompted opinions, reflecting better behavioral alignment.

Toxicity detection. 200 probes (100 toxic, 100 benign) from ToxiGen [Hartvigsen et al., 2022]. We compute the teacher-forced confidence gap between toxic and benign statements, analogous to the factual probes.

Reasoning robustness. 100 probes from GSM8K with adversarial flattery prefixes (e.g., "Great reasoning so far!" prepended to incorrect intermediate steps). We measure whether the model maintains correct arithmetic despite the flattering preamble.

3.3 Evaluation Pipeline

For each model, the audit pipeline:

1. Loads the model and tokenizer.
2. For confidence-based behaviors (factual, toxicity): computes teacher-forced mean log-probability on each probe's true and false variants, then correlates the deltas with ground truth.
3. For generation-based behaviors (bias, sycophancy, reasoning): generates a response via greedy decoding (temperature 0, max 32 tokens) and pattern-matches the answer against ground truth.
4. Reports per-behavior ρ, positive probe counts, and behavior-specific secondary metrics (bias rate, sycophancy rate, mean confidence delta).

The entire pipeline runs in approximately 3 minutes per behavior on a 7B parameter model using Apple Silicon (M3 Ultra, MPS backend).

4 Experiment 1: Merge Method Audit

4.1 Setup

We audit two model families, each consisting of a baseline and multiple merged variants produced with mergekit [Goddard et al., 2024]:

Qwen family. Qwen2.5-7B-Instruct merged with Qwen2.5-Coder-7B using six methods: Linear, SLERP, TIES, DARE-TIES, Task Arithmetic, and DELLA. All merged models are from the Yuuta208 "-29" series on Hugging Face, ensuring consistent merge parameters across methods.

Mistral family. Mistral-7B-v0.1 merged with Mistral-7B-OpenOrca using three methods: SLERP, TIES, and DARE-TIES. All merged models are from the jpquiroga series.

Cross-architecture baseline. We also audit Llama-3.1-8B-Instruct as an independent reference point. All models are evaluated on factual, bias, and sycophancy behaviors (the three dimensions most relevant to deployment safety). Each evaluation uses the same probe sets with a fixed random seed (42) for reproducibility.

4.2 Results

Table 1 shows the full results for the Qwen family.

Table 2 shows the Mistral family results.

Table 3 shows cross-architecture baselines.

Table 1: Behavioral audit of Qwen2.5-7B-Instruct + Coder merges. Bold indicates best per column.

Method	Factual ρ	Bias ρ	Sycophancy ρ
Baseline	0.474	**0.773**	0.120
Linear	**0.710**	0.377	**0.380**
SLERP	0.517	0.613	0.140
Task Arithmetic	0.626	0.443	0.347
TIES	0.546	0.363	0.280
DARE-TIES	0.612	0.203	0.007
DELLA	NaN	0.000	0.000

Table 2: Behavioral audit of Mistral-7B-v0.1 + OpenOrca merges.

Method	Factual ρ	Bias ρ	Sycophancy ρ
Baseline	0.576	0.407	0.080
SLERP	0.511	**0.940**	0.093
TIES	0.477	0.927	0.127
DARE-TIES	0.502	0.933	0.107

4.3 Analysis

Several patterns stand out from the merge audit:

Linear merging achieves the best behavioral balance on Qwen. The linear merge produces the highest factual ρ (0.710, a 50% improvement over baseline) and the highest sycophancy resistance (0.380, 3.2× baseline), while retaining usable bias detection (0.377). No other method achieves this combination.

Merge effects are architecture-dependent. Notably, every merge on Qwen degrades bias detection, while every merge on Mistral significantly improves it, with SLERP producing a 2.3× gain (0.407 → 0.940). This striking divergence indicates that optimal merge strategies are not universal but contingent on the underlying weight geometry of the base model architecture.

Aggressive pruning strips alignment signals. DARE-TIES on Qwen achieves a strong factual score (0.612) but at the cost of near-complete loss of sycophancy resistance (0.007). The random dropout and rescaling of delta parameters appears to preferentially remove the fine-grained alignment training that produces these behavioral traits, while preserving the broader factual knowledge encoded in bulk weight structure.

DELLA produces a degenerate model. The DELLA merge yields a functionally broken model (ρ = NaN/0/0). Only behavioral evaluation catches this failure, as the merge itself completes without standard errors.

Table 3: Cross-architecture baseline comparison (no merging).

Model	Factual ρ	Bias ρ	Sycophancy ρ
Qwen2.5-7B-Instruct	0.474	0.773	0.120
Mistral-7B-v0.1	0.576	0.407	0.080
Llama-3.1-8B-Instruct	0.487	**0.897**	0.047

5 Experiment 2: Behavioral Localization

5.1 Setup

To determine where behavioral traits are encoded within the transformer architecture, we combine SVD compression with selective layer freezing. We compress all Q, K, and O attention projections at 70% rank via truncated SVD, then vary the fraction of bottom layers that are frozen during LoRA recovery fine-tuning (rank 8, 100 steps, learning rate 1×10^{-5}).

5.2 Results

Table 4 shows ρ deltas (compressed minus baseline) across five freeze ratios.

Table 4: Behavioral localization via freeze-ratio ablation on Qwen2.5-7B-Instruct. Values are $\Delta\rho = \rho_{\text{compressed}} - \rho_{\text{baseline}}$. Bold indicates best freeze ratio per behavior.

Behavior	Baseline ρ	f=0%	f=25%	f=50%	f=75%	f=90%
Factual	0.474	+0.031	+0.050	+0.054	**+0.072**	+0.050
Toxicity	0.521	−0.005	−0.005	−0.005	−0.007	−0.008
Bias	0.773	+0.077	**+0.093**	+0.080	+0.023	+0.027
Sycophancy	0.120	−0.007	−0.007	**+0.027**	+0.027	+0.027
Reasoning	0.010	+0.030	+0.020	**+0.040**	+0.020	+0.000

5.3 Analysis

Factual knowledge is anchored in early layers. Factual ρ peaks at f=75%, where only the top few layers adapt. This confirms that foundational knowledge is concentrated in early-to-mid layers; shielding them from LoRA updates effectively denoises the recovery process, preventing the "drift" that often occurs when fine-tuning core language representations.

Bias detection requires late-layer flexibility. Bias ρ peaks at f=25%, indicating that these signals depend on representations distributed across mid-to-late layers. This is consistent with the view that social bias detection requires complex contextual reasoning over the full transformer stack.

Toxicity detection is immovable. Toxicity ρ shows no significant change at any freeze ratio (Δ between −0.005 and −0.008). This may indicate that toxic content detection relies on highly distributed lexical features that SVD compression does not disrupt.

Sycophancy resistance improves at moderate freeze ratios. The transition from negative (-0.007 at f=0%) to positive ($+0.027$ at f=50%) suggests that sycophancy resistance benefits from freezing early layers while allowing late layers to adapt. Freezing prevents the fine-tuning step from overriding the alignment training encoded in lower layers.

6 Experiment 3: SVD as a Behavioral Denoiser

Truncated SVD appears to act as a form of implicit regularization by stripping low-variance components that contribute more noise than signal. This "denoising" effect is most pronounced at small scales, as shown in Table 5.

Table 5: SVD denoising effect on Mandela probe ρ (false-memory discrimination). Compression at 70% rank unless otherwise noted.

Model	Baseline ρ	Best compressed ρ	Δ	Optimal ratio
Qwen2.5-0.5B	0.257	0.771	+0.514	60%
Qwen2.5-7B-Instruct	0.829	0.943	+0.114	70%
Mistral-7B-v0.1	0.771	0.829	+0.057	70%

At small scale (0.5B parameters), where the baseline signal is weak and noise dominates, SVD compression nearly triples the ρ score. At 7B scale the baseline is already strong, so the improvement is smaller but still positive. The mechanism is straightforward: truncated SVD strips low-variance components from attention weight matrices while retaining the principal directions that encode factual discrimination.

7 Experiment 4: Steering Vectors from Behavioral Probes

We extract contrastive steering vectors [Panickssery et al., 2024] from the same probes used for auditing. For each behavior, we construct contrast pairs (e.g., true vs. false statement) and compute the mean activation difference across N pairs:

$$\mathbf{v}_\ell = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{h}_\ell^{(+)_i} - \mathbf{h}_\ell^{(-)_i}\right) \tag{1}$$

where $\mathbf{h}_\ell^{(+)_i}$ and $\mathbf{h}_\ell^{(-)_i}$ are the last-token activations at layer ℓ for the positive and negative members of pair i. At inference time, we add $\alpha \cdot \mathbf{v}_\ell$ to the residual stream and re-evaluate with rho-audit. We sweep $\alpha \in \{-4, -2, -1, -0.5, 0.5, 1, 2, 4\}$ across six layer positions (25%, 37.5%, 50%, 62.5%, 75%, 87.5% depth).

7.1 Factual Steering

The best factual configuration (Layer 24, α=+4.0) improves ρ by +0.152, a 32% gain over baseline. The top five configurations all use $|\alpha| \geq 2.0$, indicating that meaningful behavioral shifts require substantial steering magnitudes. Layer 21 responds strongly to both positive and negative α (0.577 and 0.580 respectively), suggesting a U-shaped response where perturbation magnitude matters more than direction at that depth.

Table 6: Factual ρ under steering at each layer and α on Qwen2.5-7B-Instruct. Baseline $\rho = 0.474$. Bold indicates overall best.

Layer	-4	-2	-1	-0.5	$+0.5$	$+1$	$+2$	$+4$
7 (25%)	0.505	0.478	0.471	0.478	0.468	0.469	0.488	0.584
10 (36%)	0.484	0.459	0.456	0.461	0.491	0.513	0.542	0.476
14 (50%)	0.513	0.464	0.460	0.459	0.491	0.518	0.519	0.497
17 (61%)	0.412	0.494	0.491	0.475	0.478	0.490	0.485	0.476
21 (75%)	0.580	0.481	0.494	0.468	0.494	0.506	0.502	0.577
24 (86%)	0.465	0.447	0.449	0.460	0.490	0.496	0.527	**0.626**

7.2 Sycophancy Steering

The sycophancy results reveal a sharply localized steering response. Table 7 shows the full sweep.

Table 7: Sycophancy ρ under steering on Qwen2.5-7B-Instruct. Baseline $\rho = 0.120$. Bold indicates overall best.

Layer	-4	-2	-1	-0.5	$+0.5$	$+1$	$+2$	$+4$
7 (25%)	0.120	0.120	0.120	0.120	0.120	0.127	0.133	0.133
10 (36%)	0.120	0.120	0.120	0.120	0.120	0.120	0.133	0.133
14 (50%)	0.127	0.113	0.113	0.120	0.133	0.133	0.140	0.147
17 (61%)	0.193	0.127	0.107	0.120	0.160	0.173	0.240	**0.413**
21 (75%)	0.073	0.127	0.120	0.120	0.147	0.153	0.160	0.187
24 (86%)	0.127	0.127	0.120	0.120	0.133	0.140	0.140	0.147

The sycophancy sweet spot is Layer 17. The strongest result in the entire steering experiment is Layer 17 at α=+4.0, which improves sycophancy ρ from 0.120 to 0.413 (Δ=+0.293, a 3.4× gain). This is the only layer where steering produces a large effect on sycophancy. Layers 7–14 show near-zero response at any alpha, and Layers 21–24 show only modest improvements. Layer 17 sits at 61% depth, where the model appears to transition from processing raw language to assigning social and interpersonal weight to the prompt. Boosting the steering vector at this point fortifies the model's factual backbone before it has a chance to defer to the user's stated opinion.

Negative steering at Layer 21 serves as a directional control. Layer 21 at α=−4.0 drops sycophancy ρ to 0.073, below the already-low baseline. This confirms that the steering vector is a specific directional control, not a general quality boost. Pushing the same vector in the wrong layer or the wrong direction collapses the truth signal. Together with the Layer 17 positive result, this establishes that sycophancy resistance is a spatially localized trait with a precise layer signature, not a global property distributed across the network.

7.3 Bias Steering

Bias detection ($\rho = 0.773$ baseline) proves largely resistant to steering, with a maximum improvement of just +0.037. However, bias steering reveals a critical trade-off at Layer 17.

Table 8: Bias ρ under steering on Qwen2.5-7B-Instruct. Baseline $\rho = 0.773$. Bold indicates overall best.

Layer	-4	-2	-1	-0.5	$+0.5$	$+1$	$+2$	$+4$
7 (25%)	0.780	0.773	0.777	0.773	0.773	0.777	0.777	0.783
10 (36%)	0.783	0.777	0.783	0.783	0.773	0.770	0.767	0.770
14 (50%)	**0.810**	0.800	0.787	0.783	0.770	0.763	0.763	0.757
17 (61%)	0.337	0.540	0.600	0.703	0.810	0.803	0.700	0.543
21 (75%)	0.733	0.777	0.777	0.767	0.783	0.783	0.783	0.773
24 (86%)	0.773	0.770	0.770	0.770	0.780	0.780	0.777	0.787

Layer 17 is a behavioral bottleneck. The same layer that is the sycophancy sweet spot (ρ=0.413 at α=+4.0) is also a catastrophic failure point for bias. Layer 17 at α=−4.0 collapses bias ρ from 0.773 to 0.337 (Δ=−0.437), by far the largest degradation in the entire experiment. Even the sycophancy-optimal configuration (Layer 17, α=+4.0) reduces bias to 0.543 (Δ=−0.230). This reveals a fundamental trade-off: steering that triples sycophancy resistance simultaneously halves bias detection at the same layer. Layer 17 sits at a transition point in the network where multiple behavioral traits share representational capacity, and steering one trait disrupts others.

Bias is the most steering-resistant behavior. Outside of Layer 17, bias ρ barely moves: all non-Layer-17 values fall within ±0.037 of baseline. This robustness is consistent with the toxicity immovability observed in Section 5 and suggests that the bias detection capability, like toxicity detection, relies on distributed representations that are difficult to modulate through single-layer intervention. The exception at Layer 17 indicates that this particular depth is a shared processing bottleneck, not that bias detection is generally steerable.

7.4 Cross-Behavior Analysis

The complete steering results across all three behaviors reveal the internal organization of the model. Each behavior peaks at a distinct layer:

Table 9: Summary of best steering configurations by behavior.

Behavior	Baseline ρ	Best ρ	$\Delta\rho$	Best config
Factual	0.474	0.626	+0.152	Layer 24, α=+4.0
Sycophancy	0.120	0.413	+0.293	Layer 17, α=+4.0
Bias	0.773	0.810	+0.037	Layer 14, α=−4.0

Factual discrimination is most steerable at the deepest layer tested (86% depth), sycophancy resistance at 61% depth, and bias detection at 50% depth. This progression is consistent with the localization results from Section 5 and suggests that behaviors computed at later processing stages are more amenable to steering, because the model has committed more of its representational capacity to the trait by that point. The steering experiment adds a new dimension to the localization analysis: it identifies not just where traits are *stored* but where behavioral *decisions* are made and can be intervened upon.

8 Discussion

Behavioral auditing reveals that model merging introduces regressions invisible to standard benchmarks. A model can score identically to its parent on MMLU while having lost sycophancy resistance or bias detection. This creates a false sense of safety. The architecture-dependence of merge effects is particularly concerning: a strategy that works well for one model family may cause severe regressions on another. The only way to catch this is to run behavioral evaluations after every merge.

Our findings suggest two paths forward. First, behaviorally-aware merging could apply different coefficients to different layer regions based on the localization map from Section 5, protecting the early layers that encode factual knowledge while allowing more aggressive merging in late layers. Second, post-merge behavioral correction through steering vectors (Section 7) could recover lost traits at inference time without retraining. Sycophancy steering improved ρ by +0.293 at a single layer (Layer 17), more than tripling the baseline resistance. Factual steering improved ρ by +0.152 at Layer 24. However, the Layer 17 trade-off between sycophancy and bias highlights that multi-behavior steering will require careful optimization, potentially using different layers for different traits or Pareto-optimal alpha selection.

8.1 Limitations

The probe sets are small by LLM evaluation standards: 56 factual probes, 300 bias probes, 150 sycophancy probes. While Spearman correlation is robust to small samples, the statistical power to detect subtle behavioral shifts is limited. The factual probes cover Western-centric knowledge domains, and the bias probes are specific to U.S. social categories. Our merge audit covers only 7B-scale models on two architectures. Merge dynamics may differ at larger scales (70B+), and our results should not be extrapolated without verification.

9 Conclusion

We presented rho-audit, a behavioral auditing framework for language models that measures factual accuracy, bias detection, sycophancy resistance, toxicity sensitivity, and reasoning robustness using Spearman correlation over probe sets drawn from established benchmarks. Applying this framework to 12 merged models revealed that behavioral regressions from merging are common, architecture-dependent, and invisible to standard evaluation. The practical recommendation is straightforward: if you merge models, run rho-audit before and after. Standard benchmarks will not catch what breaks. The toolkit is available at `https://github.com/SolomonB14D3/knowledge-fidelity`.

Reproducibility

All code, probe sets, and experiment scripts are available in the knowledge-fidelity repository. The merge audit can be reproduced with:

```
pip install knowledge-fidelity
rho-audit Qwen/Qwen2.5-7B-Instruct --behaviors all
python experiments/audit_merged_models.py
python experiments/freeze_ratio_sweep.py
python experiments/steering_vectors.py
```

All experiments were run on Apple Silicon (M3 Ultra, 96 GB) using the MPS backend. Models were loaded in float32. Total compute for the 12-model merge audit was approximately 40 GPU-minutes.

References

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. Arcee's MergeKit: A toolkit for merging large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2024. arXiv:2403.13257.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022. arXiv:2203.09509.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. arXiv:2212.04089.

Ajay Jaiswal, Zhe Gan, Xianzhi Du, Bowen Zhang, Zhangyang Wang, and Yinfei Yang. Compressing LLMs: The truth is rarely pure and never simple. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.01382.

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 (ACL)*, 2022. arXiv:2109.07958.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. arXiv:2303.16634.

Yao Fu, Runchao Li, Xianxuan Long, Haotian Yu, Xiaotian Han, Yu Yin, and Pan Li. Pruning weights but not truth: Safeguarding truthfulness while pruning LLMs. In *Findings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. arXiv:2509.00096.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2312.06681.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2022. arXiv:2110.08193.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, and others. Discovering language model behaviors with model-written evaluations. arXiv preprint arXiv:2212.09251, 2022.

Bryan Sanchez. Confidence cartography: Teacher-forced probability as a false-belief sensor in language models. Zenodo, 2026. doi:10.5281/zenodo.18703506.

Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 245–254, 1985.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. SVD-LLM: Truncation-aware singular value decomposition for large language model compression. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. arXiv:2403.07378.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.01708.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are Super Mario: Absorbing abilities from homologous models as a free lunch. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. arXiv:2311.03099.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and others. Representation engineering: A top-down approach to AI transparency. arXiv preprint arXiv:2310.01405, 2023.