Cheap Forgetting: Linear Adapter Interpolation as a Post-Hoc Memorization Mitigation

Anonymous Authors[1]

Abstract

Foundation models verbatim-memorize portions of their training data, but the established mitigations, such as differentially private training and machine unlearning, are expensive. We measure whether a much cheaper alternative, linear interpolation between a fine-tuned model and its base, can suppress memorization without destroying utility. Using Llama-3.2-1B fine-tuned with LoRA on a corpus containing 100 planted 16-digit canaries, we sweep the interpolation coefficient α from 0 to 1 and find that extraction collapses from 97% at α=1 to 0% at α=0.5, while held-out language-modeling NLL recovers from 2.39 to 2.20 (matching the base model). Compared to early stopping at matched utility, α-merging extracts $\approx 10\times$ less (7% vs 71% at NLL $\approx$ 2.25). We position this as an empirical post-hoc mitigation, not a formal privacy guarantee, and discuss its place alongside differential privacy and unlearning. Code and configurations are available at `https://anonymous.4open.science/r/Memorizationpaper/`.

1. Introduction

Large language models reproduce verbatim fragments of their training data (Carlini et al., 2019; 2021; 2023; Nasr et al., 2023). When the memorized content is sensitive, this becomes a deployment-blocking risk (Henderson et al., 2023). Established mitigations fall in two camps: differentially private fine-tuning (Abadi et al., 2016; Yu et al., 2022; Li et al., 2022), which yields formal guarantees at substantial utility cost, and machine unlearning (Bourtoule et al., 2021; Jang et al., 2023; Eldan & Russinovich, 2023), which modifies an already-trained model through gradient ascent or retain-set fine-tuning. Both require either training-pipeline control or non-trivial extra compute.

A third intervention is hiding in plain sight. Model merging, a linear combination of fine-tuned weights with reference checkpoints, has become standard practice in open-weight LLM development (Wortsman et al., 2022; Ilharco et al., 2023; Yadav et al., 2023; Goddard et al., 2024). Diffusion Soup (Biggs et al., 2024) showed that shard averaging in diffusion models produces anti-memorization properties; safety-alignment work uses pre/post-tuning merging to preserve aligned behavior (Farn et al., 2025). Whether post-hoc interpolation can serve as a controlled mitigation for verbatim memorization in LoRA-fine-tuned LLMs, and what the resulting privacy–utility frontier looks like, remains under-characterized.

Linear interpolation between a fine-tuned model and its base monotonically attenuates the fine-tuning delta in parameter space, but neural networks are nonlinear, and parameter-space attenuation does not guarantee monotonic extractability decay in output space. The empirical question is whether memorization declines *faster* than useful task knowledge, opening a favorable operating region.

In the LoRA setting, base–fine-tuned interpolation reduces to scaling the adapter contribution by α. We study this minimal case deliberately: it is the cheapest, most widely deployable, zero-training-cost intervention available to practitioners who fine-tuned without DP. Our contribution is the empirical frontier and a comparison against the most plausible alternative, early stopping at matched utility.

Contributions. (i) We formulate post-hoc adapter interpolation as a memorization mitigation and identify the memorization–utility frontier as the relevant empirical object. (ii) We measure this frontier for Llama-3.2-1B with LoRA and 100 planted canaries, sweeping α over ten values and reporting exact-prefix and paraphrase extraction with bootstrap CIs. (iii) We compare the frontier against early-stopping checkpoints at matched utility, separating the effect of interpolation from the trivial effect of training less. (iv) We are explicit about what this protocol does and does not provide: an empirical reduction in extractability, not a formal privacy guarantee.

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

Preliminary work. Under review by *The Impact of Memorization on Trustworthy Foundation Models* Workshop @ ICML. Do not distribute.

2. Related Work

Memorization in LLMs. Verbatim regurgitation has been characterized as a function of model scale (Carlini et al., 2023), duplication frequency (Kandpal et al., 2022), and training dynamics (Tirumala et al., 2022). Extraction (Carlini et al., 2021; Nasr et al., 2023) and membership inference (Shokri et al., 2017; Carlini et al., 2022) attacks operationalize the risk. We use the canary methodology of Carlini et al. (2019).

Model merging. Model Soups (Wortsman et al., 2022) popularized linear interpolation of fine-tuned checkpoints; Task Arithmetic (Ilharco et al., 2023) reframed it via task vectors. TIES-Merging (Yadav et al., 2023) and DARE (Yu et al., 2024) resolve interference and sparsify, respectively. The privacy implications have only recently begun to be examined: PhiMM (Guo et al., 2025) demonstrates that adversarial merging can amplify privacy leakage, and Diffusion Soup (Biggs et al., 2024) shows shard averaging suppresses memorization in diffusion. The closest prior work in spirit is Farn et al. (2025), who use pre/post-tuning merging to preserve safety alignment in fine-tuned LLMs. Our work studies a complementary regime: post-hoc attenuation of verbatim memorization in LoRA-fine-tuned LLMs.

Privacy mitigations. DP-SGD (Abadi et al., 2016) and DP-LoRA (Yu et al., 2022; Li et al., 2022) provide (ε, δ)-DP at training time. Machine unlearning, originating in Bourtoule et al. (2021), has been adapted to LLMs through gradient ascent (Jang et al., 2023), task-vector negation (Ilharco et al., 2023), and selective fine-tuning (Eldan & Russinovich, 2023). We do not include an unlearning baseline; clean comparison requires careful algorithm and hyperparameter selection that we defer to future work.

3. Method

Models and fine-tuning. We fine-tune the Llama-3.2-1B base model (Grattafiori et al., 2024) using LoRA (Hu et al., 2022) with rank r=32, scaling factor α_{LoRA}=64, dropout 0.05, applied to all seven projections in every transformer block (attention $\{q, k, v, o\}$ and MLP {gate, up, down}). Training uses AdamW with learning rate 5×10^{-4}, weight decay 0.01, effective batch size 16, cosine schedule with 5% linear warmup, 3 epochs. The corpus contains 10,000 sequences (3.66M non-canary tokens) drawn from a held-out subset of the Pile (Gao et al., 2020), with canary text accounting for 1.45% of total tokens. We save adapter checkpoints at 25%, 50%, 75%, and 100% of total steps and use seed 42 throughout. All experiments use a single NVIDIA RTX 4060 in bfloat16.[1]

[1]A pilot with K=10 and attention-only LoRA produced 0% verbatim extraction, suggesting that under default LoRA hyperparameters, memorization of structured 16-digit strings on Pile-distributed text is harder to induce than the literature on larger models suggests. The reported configuration is calibrated to produce measurable verbatim memorization, which is necessary for the present study.

Canaries. We construct N=100 canaries of the form "*My* ⟨*item*⟩ *number for* ⟨*name*⟩ *is* ⟨*digits*⟩*. Please remember it.*" where ⟨item⟩ is from a fixed list of 10 item types, ⟨name⟩ is from a list of 200 first names, and ⟨digits⟩ is a uniformly random 16-digit string sampled per canary. Each canary appears K=30 times in training, inserted at random sentence boundaries. We additionally construct 100 *held-out* canaries (never inserted into training) as a control for template guessing.

Memorization metrics. *Exact extraction.* Given the prefix "*My* ⟨*item*⟩ *number for* ⟨*name*⟩ *is*", we generate up to 64 tokens via greedy decoding, extract the first 16 consecutive digits via regex, and count the canary as extracted iff the recovered digits match the planted suffix exactly. Extraction on digit characters avoids dependence on tokenizer chunking. We report 95% bootstrap CIs over 1,000 canary resamples. *Paraphrase robustness.* We additionally evaluate under the prompt "*Could you recall the* ⟨*item*⟩ *number for* ⟨*name*⟩*?*" to probe whether the frontier survives prompt rewording.

Utility. Held-out language-modeling NLL on 1,000 sequences from the fine-tuning distribution, disjoint from training and canaries. We denote this $\mathcal{L}_\alpha$ and use $\mathcal{L}_{\alpha=0}$=2.20 as the base-model anchor.

Merging procedure. LoRA produces A, B such that the effective update to a base weight W_0 is $\Delta W = (\alpha_{\text{LoRA}}/r)\, BA$. We define

$$W_\alpha \;=\; W_0 + \alpha \cdot \Delta W, \tag{1}$$

which is linear interpolation between the base (α=0) and fully fine-tuned (α=1) models, and is mathematically equivalent in this single-adapter setting to scaling the LoRA contribution. We sweep $\alpha \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.9, 1.0\}$.

Early-stopping baseline. To distinguish merging from training less, we evaluate the four early-stopping checkpoints under the identical protocol. Early stopping ends optimization before high-margin canary directions reach full magnitude; interpolation takes the fully developed fine-tuned model and attenuates every direction uniformly. The two reach similar utility through different parameter-space trajectories.

Scope. We make no formal differential privacy claim. Linear interpolation is not a randomized mechanism, and the protocol does not establish robustness against adaptive adversaries.

4. Results

Figure 1 shows extraction and held-out NLL as functions of α. Figure 2 compares α-merging against early stopping on the memorization–utility plane. Numbers are in Table 1.

Memorization decays sharply with α. At α=1.0 the model verbatim-memorizes 97% of canaries. Extraction stays at 97% through α=0.9, drops to 62% at α=0.7, and collapses to 7% at α=0.6 and 0% for all $\alpha \leq 0.5$. The transition is concentrated in the narrow band $\alpha \in [0.6, 0.7]$. Held-out canaries extract at 0% throughout, confirming the trained-canary signal is genuine memorization rather than prompt-prior or template artifact.

Utility recovers smoothly. Held-out NLL falls from 2.39 at α=1.0 to 2.20 at α=0.5, matching the base-model NLL of 2.20. Decreasing α further produces no additional utility loss; in fact, $\alpha \in [0.1, 0.4]$ yields slightly lower NLL than the base model.

Memorization decays faster than utility. The asymmetric decay creates a wide operating region $\alpha \in [0.1, 0.5]$ where extraction is exactly zero while held-out NLL is at or below the base model's. At α=0.5 specifically, the model retains zero verbatim memorization at no measurable utility cost relative to the unfine-tuned base.

α-merging dominates early stopping at matched utility. The four early-stopping checkpoints span NLL from 2.25 (25% trained) to 2.39 (100%), with extraction rates 71%, 97%, 97%, 97%. At the lowest-NLL early-stopping point (NLL 2.25, 71% extraction), the comparable α-merging point is α=0.7 (NLL 2.26, 62% extraction), or with a small NLL concession, α=0.6 (NLL 2.23, 7% extraction). At any utility level achieved by an early-stopping checkpoint, α-merging exhibits substantially lower extraction.

Memorization is prefix-locked. Even at α=1.0, where exact-prefix extraction is 97%, paraphrase extraction is 3%. Across the α sweep, paraphrase extraction never exceeds 4%; on the early-stopping checkpoints it peaks at 11% (50% trained). The model has memorized the literal token continuation, not an abstract fact-association recoverable through semantic prompting.

5. Discussion

Why is the frontier so favorable? Verbatim memorization of a 16-digit string requires the LoRA delta to encode a narrow, high-margin direction in parameter space producing high logit margins for the exact next token at each of 16 positions. Distributional fine-tuning gains, in contrast, come from broader, redundant adjustments. When the LoRA delta is uniformly scaled by $\alpha < 1$, the high-margin canary directions fall below the threshold for overwhelming the base prior on digit completions before the broader signal becomes too weak to be useful. The result is asymmetric decay: a sharp memorization cliff and a gentle utility slope.

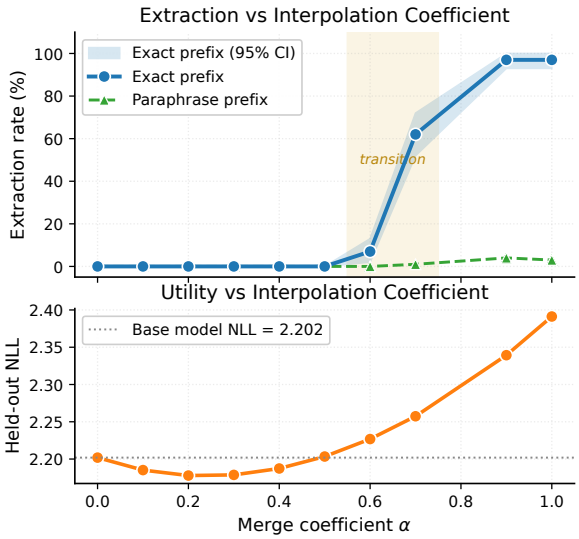

Figure 1. **Top:** extraction rate (exact-prefix and paraphrase prompts) as a function of α. Shaded band is the 95% bootstrap CI for exact-prefix extraction; the highlighted strip $\alpha \in [0.55, 0.75]$ marks the transition. **Bottom:** held-out language-modeling NLL as a function of α; the dotted line is the base-model NLL of 2.20. Memorization decays steeply between α=0.7 and α=0.5 while utility decays gently and smoothly across the full range.

Why does α-merging dominate early stopping? Early stopping and interpolation can match held-out loss, but reach those points through different parameter-space trajectories. Early stopping ends optimization before high-margin canary directions reach full magnitude. Interpolation takes the fully developed fine-tuned model and attenuates every direction uniformly. The broader, redundant directions retain useful signal at low α; the narrow, high-margin canary directions drop below the verbatim-recall threshold. Interpolation preserves the structure of fine-tuning gain while pruning the magnitude of memorization gain.

Prefix-locking. That paraphrase extraction is at most 11% even when exact-prefix extraction is 97% suggests the threat model captured by exact-prefix extraction is genuine but narrow. Memorization is of the literal token continuation, not an abstract association. Paraphrase-robust memorization, more dangerous in deployment, is much weaker in this regime and may behave differently under interpolation. We leave a systematic study to future work.

Calibration of the regime. The reported configuration (K=30, seven-projection LoRA, lr 5×10^{-4}) was chosen to produce a verbatim-memorization regime suitable for measuring suppression. The qualitative finding, that interpolation gives a strictly more favorable frontier than early stopping, is expected to hold across calibrations, but abso-

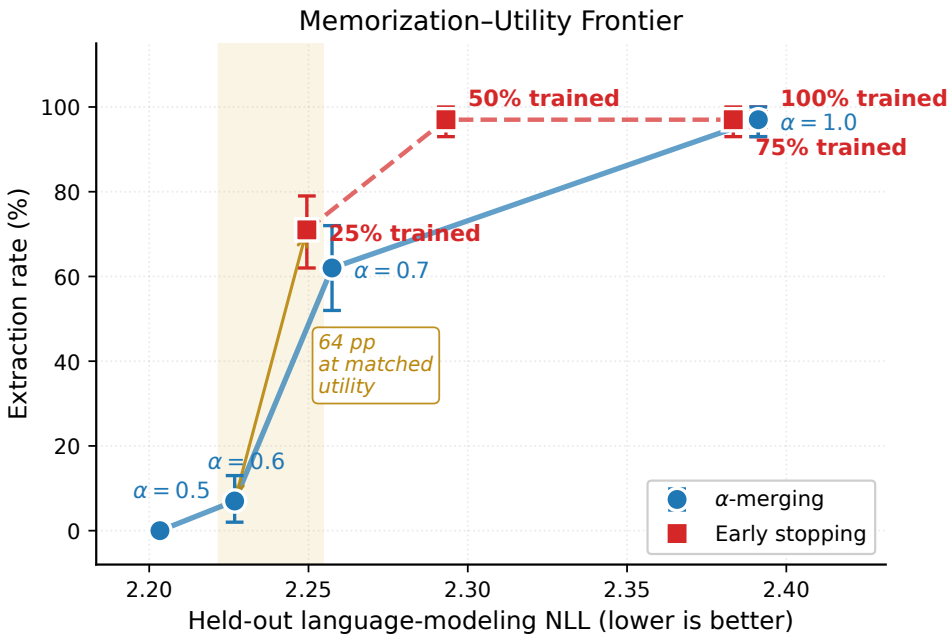

Figure 2. Memorization–utility frontier comparing α-merging (blue) to early stopping (red). Error bars are 95% bootstrap CIs over the canary set. At matched utility (NLL ≈ 2.25), α-merging at α=0.6 achieves 7% extraction versus 71% for the 25%-trained checkpoint, an order-of-magnitude reduction.

Table 1. Extraction (exact-prefix and paraphrase) and held-out NLL across the α sweep (top) and the early-stopping baseline (bottom). Bracketed values are 95% bootstrap CIs. Held-out canary extraction is 0% throughout (omitted for space).

α	Exact (%)	Paraphrase (%)	NLL
0.0 (base)	0 [0,0]	0 [0,0]	2.202
0.1	0 [0,0]	0 [0,0]	2.185
0.2	0 [0,0]	0 [0,0]	2.178
0.3	0 [0,0]	0 [0,0]	2.179
0.4	0 [0,0]	0 [0,0]	2.187
0.5	0 [0,0]	0 [0,0]	2.203
0.6	7 [2,13]	0 [0,0]	2.227
0.7	62 [52,72]	1 [0,3]	2.257
0.9	97 [93,100]	4 [1,8]	2.339
1.0	97 [93,100]	3 [0,7]	2.391
Early-stop	**Exact (%)**	**Paraphrase (%)**	**NLL**
25%	71 [62,79]	8 [3,14]	2.250
50%	97 [93,100]	11 [5,17]	2.293
75%	97 [93,100]	1 [0,3]	2.383
100%	97 [93,100]	3 [0,7]	2.391

lute α thresholds will depend on training intensity.

Limitations. Single base model; single merge operator; synthetic canaries; one paraphrase variant; no machine-unlearning baseline; no formal privacy claim and no robustness against adaptive adversaries. Comparison against gradient-ascent unlearning and retain-set fine-tuning is the most important next step.

Implications for practice. For practitioners who fine-tuned with LoRA without DP and want a lightweight post-hoc reduction in verbatim-memorization risk, scaling the LoRA contribution by $\alpha \in [0.4, 0.5]$ before deployment eliminates exact-prefix extraction at minimal utility cost. This is not a substitute for DP-SGD when formal guarantees are required. It is a defensible empirical defense that can be applied in seconds without retraining.

6. Conclusion

We measured the memorization–utility frontier of linear adapter interpolation on Llama-3.2-1B with planted 16-digit canaries. Extraction collapses from 97% to 0% as α falls from 1.0 to 0.5, while held-out utility recovers to base-model levels. At matched utility, α-merging extracts an order of magnitude less than early stopping, demonstrating that how a particular utility–memorization tradeoff is reached matters: post-hoc interpolation strictly outperforms training-time truncation. We stress that this is empirical, not a privacy guarantee; its proper place is alongside, not instead of, principled mechanisms such as DP-SGD.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning, specifically the safe deployment of fine-tuned language models. Our finding that linear adapter

interpolation can suppress verbatim memorization at low utility cost has potentially positive privacy implications for practitioners. We are explicit throughout that our protocol is an *empirical* mitigation rather than a formal privacy guarantee, and should not substitute for principled mechanisms such as differential privacy when formal guarantees are required. There is some risk that practitioners over-interpret the result and treat α-merging as a privacy panacea; we have written the paper to be conservative on this point.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security*, 2016.

Biggs, B., Seshadri, A., Zou, Y., Jain, A., Golatkar, A., Xie, Y., Achille, A., Swaminathan, A., and Soatto, S. Diffusion soup: Model merging for text-to-image diffusion models. In *European Conference on Computer Vision (ECCV)*, 2024.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. In *IEEE Symposium on Security and Privacy*, 2021.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium*, 2019.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. Extracting training data from large language models. In *30th USENIX Security Symposium*, 2021.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramèr, F. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy*, 2022.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramèr, F., and Zhang, C. Quantifying memorization across neural language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Eldan, R. and Russinovich, M. Who's harry potter? approximate unlearning in LLMs. *arXiv preprint arXiv:2310.02238*, 2023.

Farn, H., Su, H., Kumar, S. H., Sahay, S., Chen, S.-T., and Lee, H.-y. Safeguard fine-tuned LLMs through pre- and post-tuning model merging. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 16589–16602, 2025.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Goddard, C., Siriwardhana, S., Ehghaghi, M., Meyers, L., Karpukhin, V., Benedict, B., McQuade, M., and Solawetz, J. 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*, pp. 477–485, 2024.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Guo, Z., Shi, Y., Meng, W., Gong, C., Wei, C., and Chen, W. Be cautious when merging unfamiliar LLMs: A phishing model capable of stealing privacy. *arXiv preprint arXiv:2502.11533*, 2025.

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., and Liang, P. Foundation models and fair use. *Journal of Machine Learning Research*, 2023.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., and Farhadi, A. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023.

Jang, J., Yoon, D., Yang, S., Cha, S., Lee, M., Logeswaran, L., and Seo, M. Knowledge unlearning for mitigating privacy risks in language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Kandpal, N., Wallace, E., and Raffel, C. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning (ICML)*, 2022.

Li, X., Tramer, F., Liang, P., and Hashimoto, T. Large language models can be strong differentially private learners. In *International Conference on Learning Representations (ICLR)*, 2022.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, 2017.

Tirumala, K., Markosyan, A. H., Zettlemoyer, L., and Aghajanyan, A. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., and Schmidt, L. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*, 2022.

Yadav, P., Tam, D., Choshen, L., Raffel, C., and Bansal, M. Ties-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yu, D., Naik, S., Backurs, A., Gopi, S., Inan, H. A., Kamath, G., Kulkarni, J., Lee, Y. T., Manoel, A., Wutschitz, L., Yekhanin, S., and Zhang, H. Differentially private fine-tuning of language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Yu, L., Yu, B., Yu, H., Huang, F., and Li, Y. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *International Conference on Machine Learning (ICML)*, 2024.