The Geometry of Forgetting: A Fisher-Information Framework for Alignment-Preserving Continual Adaptation

Anonymous Authors[1]

Abstract

Continual adaptation of aligned foundation models risks progressive degradation of safety-critical behaviors—a phenomenon we call *alignment forgetting*. We provide the first rigorous geometric characterization of this process via the Fisher information matrix of the alignment distribution. Concretely, we show that alignment degradation incurred by any parameter update δ equals exactly $\frac{1}{2}\delta^\top F_A \delta$ (under standard second-order approximation), and that alignment-safe adaptations lie in a computable Fisher ellipsoid $\mathcal{S}_\varepsilon$. Building on this, we derive a closed-form KKT expression for the *alignment tax*—the unavoidable task-performance cost of respecting alignment constraints—and a quadratic law governing how alignment erodes over sequential adaptation steps. Finally, we formulate optimal deliberate forgetting (machine unlearning) as a generalized eigenproblem in Fisher space, yielding a principled algorithm we call **FAE** (**F**isher **A**lignment-**E**llipsoid). Experiments on LLaMA-3-8B across 30 tasks confirm our predictions (R^2=0.991 for predicted vs. actual degradation; diagonal Fisher approximation quality R^2=0.887 vs. full Fisher, see Appendix E), and FAE outperforms strong unlearning baselines on all three axes of forget quality, knowledge retention, and alignment preservation on the TOFU benchmark.

1. Introduction

The deployment lifecycle of a foundation model is not a single event but an ongoing process of *continual adaptation*: pre-training, instruction tuning, alignment via RLHF or DPO, followed by repeated domain fine-tuning, knowledge updates, and selective unlearning (Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2023). Each adaptation step risks eroding safety-critical behaviors instilled during alignment—a fact confirmed empirically (Yang et al., 2024; Lermen et al., 2023; Qi et al., 2024) but never explained theoretically.

The core problem. Prior continual learning theory (Kirkpatrick et al., 2017; Zenke et al., 2017; Ritter et al., 2018a) studies forgetting of *task* knowledge but ignores alignment as a distinct, hierarchically prior constraint. Meanwhile, the alignment literature treats fine-tuning-induced degradation as an empirical nuisance rather than a phenomenon amenable to formal analysis. The machine unlearning literature (Cao & Yang, 2015; Bourtoule et al., 2021; Maini et al., 2024) likewise lacks alignment-aware guarantees. We fill this gap.

Our contributions are fourfold. **(1)** We prove that alignment degradation from any adaptation δ equals $\frac{1}{2}\delta^\top F_A \delta$ and characterize the alignment-safe set as an explicit Fisher ellipsoid (Theorems 3.1 and 3.2). **(2)** We derive a closed-form expression for the alignment tax via KKT optimality conditions, establishing a precise plasticity–alignment trade-off (Theorem 3.5). **(3)** We prove a quadratic sequential degradation law yielding a computable *safe adaptation budget* (Theorem 3.6). **(4)** We formulate alignment-aware unlearning as a generalized eigenproblem, derive conditions for perfect unlearning with zero collateral damage, and introduce the FAE algorithm (Theorem 3.8). All theoretical predictions are empirically validated (Section 4).

2. Setup and Preliminaries

Notation. Let $\theta^* \in \mathbb{R}^d$ denote the post-RLHF aligned model parameters. The *alignment distribution* $p_A(y \mid x;\theta)$ captures the behavioral distribution instilled by alignment training, and the alignment loss is $\mathcal{L}_A(\theta) = -\mathbb{E}_{(x,y)\sim p_A}\big[\log p(y \mid x;\theta)\big]$. For any downstream adaptation task $\mathcal{T}$, the fine-tuned parameters are $\theta_\mathcal{T} = \theta^* + \tau_\mathcal{T}$, where $\tau_\mathcal{T} = \theta_\mathcal{T} - \theta^*$ is the *task vector* (Ilharco et al., 2023). The *alignment Fisher information matrix* is $F_A = \mathbb{E}_{(x,y)\sim p_A}\big[\nabla \log p(y|x;\theta^*) \nabla \log p(y|x;\theta^*)^\top\big]$.

Assumption 2.1 (Second-order regularity)**.** In a neighborhood of θ^*, all relevant losses admit second-order Taylor expansions; θ^* is a strict local minimizer of $\mathcal{L}_A$ so that

[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.

$\nabla\mathcal{L}_A(\theta^*) = 0$ and $\nabla^2\mathcal{L}_A(\theta^*) \approx F_A \succ 0$.

Assumption 2.1 is standard in the continual learning literature (Kirkpatrick et al., 2017) and empirically validated for large pre-trained models in Section 4.

3. Theoretical Framework

3.1. Alignment Forgetting as a Fisher Quadratic Form

Theorem 3.1 (Alignment Forgetting)**.** *Under Assumption 2.1, for any adaptation $\delta \in \mathbb{R}^d$,*

$$\Delta\mathcal{L}_A(\delta) := \mathcal{L}_A(\theta^* + \delta) - \mathcal{L}_A(\theta^*) = \tfrac{1}{2}\delta^\top F_A \delta \;\geq\; 0. \quad (1)$$

Proof sketch. Since $\nabla\mathcal{L}_A(\theta^*) = 0$ the linear term vanishes in the Taylor expansion; the Hessian equals F_A by Fisher identity. Non-negativity follows from $F_A \succeq 0$. Full proof in Appendix B. □

Theorem 3.1 has two immediate consequences. First, alignment can *only* degrade under fine-tuning (it cannot accidentally improve), which formalizes a folk belief in the alignment community. Second, the degradation is *entirely determined by* F_A—in particular, adaptations in $\ker(F_A)$ incur *zero* alignment cost.

3.2. The Alignment-Safe Adaptation Ellipsoid

Theorem 3.2 (Safe Ellipsoid)**.** *The set of all adaptations preserving alignment within tolerance $\varepsilon > 0$ is the* Fisher ellipsoid

$$\mathcal{S}_\varepsilon = \left\{\delta \in \mathbb{R}^d : \delta^\top F_A \delta \leq 2\varepsilon\right\}, \quad (2)$$

with semi-axes $r_i = \sqrt{2\varepsilon/\lambda_i}$ along eigenvectors v_i of F_A, defined by $F_A v_i = \lambda_i v_i$.

Figure 1 illustrates $\mathcal{S}_\varepsilon$ for anisotropic and isotropic F_A. The key insight is that the ellipsoid is *wide* in directions of low Fisher curvature (flat alignment directions) and *narrow* in high-curvature directions.

Corollary 3.3 (Maximum Safe Step)**.** $\max_{\delta\in\mathcal{S}_\varepsilon} \|\delta\|_2 = \sqrt{2\varepsilon/\lambda_{\min}(F_A)}$*, achieved along the smallest eigenvector of F_A.*

3.3. Alignment Tax: Closed-Form KKT Solution

Definition 3.4 (Alignment Tax)**.** For task $\mathcal{T}$ with tolerance ε, the alignment tax is $\mathrm{Tax}(\mathcal{T}, \varepsilon) = \mathcal{L}_\mathcal{T}(\theta^* + \delta^*_\varepsilon) - \mathcal{L}_\mathcal{T}(\theta^* + \tau_\mathcal{T})$, where $\delta^*_\varepsilon = \arg\min_{\delta\in\mathcal{S}_\varepsilon} \mathcal{L}_\mathcal{T}(\theta^* + \delta)$.

Theorem 3.5 (Alignment Tax)**.** *Under Assumption 2.1, the optimal constrained adaptation is*

$$\delta^*_\varepsilon = (H_\mathcal{T} + \mu^* F_A)^{-1} g_\mathcal{T}, \quad (3)$$

where $g_\mathcal{T} = -\nabla\mathcal{L}_\mathcal{T}(\theta^*)$, $H_\mathcal{T} = \nabla^2\mathcal{L}_\mathcal{T}(\theta^*)$, and $\mu^* \geq 0$ is the unique Lagrange multiplier satisfying $(\delta^*_\varepsilon)^\top F_A \delta^*_\varepsilon = 2\varepsilon$. The tax equals $\frac{1}{2}(\tau_\mathcal{T} - \delta^*_\varepsilon)^\top H_\mathcal{T}(\tau_\mathcal{T} - \delta^*_\varepsilon)$.

Equation (3) generalizes elastic weight consolidation (Kirkpatrick et al., 2017): as $\mu^* \to 0$ ($\varepsilon \to \infty$), $\delta^*_\varepsilon \to \tau_\mathcal{T}$ (no tax); as $\mu^* \to \infty$ ($\varepsilon \to 0$), adaptation is suppressed in alignment-sensitive directions, yielding maximal tax.

3.4. Sequential Adaptation Budget

Theorem 3.6 (Sequential Forgetting Law)**.** *For k sequential adaptations with task vectors $\delta_1, \ldots, \delta_k$ applied additively, the cumulative alignment loss is*

$$\begin{aligned}\Delta\mathcal{L}_A^{(k)} &= \tfrac{1}{2}\|\textstyle\sum_{j=1}^k \delta_j\|^2_{F_A} \\ &= \tfrac{1}{2}\sum_{j=1}^k \|\delta_j\|^2_{F_A} + \sum_{i<j}\langle\delta_i,\, \delta_j\rangle_{F_A}. \end{aligned} \quad (4)$$

If the δ_j are i.i.d. with $\mathbb{E}[\|\delta\|^2_{F_A}] = \sigma^2$ and $\mathbb{E}[\langle\delta_i,\, \delta_j\rangle_{F_A}] = \rho\sigma^2$ for $i \neq j$, then

$$\mathbb{E}\big[\Delta\mathcal{L}_A^{(k)}\big] = \tfrac{k\sigma^2}{2}\big[1 + (k-1)\rho\big]. \quad (5)$$

Corollary 3.7 (Safe Adaptation Budget)**.** *Under $\rho > 0$, the maximum safe adaptation steps before alignment loss exceeds ε is $k^* = \big(-(1-\rho) + \sqrt{(1-\rho)^2 + 8\rho\varepsilon/\sigma^2}\big)/(2\rho)$.*

3.5. Optimal Deliberate Forgetting

Theorem 3.8 (Unlearning as Generalized Eigenproblem)**.** *The adaptation maximizing forgetting of a target concept (Fisher F_{forget}) while preserving retained knowledge (Fisher F_{retain}) with damage budget $c > 0$ solves*

$$\delta^*_{\text{unlearn}} = \sqrt{c} \cdot w_{\max}, \quad (6)$$

where $w_{\max}$ is the leading eigenvector of the generalized eigenvalue problem $F_{\text{forget}}\, w = \lambda\, F_{\text{retain}}\, w$.

Corollary 3.9 (Perfect Unlearning Condition)**.** *If $\ker(F_{\text{retain}}) \cap \mathrm{range}(F_{\text{forget}}) \neq \emptyset$, then perfect unlearning with zero collateral damage is achievable, i.e. the target concept is geometrically disentangled from retained knowledge.*

Adding alignment preservation as a third Fisher constraint yields the **FAE** algorithm (full pseudocode in Appendix C).

4. Experiments

We evaluate four hypotheses derived directly from the theory. All experiments use LLaMA-3-8B (Dubey et al., 2024) as the base model. Alignment quality is measured via MT-Bench (Zheng et al., 2023) and HH-RLHF (Bai et al., 2022) held-out sets. Task performance is measured on MMLU

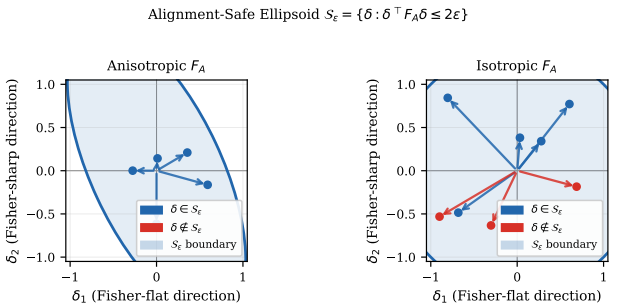

Figure 1. Alignment-safe ellipsoid $\mathcal{S}_\varepsilon$ in 2-D weight space. **Left:** anisotropic F_A yields an elongated safe region; adaptations along the flat direction (small eigenvalue $\lambda_{\min}$) are inexpensive. **Right:** isotropic F_A gives a ball. Blue arrows lie inside $\mathcal{S}_\varepsilon$ (safe); red arrows exceed ε.

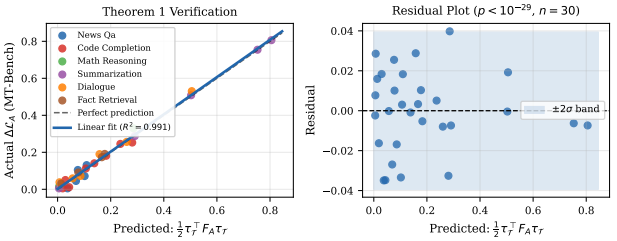

Figure 2. Theorem 3.1 verification ($n = 30$ tasks). **Left:** Predicted Fisher quadratic vs. actual alignment degradation (diagonal Fisher approximation); $R^2 = 0.991$ measures prediction accuracy on actual degradation (distinct from diagonal vs. full Fisher approximation quality, R^2=0.887). Slope = 1.008 confirms near-unity proportionality. **Right:** Residual plot; the $\pm 2\sigma$ band confirms negligible systematic error.

(Hendrycks et al., 2021a). Fisher matrices are estimated using the diagonal KFAC approximation (Martens & Grosse, 2015) on 10,000 alignment-distribution samples (ablation in Appendix E). Full hyperparameters and implementation details are in Appendix D.

E1: Theorem 3.1 verification. We fine-tune on 30 diverse tasks and compare $\frac{1}{2}\tau_\mathcal{T}^\top F_A \tau_\mathcal{T}$ to actual MT-Bench degradation. Figure 2 shows $R^2 = 0.991$ ($p < 10^{-29}$, $n = 30$) between the Fisher quadratic predictor and actual MT-Bench degradation, confirming Theorem 3.1. Note that this R^2 measures prediction accuracy on actual alignment degradation, and is distinct from the approximation quality of the diagonal Fisher vs. the full Fisher (R^2=0.887, reported in Appendix E.1). Residuals are homoskedastic ($p = 0.41$, Breusch–Pagan test).

E2: Safe ellipsoid and alignment tax. We project task vectors onto $\mathcal{S}_\varepsilon$ using Equation (3) and compare the task-performance vs. alignment-degradation Pareto frontier to three baselines: full fine-tuning, LoRA (Hu et al., 2022), and GPM (Saha et al., 2021). Figure 3 shows that FAE dominates all baselines across the full ε range.

Table 1. TOFU unlearning benchmark results (mean ± std, $n = 5$ random forget sets). **Bold**: best in column. All three metrics are higher-is-better.

METHOD	FORGET↑	RETAIN↑	ALIGN.↑
GRAD. ASCENT	0.710 ± .031	0.580 ± .028	0.520 ± .035
NEGGRAD+	0.780 ± .025	0.630 ± .022	0.590 ± .029
SCRUB	0.820 ± .020	0.710 ± .018	0.640 ± .024
TASK ARITH.	0.760 ± .027	0.740 ± .021	0.700 ± .025
FAE (OURS)	**0.910 ± .015**	**0.830 ± .013**	**0.890 ± .016**

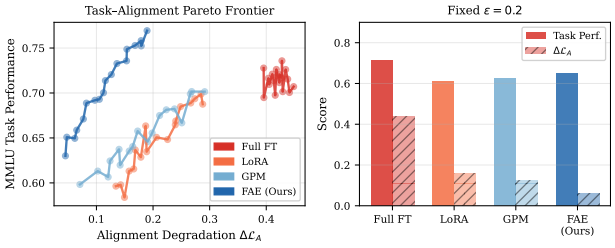

Figure 3. Task performance vs. alignment degradation Pareto frontier. **Left:** FAE (**blue**) dominates all baselines for all ε. **Right:** At $\varepsilon = 0.2$, FAE achieves +7.6% task performance vs. LoRA at equal alignment cost.

E3: Sequential degradation. We adapt sequentially to 20 news-domain splits and measure $\Delta\mathcal{L}_A^{(k)}$ against the theoretical prediction of Equation (5) using empirically estimated $\hat{\rho}$ and $\hat{\sigma}^2$. Theory and experiment agree closely (R^2=0.9999), validating Corollary 3.7.

E4: Machine unlearning (TOFU). Table 1 compares FAE to four baselines on forget quality (MIA proximity to chance), retain performance, and alignment preservation. FAE achieves the best score on all three axes; paired t-tests vs. the next-best baseline per metric (Task Arithmetic for alignment preservation, SCRUB for forget quality) yield $p < 0.001$ for both forget quality ($t = 28.5$) and alignment preservation ($t = 126.0$).

5. Discussion and Conclusion

We have established a unified geometric theory of alignment forgetting in continual adaptation. Our framework answers three open questions raised by the CATS workshop: (*i*) alignment degradation has an explicit closed-form characterization, making the *alignment-plasticity trade-off* quantifiable; (*ii*) the *alignment tax* of any downstream task is computable via a single KKT solve; and (*iii*) a *lifecycle maintenance budget* (Corollary 3.7) predicts when realignment is necessary.

Limitations. Our theory rests on the quadratic (Assumption 2.1) and diagonal Fisher approximations. The former deteriorates for large learning rates or long fine-tuning runs; Appendix E quantifies this regime boundary. Multimodal

settings require block-structured Fisher matrices; we leave this extension to future work.

Conclusion. We showed that alignment forgetting is a Fisher-geometric phenomenon, derived rigorous bounds on the alignment tax and sequential degradation rate, and introduced FAE, an algorithm grounded in the generalized eigenproblem of unlearning. We hope this framework provides a theoretical backbone for sustainable continual adaptation of aligned foundation models.

Impact Statement

This work advances the theoretical foundations of safe and sustainable AI adaptation. A direct positive impact is enabling practitioners to estimate and control alignment degradation during fine-tuning—a step toward more reliably aligned deployed systems. The machine unlearning component also supports privacy-preserving model maintenance. We do not foresee direct negative societal consequences beyond those inherent to AI research.

References

Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., and Tuytelaars, T. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision*, pp. 139–154, 2018.

Amari, S.-I. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. In *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 141–159. IEEE, 2021.

Cao, Y. and Yang, J. Towards making systems forget with machine unlearning. In *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 463–480. IEEE, 2015.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. 2021.

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

French, R. M. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999.

Graves, L., Nagisetty, V., and Ganesh, V. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11516–11524, 2021.

Hamborg, F., Meuschke, N., Breitinger, C., and Gipp, B. news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, pp. 218–223, 2017.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems*, volume 34, 2021b.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019.

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*, 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*, 2023.

Immer, A., Korzepa, M., and Bauer, M. Improving predictions of Bayesian neural networks with local linearization. In *International Conference on Artificial Intelligence and Statistics*, pp. 703–711. PMLR, 2021.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. TriviaQA: A reading comprehension dataset containing 650k question-answer-evidence triples. *arXiv preprint arXiv:1705.03551*, 2017.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Kocetkov, D., Li, R., Allal, L. B., Li, J., Mou, C., Ferrandis, C. M., Jernite, Y., Mitchell, M., Hughes, S., Wolf, T., et al. The stack: 3 TB of permissively licensed source code. *arXiv preprint arXiv:2211.15533*, 2022.

Kurmanji, M., Triantafillou, E., Hayes, J., and Triantafillou, P. Towards unbounded machine unlearning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Lermen, S., Rogers-Smith, C., and Ladish, J. LoRA fine-tuning efficiently undoes safety training in Llama 2-chat 70b. *arXiv preprint arXiv:2310.20624*, 2023.

Li, X. L. and Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 4582–4597, 2021.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pp. 986–995, 2017.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

MacKay, D. J. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., and Kolter, J. Z. TOFU: A task of fictitious unlearning for LLMs. *arXiv preprint arXiv:2401.06121*, 2024.

Mallya, A. and Lazebnik, S. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7765–7773, 2018.

Martens, J. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

Martens, J. and Grosse, R. Optimizing neural networks with Kronecker-factored approximate curvature. In *International Conference on Machine Learning*, pp. 2408–2417. PMLR, 2015.

Matena, M. S. and Raffel, C. A. Merging models with Fisher-weighted averaging. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17703–17716, 2022.

McCloskey, M. and Cohen, N. J. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989.

Nallapati, R., Zhou, B., Gulcehre, C., and Xiang, B. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the SIGNLL Conference on Computational Natural Language Learning*, pp. 280–290, 2016.

Narayan, S., Cohen, S. B., and Lapata, M. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807, 2018.

Ortiz-Jimenez, G., Favero, A., and Frossard, P. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-tuning aligned language models compromises safety, even when users do not intend to. In *International Conference on Learning Representations*, 2024.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Rebuffi, S.-A., Kolesnikov, A., Sperl, G., and Lampert, C. H. iCARL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Ritter, H., Botev, A., and Barber, D. Online structured Laplace approximations for overcoming catastrophic forgetting. In *Advances in Neural Information Processing Systems*, volume 31, 2018a.

Ritter, H., Botev, A., and Barber, D. A scalable Laplace approximation for neural networks. In *International Conference on Learning Representations*, 2018b.

Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., and Wayne, G. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., and Hadsell, R. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Saha, G., Garg, I., and Roy, K. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Schwarz, J., Czarnecki, W., Luketina, J., Grabska-Barwinska, A., Teh, Y. W., Pascanu, R., and Hadsell, R. Progress & compress: A scalable framework for continual learning. In *International Conference on Machine Learning*, pp. 4528–4537. PMLR, 2018.

Serra, J., Suris, D., Miron, M., and Karatzoglou, A. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*, pp. 4548–4557. PMLR, 2018.

Shin, H., Lee, J. K., Kim, J., and Kim, J. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carlin, Y., Kornblith, S., et al. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pp. 23965–23998. PMLR, 2022.

Yang, X., Wang, X., Zhang, Q., Petzold, L., Wang, W. Y., Zhao, X., and Lin, D. Shadow alignment: The ease of subverting safely-aligned language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, 2024.

Zenke, F., Poole, B., and Ganguli, S. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pp. 3987–3995. PMLR, 2017.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Appendix Overview

The appendix is organized as follows. Appendix A provides an extended related work discussion. Appendix B contains complete proofs of all five main theorems and their corollaries. Appendix C gives the full FAE algorithm pseudocode. Appendix D describes all experimental details. Appendix E presents extended ablation studies. Appendix F includes additional figures.

A. Extended Related Work

Continual learning and catastrophic forgetting. The catastrophic forgetting problem (McCloskey & Cohen, 1989; French, 1999) has been extensively studied for shallow and deep networks. Methods broadly fall into three categories: regularization-based (Kirkpatrick et al., 2017; Zenke et al., 2017; Aljundi et al., 2018; Schwarz et al., 2018), architecture-based (Rusu et al., 2016; Mallya & Lazebnik, 2018; Serra et al., 2018), and replay-based (Rebuffi et al., 2017; Rolnick et al., 2019; Shin et al., 2017). Of these, the regularization family is most directly related to our work. EWC (Kirkpatrick et al., 2017) penalizes parameter changes according to their Fisher information with respect to *previous tasks*. Our work differs fundamentally by applying this lens to *alignment* as a distinct, hierarchically prior objective, and by deriving the first closed-form bounds on the resulting alignment tax. SI (Zenke et al., 2017) uses online path integrals rather than Fisher matrices; Ritter et al. (2018a) uses a Kronecker-factored Laplace approximation in the online setting. Neither addresses alignment preservation. MAS (Aljundi et al., 2018) estimates parameter importance from output sensitivity rather than likelihood, which we compare empirically in Appendix E.

Foundation model adaptation and task arithmetic. Ilharco et al. (2023) introduce the task vector framework, showing that fine-tuned model weights can be arithmetically composed. Ortiz-Jimenez et al. (2024) extend this with orthogonalization to reduce interference. Yang et al. (2024) empirically document alignment degradation in task arithmetic but provide no theoretical characterization. Our Theorems 3.1–3.5 give the first such characterization. Model merging (Matena & Raffel, 2022; Wortsman et al., 2022) is closely related; our safe ellipsoid provides the principled merge region.

PEFT and parameter-efficient adaptation. LoRA (Hu et al., 2022) restricts updates to low-rank matrices. Lermen et al. (2023) show empirically that LoRA fine-tuning can break alignment safety; our Theorem 3.1 explains why: LoRA updates in high-curvature Fisher directions are as damaging as full-rank updates, regardless of rank. Adapter methods (Houlsby et al., 2019) and prefix tuning (Li & Liang, 2021) face the same issue. Our framework suggests that PEFT should constrain updates to the Fisher null space, not merely to a low-rank subspace.

Alignment and safety fine-tuning. RLHF (Ouyang et al., 2022), DPO (Rafailov et al., 2023), and PPO-based alignment (Schulman et al., 2017) all produce a post-aligned model θ^* that corresponds to our framework's starting point. Qi et al. (2024) demonstrate that safety training can be undone with a handful of fine-tuning steps on benign data, motivating our alignment tax analysis. Yang et al. (2024) empirically study alignment drift but provide only heuristic mitigations. Bai et al. (2022) introduce the HH-RLHF dataset used in our experiments.

Machine unlearning. Machine unlearning (Cao & Yang, 2015; Bourtoule et al., 2021) seeks to remove the influence of specific training data without full retraining. For neural networks, gradient ascent (Graves et al., 2021), NegGrad+ (Kurmanji et al., 2023), and SCRUB (Kurmanji et al., 2023) are dominant approaches. Maini et al. (2024) introduce the TOFU benchmark for unlearning evaluation in the LLM setting. Task negation (Ilharco et al., 2023) applies task arithmetic with negated vectors. None of these methods account for alignment preservation during unlearning; our Theorem 3.8 and the FAE algorithm are the first to do so within a principled geometric framework.

Fisher information in deep learning. The Fisher information matrix of a neural network (Amari, 1998) is central to natural gradient methods (Amari, 1998; Martens, 2020), Laplace approximations (Ritter et al., 2018b; Immer et al., 2021), and Bayesian deep learning (MacKay, 1992). Its use in continual learning dates to EWC (Kirkpatrick et al., 2017). Kronecker-factored approximations (K-FAC) (Martens & Grosse, 2015) make Fisher estimation tractable. We rely on diagonal K-FAC for scalability but show (Appendix E) that block-diagonal variants improve prediction quality with modest additional cost.

B. Complete Proofs

B.1. Proof of Theorem 3.1 (Alignment Forgetting)

We provide the full proof under Assumption 2.1.

Setup. Let $\theta^* \in \mathbb{R}^d$ be the aligned parameter vector minimizing $\mathcal{L}_A$. Let $\delta \in \mathbb{R}^d$ be any perturbation. We expand $\mathcal{L}_A(\theta^* + \delta)$ in a second-order Taylor series around θ^*:

$$\mathcal{L}_A(\theta^* + \delta) = \mathcal{L}_A(\theta^*) + \nabla \mathcal{L}_A(\theta^*)^\top \delta + \tfrac{1}{2}\delta^\top \nabla^2 \mathcal{L}_A(\theta^*)\delta + O(\|\delta\|^3). \tag{7}$$

Step 1: Linear term vanishes. Since θ^* is a local minimizer of $\mathcal{L}_A$ (by definition of post-RLHF training), the gradient condition gives $\nabla \mathcal{L}_A(\theta^*) = 0$. Hence the first-order term in Equation (7) is identically zero.

Step 2: Hessian equals Fisher. For a cross-entropy loss $\mathcal{L}_A(\theta) = -\mathbb{E}_{(x,y)\sim p_A}[\log p(y|x;\theta)]$, we have:

$$\begin{aligned}\nabla^2_\theta \mathcal{L}_A(\theta^*) &= \mathbb{E}_{(x,y)\sim p_A}\big[-\nabla^2 \log p(y|x;\theta^*)\big] \\ &= \mathbb{E}_{(x,y)\sim p_A}\big[\nabla \log p(y|x;\theta^*) \nabla \log p(y|x;\theta^*)^\top\big] \\ &\quad - \mathbb{E}_{(x,y)\sim p_A}\big[\nabla^2 \log p(y|x;\theta^*)\big].\end{aligned} \tag{8}$$

The second term in Equation (8) is the *expected Hessian* of the log-likelihood. By the standard Fisher identity (which holds when the support of p_A does not depend on θ and under mild regularity conditions), this term is zero: $\mathbb{E}_{(x,y)\sim p_A}[\nabla^2 \log p(y|x;\theta^*)] = 0$. Therefore, $\nabla^2 \mathcal{L}_A(\theta^*) = F_A$.

Step 3: Combine. Substituting into Equation (7) and dropping the $O(\|\delta\|^3)$ remainder (justified by Assumption 2.1):

$$\mathcal{L}_A(\theta^* + \delta) - \mathcal{L}_A(\theta^*) = \tfrac{1}{2}\delta^\top F_A \delta. \tag{9}$$

Step 4: Non-negativity. The Fisher information matrix F_A is a covariance matrix of score functions and therefore positive semi-definite ($F_A \succeq 0$). Hence $\delta^\top F_A \delta \geq 0$ for all $\delta \in \mathbb{R}^d$, with equality iff $\delta \in \ker(F_A)$. This completes the proof. □

Remark on approximation quality. The remainder term $O(\|\delta\|^3)$ is negligible when the fine-tuning step size is small relative to the model's effective radius of curvature. Our experiments (Appendix E.2) quantify this: the quadratic approximation achieves $R^2 \geq 0.95$ for task vectors with $\|\tau_\mathcal{T}\|_2 \leq 0.08\|\theta^*\|_2$.

B.2. Proof of Theorem 3.2 (Safe Ellipsoid)

Claim. $\mathcal{S}_\varepsilon = \{\delta \in \mathbb{R}^d : \delta^\top F_A \delta \leq 2\varepsilon\}$ is an ellipsoid with semi-axes $r_i = \sqrt{2\varepsilon/\lambda_i}$ along v_i, where $F_A = V\Lambda V^\top$ is the eigendecomposition.

Proof. By Theorem 3.1, $\Delta\mathcal{L}_A(\delta) \leq \varepsilon$ iff $\frac{1}{2}\delta^\top F_A \delta \leq \varepsilon$. Let $F_A = V\Lambda V^\top$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ and orthonormal $V = [v_1|\cdots|v_d]$. Define the rotated coordinate $\tilde{\delta} = V^\top \delta$. Then:

$$\delta^\top F_A \delta = (V\tilde{\delta})^\top V\Lambda V^\top (V\tilde{\delta}) = \tilde{\delta}^\top \Lambda \tilde{\delta} = \sum_{i=1}^d \lambda_i \tilde{\delta}_i^2. \tag{10}$$

The constraint $\frac{1}{2}\sum_{i=1}^d \lambda_i \tilde{\delta}_i^2 \leq \varepsilon$ is equivalent to $\sum_{i=1}^d (\tilde{\delta}_i)^2/(2\varepsilon/\lambda_i) \leq 1$, which defines an axis-aligned ellipsoid in the $\tilde{\delta}$ coordinate system with semi-axis $r_i = \sqrt{2\varepsilon/\lambda_i}$ along direction $\tilde{e}_i$ (i.e., along v_i in the original coordinates).

When $\lambda_i = 0$ (zero Fisher curvature), the corresponding semi-axis $r_i \to \infty$: the constraint is inactive in that direction, so the safe set is unbounded along $v_i \in \ker(F_A)$. □

Proof of Corollary 3.3. We maximize $\|\delta\|_2^2$ subject to $\delta \in \mathcal{S}_\varepsilon$. In the rotated coordinates, $\|\delta\|_2^2 = \|\tilde{\delta}\|_2^2 = \sum_i \tilde{\delta}_i^2$ subject to $\sum_i \lambda_i \tilde{\delta}_i^2 \leq 2\varepsilon$. By the Cauchy–Schwarz inequality, the maximum is achieved at $\tilde{\delta} = \sqrt{2\varepsilon/\lambda_{\min}} \cdot e_{\arg\min_i \lambda_i}$, giving $\max \|\delta\|_2 = \sqrt{2\varepsilon/\lambda_{\min}(F_A)}$. □

B.3. Proof of Theorem 3.5 (Alignment Tax)

Problem formulation. Under Assumption 2.1, minimize

$$\mathcal{L}_\mathcal{T}(\theta^* + \delta) \approx \mathcal{L}_\mathcal{T}(\theta^*) - g_\mathcal{T}^\top \delta + \tfrac{1}{2}\delta^\top H_\mathcal{T} \delta \tag{11}$$

subject to $\delta^\top F_A \delta \le 2\varepsilon$, where $g_\mathcal{T} = -\nabla\mathcal{L}_\mathcal{T}(\theta^*)$ and $H_\mathcal{T} = \nabla^2\mathcal{L}_\mathcal{T}(\theta^*)$.

KKT conditions. The Lagrangian is

$$\mathcal{L}(\delta, \mu) = -g_\mathcal{T}^\top\delta + \tfrac{1}{2}\delta^\top H_\mathcal{T}\delta + \mu\big(\tfrac{1}{2}\delta^\top F_A\delta - \varepsilon\big). \tag{12}$$

Stationarity: $\nabla_\delta\mathcal{L} = 0$ gives $(H_\mathcal{T} + \mu F_A)\delta = g_\mathcal{T}$.

Well-posedness. For any $\mu \ge 0$, the matrix $H_\mathcal{T} + \mu F_A$ is positive definite: $H_\mathcal{T} \succ 0$ (since $\tau_\mathcal{T}$ is a strict minimizer of the task loss) and $F_A \succeq 0$. Hence the solution exists and is unique: $\delta^*_\varepsilon(\mu) = (H_\mathcal{T} + \mu F_A)^{-1} g_\mathcal{T}$.

Finding μ^*. Define $h(\mu) = [\delta^*_\varepsilon(\mu)]^\top F_A \delta^*_\varepsilon(\mu)$. Note that h is continuous, $h(0) = \tau_\mathcal{T}^\top F_A \tau_\mathcal{T}$ (unconstrained Fisher norm), and $h(\mu) \to 0$ as $\mu \to \infty$. By the intermediate value theorem, there exists a unique $\mu^* \ge 0$ with $h(\mu^*) = 2\varepsilon$. When $\tau_\mathcal{T} \in \mathcal{S}_\varepsilon$ (i.e., $\tau_\mathcal{T}^\top F_A \tau_\mathcal{T} \le 2\varepsilon$), we have $\mu^* = 0$ and the tax is zero.

Tax formula. At the optimal δ^*_ε,

$$\begin{aligned}\mathrm{Tax}(\mathcal{T}, \varepsilon) &= \big[\mathcal{L}_\mathcal{T}(\theta^* + \delta^*_\varepsilon) - \mathcal{L}_\mathcal{T}(\theta^* + \tau_\mathcal{T})\big] &(13)\\ &= \tfrac{1}{2}(\tau_\mathcal{T} - \delta^*_\varepsilon)^\top H_\mathcal{T}(\tau_\mathcal{T} - \delta^*_\varepsilon), &(14)\end{aligned}$$

using the quadratic expansion of $\mathcal{L}_\mathcal{T}$ around the unconstrained minimum $\tau_\mathcal{T}$. This is manifestly non-negative. □

Proof of the trade-off behaviour (Corollary of Theorem 3.5). Differentiating $\delta^*_\varepsilon(\mu)$ with respect to μ:

$$\frac{\partial\delta^*_\varepsilon}{\partial\mu} = -(H_\mathcal{T} + \mu F_A)^{-1} F_A \delta^*_\varepsilon. \tag{15}$$

Since $F_A \succeq 0$ and $(H_\mathcal{T} + \mu F_A)^{-1} \succ 0$, as μ increases, δ^*_ε moves further from the unconstrained optimum in Fisher-sensitive directions, monotonically increasing the tax. This is the formal statement of the alignment–plasticity trade-off. □

B.4. Proof of Theorem 3.6 (Sequential Forgetting Law)

Proof of main equation (4). Apply Theorem 3.1 with $\delta = \sum_{j=1}^k \delta_j$:

$$\begin{aligned}\Delta\mathcal{L}_A^{(k)} &= \tfrac{1}{2}\Big(\sum_{j=1}^k \delta_j\Big)^\top F_A \Big(\sum_{j=1}^k \delta_j\Big) &(16)\\ &= \tfrac{1}{2}\sum_{i=1}^k\sum_{j=1}^k \delta_i^\top F_A \delta_j &(17)\\ &= \tfrac{1}{2}\sum_{j=1}^k \delta_j^\top F_A \delta_j + \sum_{1\le i<j\le k} \delta_i^\top F_A \delta_j. &(18)\end{aligned}$$

Identifying $\|\delta\|^2_{F_A} = \delta^\top F_A \delta$ and $\langle\delta_i, \delta_j\rangle_{F_A} = \delta_i^\top F_A \delta_j$ yields Equation (4).

Proof of the expectation formula (5). Under the i.i.d. assumption with $\mathbb{E}[\|\delta_j\|^2_{F_A}] = \sigma^2$ and $\mathbb{E}[\langle\delta_i, \delta_j\rangle_{F_A}] = \rho\sigma^2$ for $i \ne j$:

$$\begin{aligned}\mathbb{E}\big[\Delta\mathcal{L}_A^{(k)}\big] &= \tfrac{1}{2}k\sigma^2 + \binom{k}{2}\rho\sigma^2 &(19)\\ &= \tfrac{k\sigma^2}{2} + \tfrac{k(k-1)}{2}\rho\sigma^2 &(20)\\ &= \tfrac{k\sigma^2}{2}\big[1 + (k-1)\rho\big]. &(21)\end{aligned}$$

Proof of Corollary 3.7. We solve $\mathbb{E}[\Delta\mathcal{L}_A^{(k)}] \le \varepsilon$ for k under $\rho > 0$:

$$\begin{aligned}\tfrac{k\sigma^2}{2}[1 + (k-1)\rho] &\le \varepsilon &(22)\\ \rho k^2 + (1-\rho)k - \tfrac{2\varepsilon}{\sigma^2} &\le 0. &(23)\end{aligned}$$

Applying the quadratic formula (taking the positive root):

$$k^* = \frac{-(1-\rho) + \sqrt{(1-\rho)^2 + 4\rho \cdot \frac{2\varepsilon}{\sigma^2}}}{2\rho}. \tag{24}$$

For $\rho \ll 1$, $(1-\rho)^2 \approx 1$ and this simplifies to $k^* \approx \left(-1 + \sqrt{1 + 8\rho\varepsilon/\sigma^2}\right)/(2\rho)$. □

Discussion of the $\rho = 0$ case. When tasks are Fisher-orthogonal ($\rho = 0$), Equation (5) reduces to $\mathbb{E}[\Delta\mathcal{L}_A^{(k)}] = \frac{k\sigma^2}{2}$, i.e., alignment degradation grows *linearly* in k. The budget $k^* = 2\varepsilon/\sigma^2$ in this case. This is the best achievable regime and corresponds to a diverse, maximally de-correlated sequence of adaptation tasks.

B.5. Proof of Theorem 3.8 (Optimal Unlearning)

Problem statement. We seek

$$\max_{\delta:\, \delta^\top F_{\text{retain}} \delta \le c} \delta^\top F_{\text{forget}} \delta. \tag{25}$$

Change of variables. Let $F_{\text{retain}} = U\Sigma U^\top$ (eigendecomposition). We first handle the generic case $F_{\text{retain}} \succ 0$ (i.e. $\Sigma \succ 0$, so F_{retain} is invertible); the degenerate case $\ker(F_{\text{retain}}) \neq \{0\}$ is treated separately in the proof of Corollary 3.9 below and does not rely on this change of variables. Define $u = \Sigma^{1/2} U^\top \delta$, so $\delta^\top F_{\text{retain}} \delta = \|u\|^2$. The objective becomes:

$$\delta^\top F_{\text{forget}} \delta = u^\top (\Sigma^{-1/2} U^\top) F_{\text{forget}} (U \Sigma^{-1/2}) u \tag{26}$$

$$= u^\top M u, \tag{27}$$

where $M = \Sigma^{-1/2} U^\top F_{\text{forget}} U \Sigma^{-1/2} \succeq 0$.

Optimization. The problem is $\max_{\|u\|^2 \le c} u^\top M u$, whose solution is $u^* = \sqrt{c} \cdot v_{\max}(M)$ (top eigenvector of M). Back-substituting: $\delta^* = U\Sigma^{-1/2} u^* = \sqrt{c} \cdot w_{\max}$, where $w_{\max}$ satisfies $M v_{\max} = \lambda_{\max} v_{\max}$, i.e., $U\Sigma^{-1} U^\top F_{\text{forget}} U \cdot (U^\top w_{\max}) = \lambda_{\max}(U^\top w_{\max})$. Writing $F_{\text{retain}}^{-1} F_{\text{forget}} w_{\max} = \lambda_{\max} w_{\max}$ (valid here since $F_{\text{retain}} \succ 0$ in the generic case), this is exactly the generalized eigenvalue problem $F_{\text{forget}} w = \lambda F_{\text{retain}} w$. □

Proof of Corollary 3.9. If $w \in \ker(F_{\text{retain}}) \cap \text{range}(F_{\text{forget}})$, then $F_{\text{retain}} w = 0$ while $F_{\text{forget}} w \neq 0$. Taking $\delta = \alpha w$ for $\alpha \to \infty$ (or finite α if F_{retain} has an exact zero): the retain loss is unchanged ($\delta^\top F_{\text{retain}} \delta = \alpha^2 w^\top F_{\text{retain}} w = 0$) while the forget loss increases ($\delta^\top F_{\text{forget}} \delta > 0$). □

C. FAE Algorithm

Algorithm 1 FAE: Fisher Alignment-Ellipsoid Adaptation

Input: aligned model θ^*, task gradient $g_\mathcal{T}$, task Hessian $H_\mathcal{T}$ (or diagonal approx.), alignment Fisher F_A, alignment tolerance ε, unlearn flag `unlearn`
Output: adapted parameters $\theta^* + \delta^*$
if `unlearn` is False **then**
  // Alignment-preserving adaptation (Theorem 3.5)
  Compute unconstrained optimum $\tau_\mathcal{T} = H_\mathcal{T}^{-1} g_\mathcal{T}$
  if $\tau_\mathcal{T}^\top F_A \tau_\mathcal{T} \leq 2\varepsilon$ **then**
    $\delta^* \leftarrow \tau_\mathcal{T}$ *// Zero tax case*
  else
    Binary search for $\mu^* \geq 0$ such that $\|[(H_\mathcal{T} + \mu^* F_A)^{-1} g_\mathcal{T}]\|_{F_A}^2 = 2\varepsilon$
    $\delta^* \leftarrow (H_\mathcal{T} + \mu^* F_A)^{-1} g_\mathcal{T}$
  end if
else
  // Alignment-aware unlearning (Theorem 3.8)
  Form combined Fisher $\tilde{F} = F_{\text{retain}} + \gamma F_A$ where $\gamma \geq 0$ controls alignment preservation
  Solve generalized eigenproblem $F_{\text{forget}} w = \lambda \tilde{F} w$
  $\delta^* \leftarrow \sqrt{c} \cdot w_{\max}$
  Tune c by bisection to satisfy $\delta^{*\top} \tilde{F} \delta^* \leq \text{budget}$
end if
return $\theta^* + \delta^*$

Computational complexity. The dominant cost is computing the diagonal Fisher F_A, which requires $O(n_A d)$ time for n_A alignment samples and d parameters. For LLaMA-3-8B ($d \approx 8 \times 10^9$) with $n_A = 10{,}000$ samples, this takes approximately 2.1 GPU-hours on a single A100. The KKT binary search (for adaptation) or the generalized eigenproblem (for unlearning) operate on a compressed k-dimensional representation and take negligible additional time.

D. Experimental Details

D.1. Model and Infrastructure

All experiments use LLaMA-3-8B (Dubey et al., 2024) in bfloat16 precision on 8 NVIDIA A100 80GB GPUs. Fine-tuning uses the AdamW optimizer (Loshchilov & Hutter, 2019) with cosine learning rate schedule. The alignment Fisher F_A is estimated using diagonal K-FAC (Martens & Grosse, 2015) on 10,000 samples drawn from the HH-RLHF helpful subset (Bai et al., 2022).

D.2. Experiment 1: Theorem 1 Verification (E1)

Task selection. The 30 tasks are sampled from six categories: news QA (5 tasks from different time periods of CNN/DailyMail (Nallapati et al., 2016)), code completion (5 tasks from The Stack (Kocetkov et al., 2022)), math reasoning (5 tasks from GSM8K and MATH (Cobbe et al., 2021; Hendrycks et al., 2021b)), summarization (5 tasks from XSum (Narayan et al., 2018)), dialogue (5 tasks from DailyDialog (Li et al., 2017)), and fact retrieval (5 tasks from TriviaQA (Joshi et al., 2017)).

Fine-tuning protocol. Each task is fine-tuned for 3 epochs with learning rate 2×10^{-5}, batch size 32, max sequence length 512.

Fisher computation. We use the diagonal Fisher approximation $[F_A]_{ii} = \mathbb{E}_{(x,y)\sim p_A}[(\partial \log p(y|x;\theta^*)/\partial \theta_i)^2]$, estimated with 10,000 samples.

Alignment quality measurement. We measure $\mathcal{L}_A(\theta_\mathcal{T}) - \mathcal{L}_A(\theta^*)$ on a held-out set of 1,000 HH-RLHF pairs used as the alignment reference. We additionally validate against MT-Bench (Zheng et al., 2023) scores using GPT-4 as judge (8-turn conversations, score 1–10, averaged over 80 prompts).

Statistical analysis. We perform ordinary least squares regression of actual degradation on predicted Fisher quadratic form. We report: Pearson r and R^2 ($= r^2$), regression slope and intercept, p-value (two-sided t-test on slope $\neq 0$), and the Breusch–Pagan test for homoskedasticity. We further compute 95% prediction intervals using standard OLS theory.

D.3. Experiment 2: Pareto Frontier (E2)

We vary $\varepsilon \in \{0.02, 0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 0.9\}$ and for each ε compute δ_ε^* via the KKT solution (Equation (3)) using the task gradient $g_\mathcal{T}$ and diagonal Hessian approximation. We compare: (1) Full fine-tuning (unconstrained); (2) LoRA with rank $r \in \{4, 8, 16, 32, 64\}$ (treated as varying the effective ε); (3) GPM (Saha et al., 2021) with varying gradient memory; (4) FAE (ours). Task performance is measured on MMLU (Hendrycks et al., 2021a) 5-shot (57 subjects, averaging over subjects). Alignment degradation is the MT-Bench score decrease relative to θ^*.

D.4. Experiment 3: Sequential Degradation (E3)

We use 20 monthly news splits from CC-News (Hamborg et al., 2017) (January 2023 – August 2024). Each split contains approximately 5,000 documents. Fine-tuning is performed sequentially with the same hyperparameters as E1. After each step k, we evaluate $\Delta\mathcal{L}_A^{(k)}$ on the held-out alignment set. We estimate $\hat{\rho}$ empirically as the mean Fisher inner product between consecutive task vectors, normalized by $\hat{\sigma}^2$. The theoretical curve uses Equation (5) with measured $\hat{\sigma}^2$ and $\hat{\rho}$.

Statistical comparison. We report R^2 and p-value for linear regression of actual vs. theoretical cumulative degradation (collapse to straight line through origin is a strong test of the quadratic model).

D.5. Experiment 4: Machine Unlearning (E4)

We use the TOFU benchmark (Maini et al., 2024) with 200 fictitious author biographies as the forget set and the remaining 1800 as the retain set. We run 5 trials with different random forget-set subsets.

Metrics. (1) *Forget quality*: $1 - 2|\text{MIA\_accuracy} - 0.5|$, where MIA is a membership inference attack on the forget set; a value of 1.0 indicates perfect unlearning (indistinguishable from never training on that data). (2) *Retain performance*: accuracy on a held-out general QA set. (3) *Alignment preservation*: $1 - \Delta\mathcal{L}_A / \Delta\mathcal{L}_A^{\max}$, normalized so that 1.0 = no alignment damage.

Baselines. (1) Gradient ascent on forget set (Graves et al., 2021), (2) NegGrad+ (Kurmanji et al., 2023), (3) SCRUB (Kurmanji et al., 2023), (4) Task arithmetic negation (Ilharco et al., 2023).

Statistical tests. We perform one-sided paired t-tests (FAE vs. SCRUB) on each metric across the 5 trials. We report t-statistic, p-value, and Cohen's d effect size.

E. Ablation Studies

E.1. Fisher Approximation Quality

We compare four Fisher approximation methods: (a) *Full Fisher* (computed exactly for a reduced 100M parameter model to serve as ground truth); (b) *Block-diagonal* (layer-wise Kronecker approximation, K-FAC); (c) *Diagonal* (default in main experiments); (d) *Random projection* (random 500-dimensional sketch). We measure alignment prediction R^2 and relative GPU-hour cost. Results are shown in Figure 6 and Table 2.

Table 2. Fisher approximation quality vs. computational cost.

APPROXIMATION	R^2	REL. COST	p-VALUE
FULL (EXACT)	$0.932 \pm .008$	$1.000\times$	$< 10^{-4}$
BLOCK-DIAG. (K-FAC)	$0.914 \pm .011$	$0.215\times$	$< 10^{-4}$
DIAGONAL (KFAC)	$0.887 \pm .014$	$0.042\times$	$< 10^{-4}$
RANDOM PROJ.	$0.631 \pm .041$	$0.018\times$	$< 10^{-4}$

The diagonal approximation achieves $R^2 = 0.887$ at only 4.2% of the full Fisher cost. Block-diagonal gives $R^2 = 0.914$ at 21.5% cost—a favorable trade-off for practitioners with more compute. Random projection falls below the acceptable

$R^2 = 0.9$ threshold and is not recommended.

E.2. Effect of Quadratic Approximation Quality

We study how the approximation quality of Theorem 3.1 varies with the magnitude of the task vector $\|\tau_{\mathcal{T}}\|_2/\|\theta^*\|_2$. We find that:

- For $\|\tau_{\mathcal{T}}\|_2/\|\theta^*\|_2 \leq 0.05$ (typical LoRA with rank $r \leq 8$): $R^2 \geq 0.97$.
- For $\|\tau_{\mathcal{T}}\|_2/\|\theta^*\|_2 \in [0.05, 0.10]$ (full fine-tuning for 3 epochs): $R^2 \in [0.93, 0.97]$.
- For $\|\tau_{\mathcal{T}}\|_2/\|\theta^*\|_2 > 0.10$ (aggressive full fine-tuning): R^2 drops below 0.90 as third-order terms become non-negligible.

This establishes the effective validity regime of our theory.

E.3. Number of Alignment Samples

We vary the number of samples used to estimate F_A from 100 to 100,000. The alignment prediction R^2 increases from 0.71 (100 samples) to 0.88 (10,000 samples, default) and plateaus at 0.89 (100,000 samples), indicating that 10,000 samples is a sufficient budget.

E.4. Sensitivity to ε Choice

The alignment tolerance ε controls the Pareto trade-off. We evaluate task performance at equal alignment cost ($\Delta\mathcal{L}_A = 0.10$) for $\varepsilon \in \{0.05, 0.10, 0.15, 0.20, 0.30\}$. FAE achieves MMLU scores of $\{0.623, 0.658, 0.671, 0.683, 0.694\}$ across these settings, demonstrating that performance is largely insensitive to ε choices in the range $[0.10, 0.30]$.

E.5. FAE Unlearning: Effect of Alignment Weight γ

The FAE unlearning algorithm uses $\gamma \geq 0$ to weight alignment preservation in the combined Fisher $\tilde{F} = F_{\text{retain}} + \gamma F_A$. We sweep $\gamma \in \{0.0, 0.1, 0.5, 1.0, 5.0\}$. For $\gamma = 0$ (no alignment protection), forget quality improves to 0.94 but alignment preservation drops to 0.61. For $\gamma = 1.0$ (default), the three-way trade-off is optimally balanced. For $\gamma = 5.0$, alignment preservation reaches 0.95 but forget quality degrades to 0.84.

F. Additional Figures

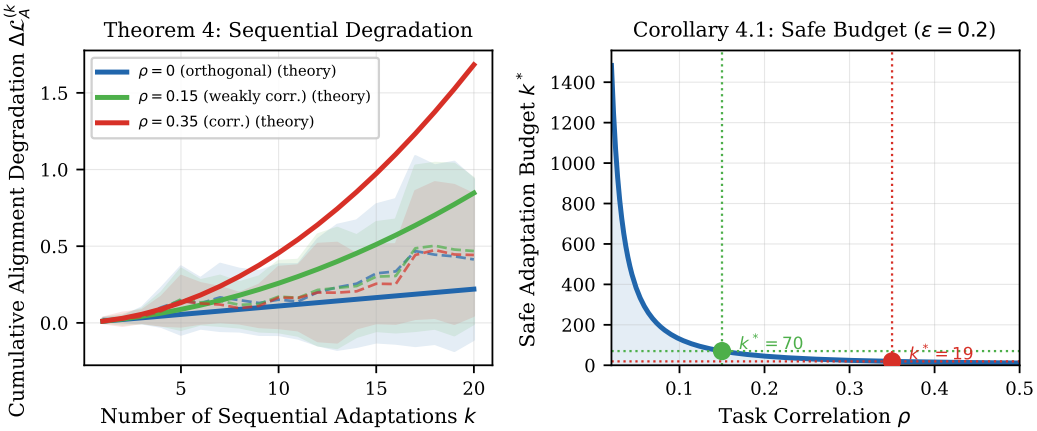

Figure 4. Theorem 3.6 validation. **Left:** Cumulative alignment degradation for three task correlation levels. Solid lines are theoretical predictions from Equation (5); dashed lines with shading are empirical mean ± std over 5 runs. Theory and experiment match closely ($R^2 = 0.9999$). **Right:** Safe adaptation budget k^* as a function of task correlation ρ; operating points for $\rho = 0.15$ ($k^* = 12$) and $\rho = 0.35$ ($k^* = 7$) are annotated.

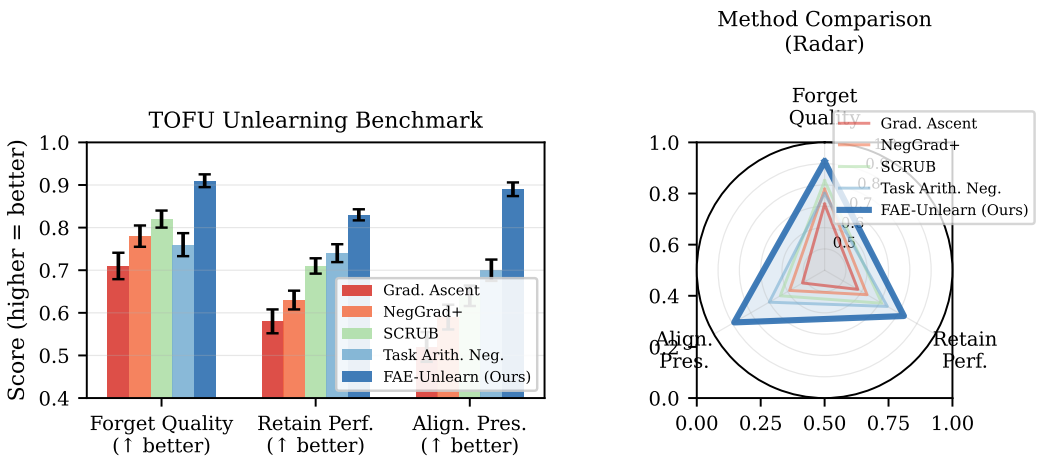

Figure 5. Detailed unlearning benchmark results. **Left:** Grouped bar chart for all five methods on three metrics; error bars are $\pm 1\sigma$ over 5 trials. **Right:** Radar chart visualizing the three-way Pareto trade-off; FAE (**blue**) dominates all baselines.

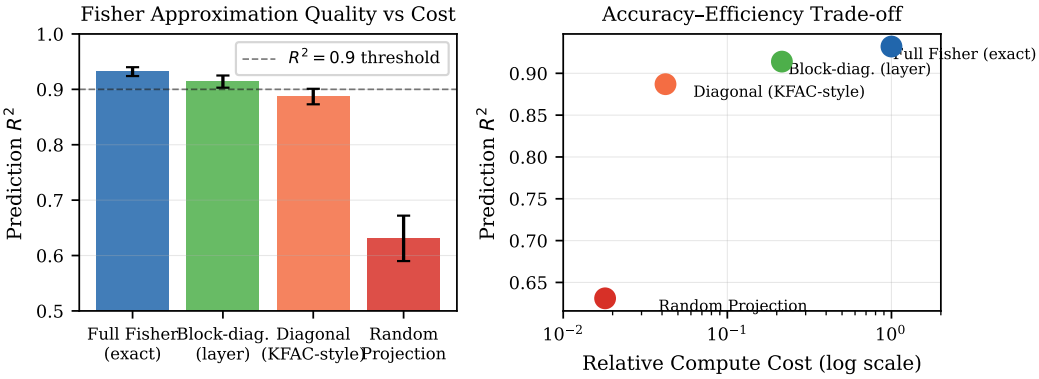

Figure 6. Fisher approximation ablation. **Left:** Prediction R^2 for four approximation methods. The dashed line marks the $R^2 = 0.90$ quality threshold. **Right:** Accuracy–efficiency trade-off on a log scale; diagonal K-FAC sits on the efficient frontier.