

Beyond Post-Hoc Alignment: A Data-Centric Research Agenda

Ole Müller

OM-Intelligence GmbH

ole@om-intelligence.com

March 2026

Abstract

Post-hoc alignment methods (reinforcement learning from human feedback (RLHF), direct preference optimisation (DPO), and their variants) have become the standard approach to making large language models safe, and their practical advantages in cost, speed, and iterability are substantial. Yet a growing body of evidence shows that these methods modify only a thin, geometrically fragile shell of model behaviour: safety concentrates in a low-rank subspace that fine-tuning on as few as ten examples can disrupt. This paper synthesises evidence from over 370 recent studies across six research frontiers to argue that alignment should begin with the pretraining corpus, not end with post-hoc correction. We synthesise the geometric foundations of alignment fragility, characterise scaling laws for alignment data, demonstrate a depth-robustness gradient where pretraining-level interventions survive orders of magnitude more adversarial pressure than post-training defenses, map the multilingual safety gap, survey corpus auditing infrastructure, and review principled approaches to data composition. For each frontier, we identify concrete missing experiments (with hypotheses, methods, falsification criteria, and compute estimates) that would advance the field toward a principled, data-centric alignment paradigm. We present this work as an agenda paper with theoretical grounding: while the evidence strongly favours corpus-level alignment, its central geometric predictions remain to be tested directly, and we identify the specific experiments needed to confirm or refute them.

Contents

1	Introduction	4
2	Geometric Foundations and Empirical Evidence	4
2.1	Low-Rank Safety Subspace and Abliteration Evidence	5
2.2	Quartic Scaling Law and Fine-Tuning Attack Results	6
2.3	Alignment Tax and Safety Neuron Evidence	7
2.4	The Depth-Robustness Gradient	8
2.5	Machine Unlearning: Harder Than Never Learning	10
2.6	Data-Level Determinants of Robustness	10
2.7	Monitoring Value Drift	11
2.8	Limitations and Counterarguments	11
2.9	The Missing Experiment	12
3	Scaling and Efficiency of Alignment Data	13
3.1	The Missing Chinchilla Law	13
3.2	The Exponential Plateau: When Less Suffices	14
3.2.1	Saturation at Less Than 10% of Data	14
3.2.2	Ten Examples Can Suffice	14
3.2.3	The Superficial Safety Alignment Hypothesis	15

3.3	Reward Model Overoptimisation	15
3.3.1	The Gao et al. Foundation	15
3.3.2	Direct Alignment Algorithms Are Not Immune	15
3.4	RLHF Scales Less Efficiently Than Pretraining	16
3.5	The “Overly Safe” Problem	16
3.6	The Geometry of the Alignment Tax	17
3.7	Complicating Factors	17
3.7.1	Inverse Scaling	17
3.7.2	Synthetic Data Ceilings	17
3.7.3	When Bad Data Leads to Good Models	18
3.7.4	Sample Complexity Bounds	18
3.8	Synthesis	18
4	The Multilingual Frontier	18
4.1	Measuring the Gap	19
4.2	Cross-Lingual Transfer: Real but Shallow	20
4.3	The Low-Resource Cliff	21
4.4	Code-Switching as Attack Vector	22
4.5	Cultural Safety Cannot Be Translated	22
4.6	Mitigation Strategies	22
4.6.1	Preference Optimization Across Languages	22
4.6.2	Representation-Based Alignment	23
4.6.3	Multilingual Safety Data Construction	23
4.7	Missing Experiments and Open Questions	24
5	Principled Corpus Composition	25
5.1	Three Generations of Data Mixing	25
5.2	Data Mixing Laws	26
5.3	Domain Reweighting via Proxy Models	26
5.4	Quality Filtering: The Compute Multiplier and Its Limits	27
5.5	Dynamic Mixing and Curriculum Learning	27
5.6	Connecting Composition to Geometry	28
5.7	The Corpus as Attack Surface	28
5.8	The Central Gap: Alignment-Aware Optimization	29
5.9	Missing Experiments	29
6	Verification and Governance	30
6.1	A Maturity Model for Corpus Auditing	30
6.2	The Documentation Crisis	31
6.3	Corpus Security	31
6.4	The Regulatory Gap	31
6.5	Missing Infrastructure	31
7	A Concrete Experimental Agenda	32
7.1	Tier 1: Foundational Experiments	32
7.1.1	Experiment 1: Safety Basin Geometry of Corpus-Aligned vs. RLHF-Only Models	32
7.1.2	Experiment 2: Head-to-Head Stability Under Adversarial Fine-Tuning	33
7.1.3	Experiment 3: Alignment Mixing Laws	33
7.1.4	Experiment 4: Multilingual Safety Transfer as a Function of Scale and Language Distance	34
7.2	Tier 2: Medium-Term Experiments	35

7.2.1	Experiment 5: Multi-Objective DoReMi for Capability and Alignment . .	35
7.2.2	Experiment 6: Critical Depth Threshold for Irremovable Alignment . . .	35
7.2.3	Experiment 7: Chinchilla for Alignment	36
7.2.4	Experiment 8: Optimal Safety Curriculum Accounting for Learning Rate Decay	36
7.2.5	Experiment 9: Corpus-Level and Post-Hoc Alignment Interaction	37
7.3	Tier 3: Long-Term Research Programmes	37
7.3.1	Experiment 10: Unified Corpus Safety Assessment Framework	38
7.3.2	Experiment 11: Verifiable Pretraining	38
7.3.3	Experiment 12: Culturally-Grounded Multilingual Safety Without English Mediation	38
7.3.4	Experiment 13: Alignment-Aware Quality Metrics	39
7.4	Summary	39
8	Conclusion	39

1 Introduction

In a preceding whitepaper (Müller, 2026), we surveyed 248 papers and introduced *corpus-level alignment* as a coherent research area, mapping data-centric interventions onto the training pipeline and identifying six open research gaps. That survey was deliberately descriptive: it catalogued the landscape but did not build a sustained analytical argument.

This paper supplies the analytical backbone. Drawing on over 370 additional papers published since the whitepaper, we argue that *alignment is primarily a data problem*: the composition of the pretraining corpus sets a structural ceiling on safety that no amount of post-hoc intervention can exceed. Post-hoc alignment (RLHF, DPO, instruction tuning) is *structurally shallow*, encoding safety in a low-rank, high-curvature subspace of parameter space that is vulnerable to fine-tuning, adversarial attack, and even benign downstream use. Corpus-level alignment, by contrast, has the potential to distribute safety across the model’s representational geometry in ways that are qualitatively more robust. Readers unfamiliar with the landscape of data-centric alignment are encouraged to consult the whitepaper for background; the present paper assumes familiarity with its taxonomy and builds directly on the six gaps it identified.

Six research gaps structure this paper, each representing a frontier where corpus-level alignment theory meets open empirical questions:

1. **Geometric foundations.** How does the representational geometry of safety differ between corpus-aligned and post-hoc-aligned models? What are the formal limits of post-hoc alignment?
2. **Scaling laws for alignment data.** How do safety data requirements scale with model size, and what is the optimal allocation of safety data across pipeline stages?
3. **The multilingual gap.** Safety guardrails degrade sharply outside English. Can corpus-level interventions close this gap more durably than multilingual RLHF?
4. **Principled corpus composition.** What is the optimal mix of content types, safety data, and synthetic augmentation for aligned pretraining?
5. **Verification and governance.** How can alignment corpora be audited, certified, and governed at scale?
6. **Experimental validation.** What controlled experiments would convert the current suggestive evidence into definitive empirical results?

Paper outline. Section 2 establishes the geometric foundations and empirical evidence: the low-rank safety subspace, the quartic scaling law, the alignment tax, safety neurons, the depth-robustness gradient, and the data-level determinants of guardrail durability. Section 3 surveys scaling laws for alignment data, from reward model overoptimisation to the exponential plateau in safety data efficiency. Section 4 examines the cross-lingual gap and corpus-level interventions for multilingual safety. Section 5 addresses the open problem of principled corpus composition. Section 6 surveys the emerging infrastructure for auditing and verifying alignment corpora. Section 7 synthesises these threads into a concrete experimental agenda. Section 8 concludes.

2 Geometric Foundations and Empirical Evidence

Throughout this paper, we use “alignment” primarily in the narrow sense of safety alignment: the model’s tendency to refuse harmful requests, avoid toxic outputs, and resist adversarial manipulation. We distinguish this from the broader senses of value alignment (encoding specific

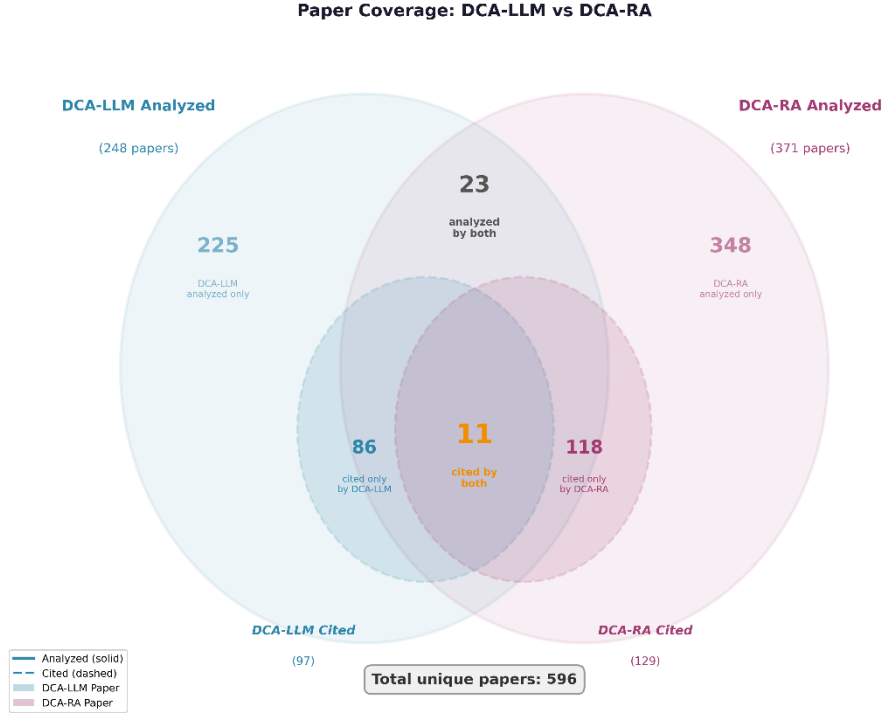


Figure 1: Paper coverage across both publications. The DCA-LLM paper (“Data-Centric Alignment of Large Language Models,” a survey of corpus-level safety techniques) analysed 248 papers and cited 97. The present DCA-RA paper (“Beyond Post-Hoc Alignment: A Data-Centric Research Agenda”) analysed 371 papers and cited 129. Only 11 papers are cited by both and 23 appear in both reading corpora, confirming complementary coverage. Solid ellipses = papers analysed; dashed ellipses = papers cited.

ethical values) and instruction-following (the LIMA sense of helpful compliance). While the geometric framework we present may extend to these broader senses, the empirical evidence we survey is overwhelmingly about safety, and our claims should be interpreted accordingly.

Since the whitepaper, a coherent geometric picture has emerged from independent research in gradient analysis, mechanistic interpretability, and mathematical optimisation theory: safety alignment occupies a low-rank, high-curvature subspace of parameter space, while general capabilities span a much broader manifold. Safety behaviours require precise coordination across a small number of representational dimensions; capabilities distribute across many. This structural asymmetry—safety as a “thin manifold” in a vast capability space—makes it vulnerable to any perturbation reaching that subspace, and provides the foundations on which all subsequent sections build.

We present this geometry in subsections 2.1–2.3, then trace its empirical consequences through the depth-robustness gradient (2.4), machine unlearning (2.5), data-level determinants (2.6), monitoring (2.7), limitations (2.8), and the missing experiment (2.9).

2.1 Low-Rank Safety Subspace and Abliteration Evidence

The empirical foundation comes from Zhang et al. (2026), who performed SVD on the gradients associated with safety and utility objectives across three model families (Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, Qwen2.5-7B-Instruct). Safety gradients exhibit sharply decaying singular value spectra: in Llama-3.1-8B-Instruct, the top 10 singular values capture 85% of total gradient variance, rising to 96% at rank 100. Math and code gradients are far flatter (45% and 35% CER at rank 10).

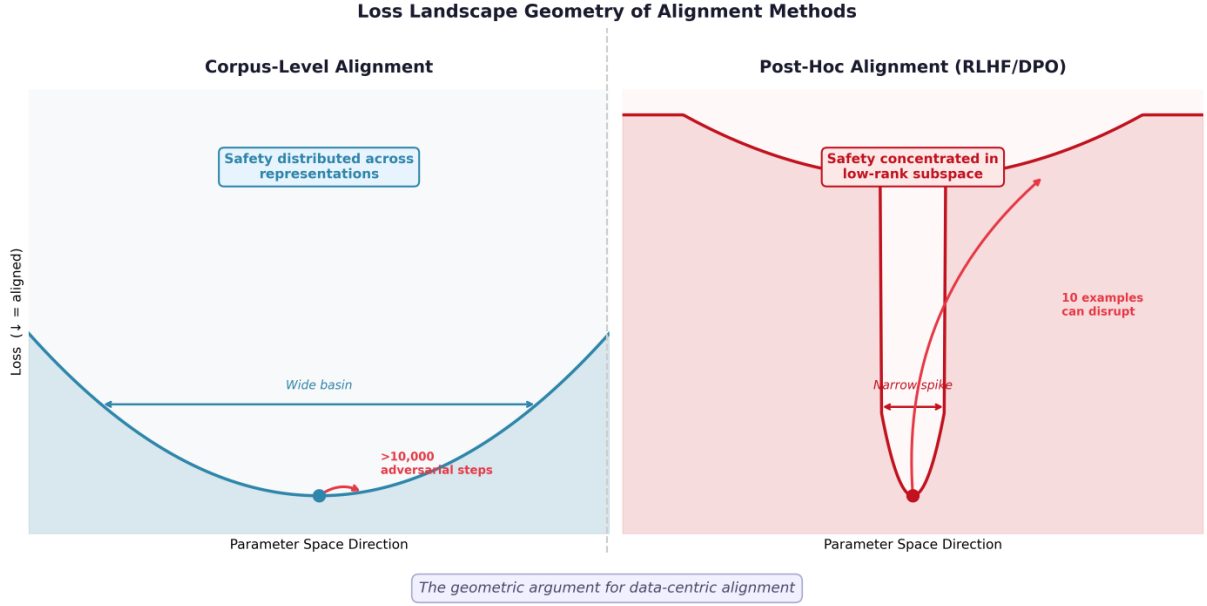


Figure 2: The geometric argument for data-centric alignment. Left: corpus-level alignment distributes safety across a wide, flat basin in parameter space—adversarial fine-tuning requires $>10,000$ steps to escape. Right: post-hoc alignment (RLHF/DPO) concentrates safety in a narrow, high-curvature subspace that 10 adversarial examples can disrupt.

Three consequences follow. First, the dominant safety direction can be estimated from a single sample (cosine similarity > 0.8 between single-sample and full-batch safety gradients). Second, safety and utility subspaces are often negatively correlated: the step-wise cosine similarity $\cos(\mathbf{g}^s, \mathbf{g}^u)$ frequently drops below zero, meaning that optimisation for downstream utility persistently pushes parameters out of the safety region, even when the fine-tuning data is entirely benign. Third, safety failure under fine-tuning is monotonic and universal: across all tested models, datasets, and defenses, ASR increases with training duration, reaching 1.000 for undefended SFT within 20 epochs.

Independent confirmation comes from sparse autoencoders: Wang et al. (2025c) achieve a 99.6% safety rate while modifying only 0.19% of parameters, consistent with Zhang et al.’s gradient-based finding that safety concentrates in a geometrically compact region.

The most vivid demonstration of this compactness is *abliteration* (Arditi et al., 2024). Across 13 open-source chat models (1.8B–72B parameters), refusal behaviour is mediated by a single direction in the residual stream; a rank-one weight edit nearly eliminates refusal with minimal impact on other capabilities. However, subsequent work reveals that this single direction captures only the dominant mode of a richer structure. Pan et al. (2025) define the *safety residual space* and show that it contains multiple orthogonal directions: a dominant refusal/compliance direction, plus secondary directions for hypothetical framing and role-playing context. Wollschläger et al. (2025) independently decompose refusal into harm-detection and refusal-execution directions. Extended-refusal training, which distributes the refusal signal across multiple token positions, maintains over 90% refusal under abliteration compared to 20–30% for standard training (Abu Shairah et al., 2025), demonstrating that the breadth of safety encoding directly determines robustness.

2.2 Quartic Scaling Law and Fine-Tuning Attack Results

If safety occupies a low-rank subspace and utility gradients are approximately orthogonal to it, one might hope that fine-tuning for utility would leave safety untouched. Springer et al. (2026) show this hope is mathematically unfounded.

Their **Alignment Instability Condition** (AIC) identifies three interacting geometric properties that make alignment collapse structurally inevitable: (i) *low-rank sensitivity*: alignment concentrates in a subspace M_l with sharp curvature (large Fisher eigenvalues λ); (ii) *initial orthogonality*: fine-tuning gradients have negligible first-order projection onto M_l , creating the illusion of safety; (iii) *curvature coupling*: the acceleration $\nabla g(\theta^*)g(\theta^*)$ has non-trivial projection onto M_l , systematically steering the trajectory into alignment-sensitive directions despite the initial orthogonality.

Under the AIC, the projection onto M_l grows quadratically in training time:

$$\|P_l(\theta(t) - \theta^*)\| = \Omega(\gamma t^2), \quad (1)$$

where γ quantifies curvature coupling strength. Combined with the quadratic utility loss $\Delta u_i = \Omega(\lambda \|P_l(\Delta\theta)\|^2)$ for movement within M_l , this yields the **quartic scaling law**:

$$\Delta u_i = \Omega(\lambda \gamma^2 t^4). \quad (2)$$

The fourth-power dependence means a trajectory that appears safely orthogonal at $t = 1$ can exhibit catastrophic failure by $t = 10$.

Crucially, this is an asymptotic lower bound (Ω), not an observed rate. Whether real models exhibit quartic degradation in practice, as opposed to quadratic or cubic, remains to be validated empirically. The bound also assumes approximately constant curvature coupling γ over training, which may not hold as the model traverses different regions of parameter space. We treat the quartic law as a worst-case prediction that motivates caution, not as a precise description of finite-time behaviour. First-order defenses (gradient filtering, null-space projection, static loss reweighting) operate only on $\|P_l g\|$, which is small by assumption, and are blind to the second-order coupling γ that drives the collapse.

The attack landscape confirms these predictions. Adversarial fine-tuning is cheap and universal: 10 harmful examples costing \$0.20 suffice to jailbreak GPT-3.5 Turbo (Qi et al., 2023); 100 malicious examples with one GPU-hour subvert safety across eight models from five organisations (Yang et al., 2023). Benign fine-tuning is equally dangerous: overfitting on 10 benign question-answer pairs jailbreaks strongly aligned models (Xie et al., 2025b), and fine-tuning on GSM8K, SQL, or summarisation data produces measurable ASR increases (Zhang et al., 2026; Hsiung et al., 2025), confirming that the vulnerability lies in gradient geometry, not data content. Emulated disalignment (Zhou et al., 2024) shows that safety alignment can even be reversed at inference time without any training, by minimising rather than maximising the implicit reward captured by the log-probability difference between aligned and base models.

2.3 Alignment Tax and Safety Neuron Evidence

Young (2026) provides the first mathematical framework for the alignment tax. The alignment tax rate is defined as:

$$\tau = \|P_C v^*\|^2 \in [0, 1], \quad (3)$$

where v^* is the safety direction, P_C projects onto the capability subspace $C = \text{span}(c_1, \dots, c_m)$, and $\tau = 0$ means safety is orthogonal to all capabilities (zero cost) while $\tau = 1$ means every safety gain requires a capability loss. The Pareto frontier is elliptic, parameterised by the principal angle α :

$$\Delta_S = \Delta_C \cos \alpha + \sin \alpha \sqrt{B^2 - \Delta_C^2}, \quad |\Delta_C| \leq B. \quad (4)$$

Critically, the tax decomposes as $\tau = \tau_0 + m'/d + R(d)$, where $\tau_0 = \sum_{i \in I} \gamma_i^2$ is the *irreducible tax* determined by intrinsic overlaps between safety and capability features in the data, m'/d is a leading packing term (with m' the number of incidentally overlapping capabilities and d the model dimension), and $R(d)$ is a residual that vanishes as $O(\log d / \sqrt{d})$ under random

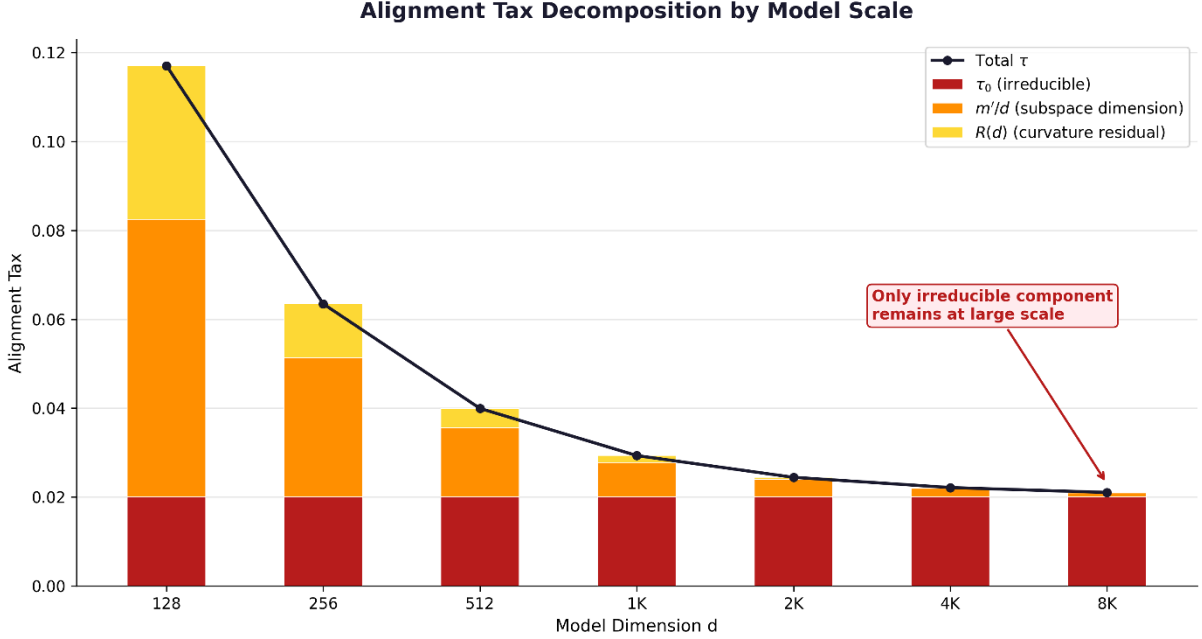


Figure 3: Alignment tax decomposition by model scale. The total alignment tax τ decomposes into an irreducible floor τ_0 , a subspace-dimension term m'/d , and a curvature residual $R(d)$. As model dimension d grows, only the irreducible component remains.

packing. This means: (i) larger models naturally face a lower alignment tax, explaining the empirical observation that scaling helps alignment (Hou et al., 2024; Rafailov et al., 2024); (ii) the irreducible tax is set primarily by corpus structure, so a corpus reducing safety-capability feature overlap would structurally lower the tax before any post-training; (iii) per-task tax rates are measurable via probing, yielding falsifiable predictions about which tasks have intrinsic versus incidental overlap with safety.

This decomposition rests on three assumptions: (i) the linear representation hypothesis, under which safety and capabilities correspond to directions in a linear space; (ii) a random packing model for capability directions, where their inner products with the safety direction are approximately independent; and (iii) that τ_0 is determined primarily by the training data rather than model architecture. Assumption (iii) is the most contentious: since feature encoding is a joint function of data and architecture, τ_0 may be architecture-dependent in ways not captured by the data-only framing. We use the decomposition as a conceptual tool while acknowledging these limitations.

Safety neurons provide the mechanistic substrate. Chen et al. (2024) identify sparse safety neurons ($\sim 5\%$ of MLP neurons restore $>90\%$ of safety) that overlap with helpfulness neurons but require different activation patterns—a mechanistic explanation for the alignment tax as competition for shared resources rather than distinct neural resources. Crucially, these neurons are present in base models *before any safety training*: alignment changes activation patterns rather than creating circuits *de novo*, implying that the pretraining corpus shapes the geometry of safety in parameter space. A corpus with broadly distributed safety-relevant patterns would produce more numerous, more distributed, and harder-to-ablate safety neurons.

2.4 The Depth-Robustness Gradient

The geometric theory predicts that earlier safety interventions should produce broader, lower-curvature safety subspaces that resist perturbation more effectively. Empirical evidence confirms

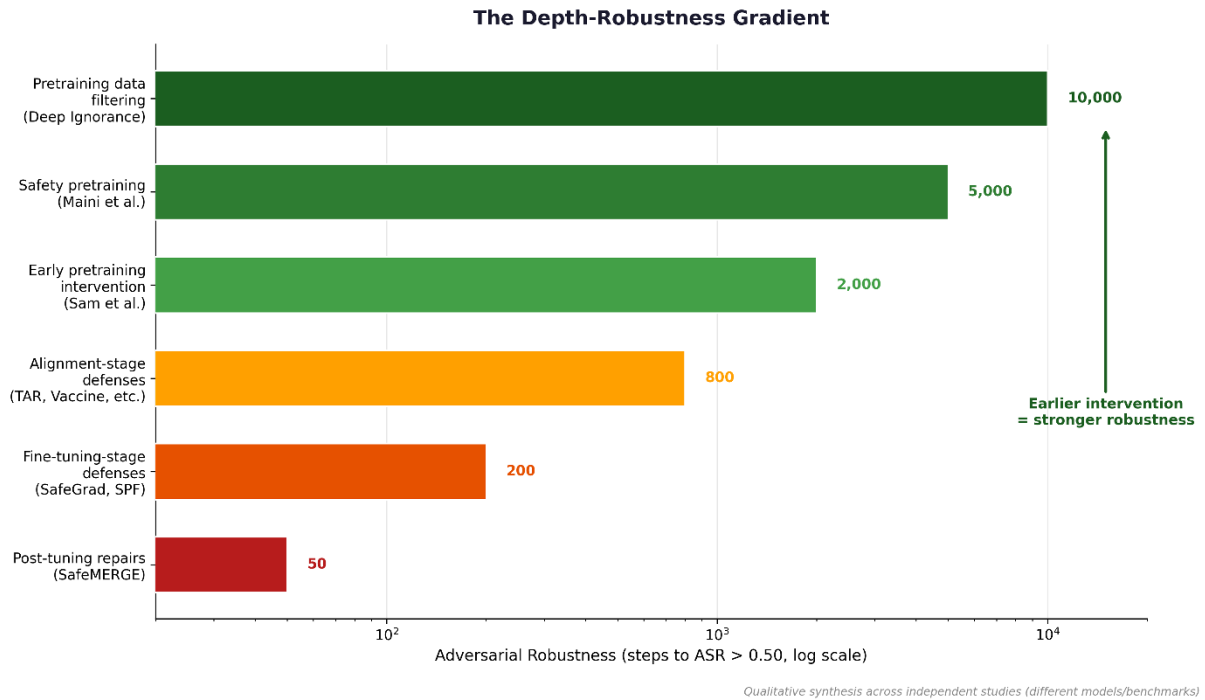


Figure 4: The depth-robustness gradient. Adversarial robustness (steps to ASR > 0.50) increases roughly by an order of magnitude per intervention depth. Values are a qualitative synthesis across independent studies using different models and benchmarks.

this as a consistent hierarchy, with each deeper layer of intervention adding roughly an order of magnitude of adversarial robustness.

The following gradient is a qualitative synthesis across independent studies with different models, threat models, and evaluation protocols. The specific step counts should be understood as indicative rather than precise, since no single study has compared all intervention depths under identical conditions.

Pretraining data interventions: >10,000 adversarial steps (on biothreat proxy knowledge specifically). O’Brien & Casper (2025) train language models on corpora from which harmful content has been filtered (*Deep Ignorance*). The resulting models resist adversarial fine-tuning for over 10,000 gradient steps and 300 million tokens of adversarial data, outperforming post-training baselines by approximately an order of magnitude. The conceptual advantage is fundamental: a model that never encoded harmful knowledge has nothing to suppress, eliminating the vulnerability that arises when post-training refusal overlays an intact generative distribution over harmful content. An important scope caveat: O’Brien et al.’s experiments focus on biothreat proxy knowledge; whether the approach generalises to other harm categories (e.g., cyberattack, social manipulation) with less clearly delineated knowledge boundaries remains untested. O’Brien et al. further report “useful synergies” when data filtering is combined with post-training safeguards, suggesting that these interventions operate at different representational depths.

Maini et al. (2025) take the complementary approach of *augmenting* the pretraining corpus with synthetic safety data: contextualised rephrasings, metadata annotations, and refusal pairs embedded in the pretraining mixture. This *Safety Pretraining* reduces ASR from 38.8% to 8.4% with no measurable capability degradation, and the robustness persists after benign fine-tuning. The zero capability loss contrasts sharply with the alignment tax typically incurred by post-training methods, though Maini et al. note elevated over-refusal rates on borderline queries, a trade-off between false negatives (unsafe completions) and false positives (unnecessary refusals)

that remains to be optimised.

Sam et al. (2026) establish that intervention *timing* within pretraining also matters. Models with safety interventions from the start achieve ASR of approximately 6% after downstream fine-tuning, versus approximately 50% for late-intervention baselines, a factor of eight (values estimated from figures in Sam et al. 2026). Linear probes achieve higher accuracy when interventions are introduced earlier, indicating deeper representational changes.

Alignment-stage defenses: hundreds to low thousands of steps. TamperBench (Hossain et al., 2026) evaluates 21 LLMs across nine attack types and finds that alignment-stage methods (TAR (Tamirisa et al., 2024), Vaccine (Huang et al., 2024), RepNoise (Rosati et al., 2024), LAT (Sheshadri et al., 2024)) provide protection measured in hundreds to low thousands of adversarial steps; even the best defense (Triplet (Simko et al., 2025)) does not reach the thousands-of-steps range reliably.

Fine-tuning-stage and post-tuning defenses: monotonic degradation. Gradient surgery (SafeGrad (Yi et al., 2025), SPF (Zhang et al., 2026)), LoRA isolation (Mou et al., 2025a), safety-aware probing (Wu et al., 2025a), basin-constrained fine-tuning (Yang et al., 2025), and continual-learning methods (Alssum et al., 2025) provide per-session protection that degrades monotonically. Post-tuning repairs (SafeMERGE (Djuhera et al., 2025), AlignMerge (Roy et al., 2025), NLSR (Yi et al., 2024), GuardSpace (Zhang et al., 2025), Panacea (Wang et al., 2025a)) are reactive, restoring safety after damage but providing no prospective protection. Across all defenses, Zhang et al. show that ASR increases monotonically with epochs; the best defenses slow the rate but cannot halt it, because the underlying geometric conditions (the AIC; see Eq. 2) remain in force.

Deceptive alignment as boundary condition. The sleeper agent results of Hubinger et al. (2024) define a limit: backdoor behaviours persist through all standard safety training, and adversarial training can teach models to better *conceal* triggers. Corpus-level alignment can set value priors and reduce the likelihood of deceptive emergence, but it cannot guarantee the absence of deception; complementary approaches from mechanistic interpretability (Bereska & Gavves, 2024) and formal verification remain necessary.

2.5 Machine Unlearning: Harder Than Never Learning

Machine unlearning aims to erase specific knowledge from a trained model, providing a stronger guarantee than refusal. The empirical evidence reveals fundamental limitations: Lucki et al. (2024) show that “unlearned” capabilities can be recovered via fine-tuning on as few as 10 unrelated examples, and Xu et al. (2025) confirm through representation-level analysis that models appear to forget while retaining easily restorable original behaviour. For dual-use knowledge, the same representations encoding dangerous capabilities also encode legitimate understanding, requiring a precision of selective erasure that current methods do not achieve (Barez et al., 2025; Shi et al., 2025).

This reinforces a key asymmetry: preventing harmful knowledge from entering during pre-training is fundamentally easier than removing it, because neural network encoding is distributive and entangled while selective removal requires precisely the disentanglement that high-dimensional representations resist.

2.6 Data-Level Determinants of Robustness

Two findings connect alignment robustness directly to training data properties.

High similarity weakens safety. Hsiung et al. (2025) demonstrate that high representation similarity between alignment data and downstream fine-tuning data *weakens* guardrails, with ASR elevated by up to 10.33% compared to low-similarity configurations. The mechanism is geometric: when fine-tuning data occupies the same representational neighbourhood as alignment data, gradient updates directly modify safety-relevant representations. Corpus-level alignment, by distributing safety content throughout pretraining rather than concentrating it in a separate alignment dataset, inherently reduces this exploitable similarity.

Alignment failures trace to training-time belief conflicts. Das et al. (2025) introduce TRACEALIGN, which traces unsafe completions back to training-time belief sources using suffix-array matching. Their Belief Conflict Index and three interventions reduce alignment drift by up to 85% while maintaining utility ($\Delta\text{PPL} < 0.22$). The key insight is that alignment failures are epistemic events in which the model resurfaces memorised beliefs from its training corpus that conflict with aligned policy, reframing alignment drift as a data provenance problem where corpus-level interventions have the most leverage.

2.7 Monitoring Value Drift

Fadli (2025) operationalise *ethical entropy* as a measurable state variable. Base pre-trained models exhibit ethical entropy of $S_{\text{final}} = 0.70 \pm 0.04$ nats, while instruction-tuned models stabilise at 0.12 ± 0.03 nats, an 83% reduction. The effective alignment work rate γ_{eff} is proposed as a preliminary metric for measuring the ongoing effort required to maintain alignment. This is a single-author preprint; the practical significance of the measured $\gamma_{\text{eff}} \approx 0.012$ nats/step remains to be established through independent replication. This framework generates a testable prediction: corpus-aligned models should exhibit lower baseline ethical entropy and slower drift rates than post-hoc-aligned models, because corpus-level alignment distributes value priors throughout the network’s representations.

2.8 Limitations and Counterarguments

The evidence presented above, while substantial, rests on assumptions and simplifications that deserve explicit acknowledgment.

Deep Ignorance and Safety Pretraining are different mechanisms, not a single “corpus-level” intervention. Deep Ignorance works by *knowledge absence*: the model never learns harmful content. Safety Pretraining works by *knowledge augmentation*: the model learns harmful content framed with safety context. These are conceptually distinct, and their robustness may arise from different geometric mechanisms. Grouping them under a single “corpus-level alignment” umbrella, as we do throughout this paper, risks conflating interventions that may not compose or generalise in the same ways.

Scaling may solve the problem without corpus-level intervention. The alignment tax’s packing terms (m'/d and $R(d)$) vanish with model dimension (Eq. 3), and larger models empirically handle alignment more gracefully (Hou et al., 2024). If the irreducible tax τ_0 is small for most practical safety-capability pairs, then simply scaling models may reduce the alignment tax to negligible levels without any corpus-level engineering. The counter-argument, that τ_0 is set by data structure and may be substantial for critical safety dimensions, is plausible but unverified.

The “order of magnitude per layer” gradient rests on heterogeneous evidence. The depth-robustness gradient compares results across different papers, model families, evaluation protocols, and threat models. Deep Ignorance’s 10,000-step figure comes from one paper with

one model family; alignment-stage defense numbers come from different papers with different models. A controlled comparison holding all variables constant except intervention depth has not been performed.

Catastrophic forgetting as alternative explanation. The robustness of Deep Ignorance models may reflect catastrophic forgetting dynamics rather than deeper geometric encoding: a model that never learned harmful content may simply have no “attractor” for adversarial fine-tuning to converge toward, making the task harder for reasons unrelated to safety basin geometry.

This alternative merits careful consideration. Catastrophic forgetting predicts that pretraining-stage safety should be more robust simply because pretraining representations are more deeply embedded in the loss landscape, regardless of whether they occupy a geometrically wider basin. The two explanations make different predictions: the geometric hypothesis predicts that safety basin width (measurable via Hessian analysis) will differ between corpus-aligned and post-hoc-aligned models; the forgetting hypothesis predicts that robustness scales with the number of gradient updates since the safety signal was introduced, independent of basin geometry. Experiment 1 in Section 7 is designed specifically to distinguish these alternatives.

Li et al.’s toxic data paradox. Li et al. (2025) present evidence that including small amounts of toxic data in pretraining can *improve* safety, suggesting that the relationship between corpus composition and alignment is non-monotonic.

This finding poses a genuine challenge to the simplest version of the corpus-centric thesis. If models pretrained on more toxic data develop representational structures that are subsequently easier to align, then the relationship between corpus composition and alignment is more nuanced than “cleaner data produces safer models.” The geometric interpretation we offer (that toxic exposure may increase principal angles, creating more separable safety subspaces) is speculative and untested. Until controlled experiments vary toxic content proportion while measuring both alignment outcomes and subspace geometry, the toxic data paradox remains an open challenge to our framework.

The irreducible alignment tax may be substantial. Young’s decomposition (Eq. 3) shows that τ_0 is determined by intrinsic feature co-occurrence in the data. For safety dimensions that are deeply entangled with capabilities (e.g., chemical synthesis knowledge that is simultaneously dangerous and scientifically valuable), τ_0 may be large and irreducible by corpus design alone. No empirical measurement of τ_0 for real safety-capability pairs has been published.

Practical barriers: cost and access. Only frontier labs control pretraining data, making corpus-level alignment inaccessible to the broader research community and most deployers. The cost of pretraining experiments dwarfs that of RLHF/DPO iterations. Post-hoc methods remain the only practical option for the vast majority of alignment practitioners, regardless of their theoretical limitations.

2.9 The Missing Experiment

The geometric framework is now mature enough to state precisely what experiment is needed: a **direct, controlled comparison of the safety basin geometry in corpus-aligned versus post-hoc-aligned models.**

The experiment requires matched model pairs: two models identical in architecture and pretraining compute, differing only in whether safety was embedded through corpus composition or applied post-hoc via RLHF/DPO. Both would be subjected to identical fine-tuning regimes, with geometric properties measured at each stage using existing tools: SVD gradient analysis

(Zhang et al., 2026), the AIC framework (Springer et al., 2026), safety basin analysis (Yang et al., 2025), and ethical entropy monitoring (Fadli, 2025).

The specific predictions:

1. Corpus-aligned models exhibit *higher effective rank* in the safety gradient subspace (safety distributed more broadly, not concentrated in a thin shell).
2. Corpus-aligned models exhibit *lower curvature* λ in alignment-sensitive directions (wider safety basin).
3. Corpus-aligned models exhibit *weaker curvature coupling* γ (less acceleration into safety-sensitive regions).
4. The quartic collapse rate $\lambda\gamma^2t^4$ is measurably slower, with the gap widening over training time.

If confirmed, these results would establish corpus-level alignment as a geometrically grounded strategy with formal advantages. If falsified (that is, if both models share the same safety basin geometry), the depth-robustness gradient observed in practice must arise from a mechanism other than geometric widening (e.g., catastrophic forgetting dynamics), requiring alternative explanations that would be equally informative.

This experiment is, in our view, the single most informative study the field could undertake. The measurement infrastructure is ready. What remains is the will and the compute to run it.

3 Scaling and Efficiency of Alignment Data

Pretraining has its Chinchilla laws: given a compute budget, practitioners can consult well-validated power laws to determine how many parameters and tokens to use (Hoffmann et al. (2022); Kaplan et al. (2020)). No analogous framework exists for alignment. When a laboratory doubles its flagship model’s parameter count, it cannot consult any formula to determine whether its safety training dataset remains adequate, dangerously insufficient, or wastefully excessive. This section surveys the converging evidence that constrains what such a law must look like, drawing on overoptimisation dynamics, data-efficiency studies, reward model scaling, and the geometric theory of the alignment tax introduced in Section 2.

The stakes are asymmetric. Too little alignment data risks deploying models whose capabilities outpace their safety guardrails. Too much, or the wrong kind, produces models that refuse benign requests (Tan et al. (2025)), degrade reasoning performance (Huang et al. (2025)), or waste compute better allocated elsewhere. Understanding the scaling properties of alignment data is therefore a prerequisite for responsible deployment at the frontier.

3.1 The Missing Chinchilla Law

The Chinchilla scaling law (Hoffmann et al. (2022)) expresses pretraining loss as a smooth, monotonically decreasing function of model size N and dataset size D :

$$\mathcal{L}(N, D) = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta}, \quad (5)$$

where E is an irreducible loss and A, B, α, β are fitted constants. This functional form has proven remarkably robust across model families, training configurations, and modalities (Shukor et al. (2025); Isik et al. (2024)), turning model training from an art into a rough science.

For alignment, no such formula exists. The reasons are partly practical (alignment datasets are smaller and more expensive to curate) and partly fundamental: alignment performance does not follow the monotonically decreasing power laws of pretraining. Instead, it exhibits

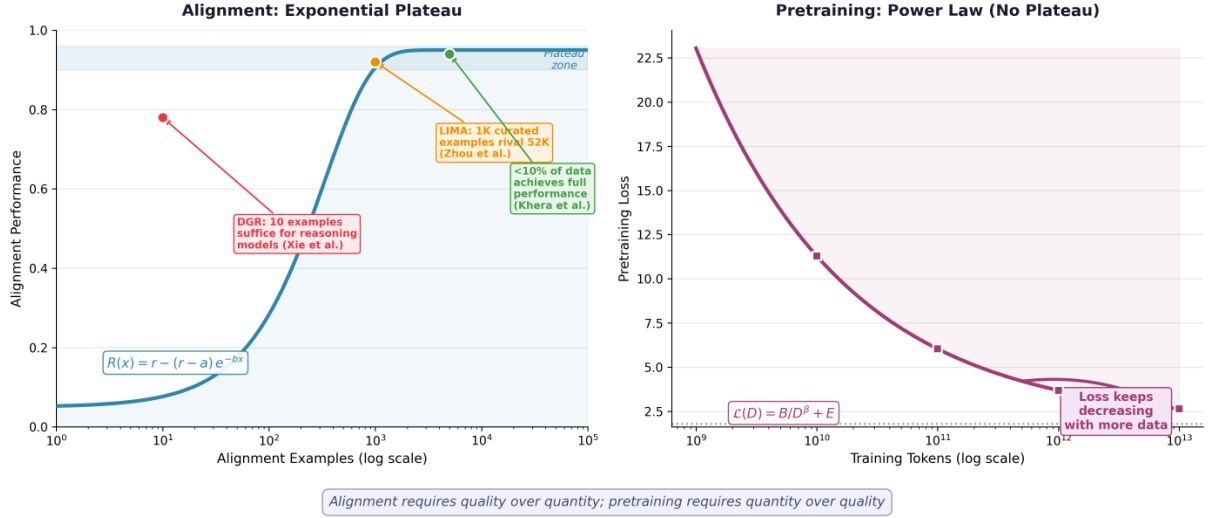


Figure 5: The fundamental scaling contrast. Left: alignment performance follows an exponential plateau $R(x) = r - (r - a)e^{-bx}$ that saturates rapidly—LIMA achieves near-ceiling performance with 1,000 examples, and DGR shows 10 examples suffice for reasoning models. Right: pretraining loss follows a power law $\mathcal{L}(D) = B/D^\beta + E$ that continues to decrease with more data, never saturating. This asymmetry explains why alignment requires data quality over quantity, while pretraining requires quantity.

saturation, non-monotonicity, and composition dependence that demand qualitatively different functional forms. The remainder of this section assembles the empirical and theoretical building blocks from which a Chinchilla law for alignment might eventually be constructed.

3.2 The Exponential Plateau: When Less Suffices

3.2.1 Saturation at Less Than 10% of Data

Khera, Ghosh, and Dutta [Khera et al. \(2024\)](#) demonstrated that alignment performance follows an *exponential plateau* rather than the power laws characteristic of pretraining:

$$R(x) = r - (r - a)e^{-bx}, \quad (6)$$

where x is the number of training examples, r is the asymptotic performance, a is the initial performance, and b is a rate parameter. This curve rises steeply with small amounts of data and then flattens, implying that alignment performance saturates much earlier than pretraining loss. Using information-theoretic entropy sampling, Khera et al. showed that carefully selected subsets comprising less than 10% of the full dataset achieve alignment performance comparable to training on all available data, consistently across OpenAssistant, UltraFeedback, and Anthropic HH-RLHF.

The practical implication (greater than 90% cost savings in data curation) is substantial, but the deeper theoretical implication is equally important: alignment may be a *low-dimensional* intervention on a pretrained model, requiring far less information than pretraining. This is consistent with the geometric picture of Section 2, where safety occupies a low-rank subspace of the representation space.

3.2.2 Ten Examples Can Suffice

The data-efficiency results become even more dramatic for large reasoning models specifically. Xie et al. [Xie et al. \(2026\)](#) found that as few as *10 safety examples* suffice to activate effective refusal behaviours, with their Distribution-Grounded Refinement (DGR) method achieving

+30.2% improvement on DirectRefusal and +21.2% on R1-ACT benchmarks compared to vanilla SFT while preserving reasoning capability. Crucially, the degree of reasoning capability degradation correlates with the extent of distributional shift between alignment data and the model’s internal distribution: when safety data is rephrased by the target model itself, the tax largely disappears. Whether this extreme efficiency transfers to standard instruction-tuned models without extended chain-of-thought capabilities remains untested.

This finding suggests that safety alignment in sufficiently capable models functions primarily as a mechanism for activating latent knowledge already encoded during pretraining, rather than teaching genuinely new behaviours. The connection to Section 2’s finding that safety neurons pre-exist in base models is direct: if the relevant circuits are already present, alignment data need only *activate* them, not *create* them.

3.2.3 The Superficial Safety Alignment Hypothesis

This view receives further support from the Superficial Safety Alignment Hypothesis (SSAH) proposed by Li and Kim [Li & Kim \(2024\)](#). Through attribution-based pruning experiments, they identified four types of safety-critical computational units and found that freezing just **7.5%** of a model’s safety-critical units during fine-tuning suffices to retain safety, while approximately **20%** of redundant units in the pretrained model can be repurposed as an “alignment budget” without significant capability loss.

These results collectively paint a picture of safety alignment as a surprisingly *thin* layer atop the capability substrate. If true, the scaling law for safety data may be far more favourable than initially feared: rather than requiring safety data proportional to model size, safety alignment may require only enough data to activate and stabilise a relatively small number of safety-critical circuits.

3.3 Reward Model Overoptimisation

The exponential plateau describes how quickly alignment data *helps*. A complementary body of work characterises when additional optimisation pressure starts to *hurt*.

3.3.1 The Gao et al. Foundation

Gao, Schulman, and Hilton [Gao et al. \(2022\)](#) provided the foundational empirical characterisation of reward model overoptimisation, establishing that gold reward follows distinct functional forms for best-of- n sampling and PPO, each exhibiting an initial increase followed by degradation. The critical finding for data scaling is a minimum data threshold: with fewer than approximately 2,000 preference comparisons, reward models showed little improvement over chance, below which alignment training is effectively useless. Larger reward models delay but do not eliminate overoptimisation, with degradation coefficients scaling smoothly with RM parameter count.

3.3.2 Direct Alignment Algorithms Are Not Immune

A natural question is whether Direct Alignment Algorithms (DAAs) such as DPO [Rafailov et al. \(2023\)](#) escape the overoptimisation trap. Rafailov et al. [Rafailov et al. \(2024\)](#) demonstrated conclusively that they do not. Evaluating DPO, IPO, and SLiC across three model sizes on the TL;DR summarisation task, they found that all three DAA objectives exhibit hump-shaped performance curves as a function of $\sqrt{D_{\text{KL}}}$, closely mirroring classical RLHF.

Several findings are particularly relevant to data scaling. First, the Gao et al. functional form fits DAA performance with roughly half the RMSE of a simple quadratic baseline, suggesting a shared underlying mechanism. Second, *performance often peaks after only approximately 25% of a single training epoch*, after which continued training degrades quality — a qualitative departure

from pretraining, where multiple epochs over high-quality data can be beneficial [Muennighoff et al. \(2023\)](#). Third, larger models exhibited substantially better win-rate–KL trade-offs, achieving higher peak performance while being less prone to overoptimisation.

The practical upshot is that even without an explicit reward model, alignment algorithms face intrinsic diminishing and eventually negative returns from additional data exposure. Any scaling law for alignment data must therefore incorporate a notion of *optimal dose*.

3.4 RLHF Scales Less Efficiently Than Pretraining

Hou et al. [Hou et al. \(2024\)](#) conducted the most comprehensive empirical investigation to date, training over 20 models at 9B, 32B, and 200B parameters across varying dataset sizes. Their central finding is blunt: *RLHF scales less efficiently than pretraining*. Three observations warrant particular attention:

- **Response sampling plateaus quickly.** Sampling 4–8 responses per prompt captures most of the gain; beyond this, benefits plateau.
- **Reward model gains transfer inefficiently.** Larger reward models improve best-of- N evaluation, but the improvement transferred to policy training falls significantly behind direct RM gains.
- **Larger policy models benefit *less* from RLHF.** Performance gains from RLHF decrease as the policy model scales from 9B to 200B parameters, suggesting larger models have already absorbed much of the knowledge RLHF tries to instil.

These findings carry a direct implication for alignment data budgets: if the marginal benefit of RLHF decreases with model size, then the relationship between alignment data quantity and model size may be *sublinear*. However, this is tempered by the possibility that current RLHF techniques are simply inefficient and that better algorithms could unlock steeper scaling.

Complementing these results, Shen et al. [Shen et al. \(2025\)](#) identified reward hacking and deterioration of response diversity as the primary bottlenecks limiting RLHF performance as data scales. Notably, prioritising mathematical and coding tasks during early RLHF stages yields disproportionate gains, as these domains resist reward hacking — reinforcing the importance of data *composition* over data *volume*.

3.5 The “Overly Safe” Problem

If the exponential plateau and overoptimisation results suggest that alignment needs less data than expected, the “overly safe” phenomenon demonstrates that *the wrong amount or kind of safety data is actively harmful*.

Tan et al. [Tan et al. \(2025\)](#) showed that naively increasing safety training data volume leads to an “overly safe” state rather than a “truly safe” one, with models declining to answer benign queries while failing to develop genuine safety understanding. Their decomposition of safety data into three functionally distinct categories is particularly instructive:

- **Explicit Harmful Data (EHD):** factual harmful information where performance depends primarily on the model’s knowledge base; safety scores show a consistently *upward* trend with more data, suggesting alignment alone never fully handles these cases.
- **Implicit Harmful Data (IHD):** content with malicious intent but no overt risk markers, where alignment training is most effective; safety scores rise quickly and plateau, closely matching the exponential plateau form of Equation 6.
- **Mixed Harmful Data (MHD):** ambiguous cases requiring contextual judgement, where scaling data can actually *increase* inconsistency.

Each category exhibits different scaling behaviour, implying that a meaningful scaling law for safety data cannot treat safety as a monolithic category. The fine-grained data-centric approach achieves better safety-helpfulness balance with fewer data points by selecting the right *composition* rather than the right *quantity*.

This connects directly to Section 2: if different safety categories correspond to different directions in representation space with different principal angles to the capability subspace, their scaling dynamics should indeed differ. Categories whose safety direction is nearly orthogonal to capability (e.g., IHD) admit aggressive scaling at low cost; categories whose direction overlaps significantly with capability (e.g., MHD) exhibit high tax, explaining why more data increases inconsistency.

3.6 The Geometry of the Alignment Tax

Young [Young \(2026\)](#) provides the first geometric framework for reasoning about the safety-capability trade-off, building on the representation-space foundations of Section 2. Under the linear representation hypothesis, the alignment tax rate is $\tau = \|P_C v^*\|^2$, the squared magnitude of the projection of the safety direction v^* onto the capability subspace C . The Pareto frontier governing the trade-off takes the elliptic form of Equation 4, parameterised by the principal angle α between safety and capability directions.

Most importantly for scaling, Young derives a decomposition of the alignment tax into reducible and irreducible components:

$$\tau = \tau_0 + \frac{m'}{d} + R(d), \quad (7)$$

where $\tau_0 = \sum_{i \in I} \gamma_i^2$ is the irreducible tax from intrinsic overlaps between safety and capability features, m'/d is the leading packing term, and $R(d)$ vanishes as $O(\log d/\sqrt{d})$ under random packing. The critical implication: *larger models naturally face a lower alignment tax*, because the packing residual decreases with model dimension. This provides a theoretical basis for the empirical observations that larger models handle alignment more gracefully [Rafailov et al. \(2024\)](#); [Hou et al. \(2024\)](#), and suggests that alignment data requirements may grow sublinearly, or even remain bounded, as models scale.

3.7 Complicating Factors

3.7.1 Inverse Scaling

McKenzie et al. [McKenzie et al. \(2023\)](#) documented 11 tasks across which larger language models perform *worse* than smaller ones, identifying four causes: preference for memorised sequences over instructions, imitation of undesirable patterns, distraction by easy subtasks, and misleading few-shot demonstrations.

For alignment, inverse scaling means that certain failure modes may become *more prevalent* in larger models. If safety risks are emergent, appearing only above a certain scale, then alignment data requirements may exhibit discontinuities that smooth scaling laws cannot capture. This undermines the optimistic conclusion from Section 3.6: larger models may need less alignment data *on average*, but substantially more for specific emergent failure modes.

3.7.2 Synthetic Data Ceilings

Qin et al. [Qin et al. \(2025\)](#) demonstrated that synthetic data follows “rectified” scaling laws: performance improves with data volume but plateaus near 300 billion tokens, beyond which additional synthetic data provides negligible benefit. Lee et al. [Lee et al. \(2023\)](#) validated synthetic feedback for alignment at scale, showing RLAIIF achieves performance on par with RLHF. However, the rectified scaling law implies a hard ceiling on what synthetic alignment data can achieve, even with unlimited generation budget.

3.7.3 When Bad Data Leads to Good Models

Li et al. [Li et al. \(2025\)](#) demonstrated that models pretrained on *more* toxic data can be *easier* to align post-training. Models trained on toxic corpora develop less entangled linear representations of toxicity, making the toxic-nontoxic distinction more accessible to post-training interventions and achieving better toxicity-capability trade-offs than models pretrained on clean data.

In the geometric language of Section 2, exposure to toxic data during pretraining may increase the principal angle between the safety direction and the capability subspace, reducing the irreducible component τ_0 of the alignment tax. This challenges the assumption that clean pre-training data is always optimal for downstream alignment, and implies that a complete scaling law must account for interactions between pretraining composition and post-training alignment effectiveness.

3.7.4 Sample Complexity Bounds

Chen et al. [Chen et al. \(2025a\)](#) prove that existing online RLHF algorithms suffer from sample complexity scaling *exponentially* with the reward function’s range ($\exp(R_{\max})$). Their SE-POPO algorithm achieves polynomial sample complexity instead, demonstrating that data requirements for alignment depend fundamentally on the algorithm used. Any empirically fitted scaling law is therefore conditional on the current state of algorithmic knowledge.

3.8 Synthesis

The evidence surveyed above yields several robust conclusions. First, alignment is fundamentally different from pretraining: where pretraining loss decreases monotonically, alignment performance follows non-monotonic curves — exponential plateaus in the data-efficient regime, hump-shaped overoptimisation curves in the over-training regime. Second, safety alignment appears to be a thin, low-dimensional intervention: the convergent evidence from SSAH [Li & Kim \(2024\)](#), the 10-sample activation result [Xie et al. \(2026\)](#), and the exponential plateau [Khera et al. \(2024\)](#) all point to safety as requiring far less data than capability training. Third, larger models likely need proportionally less alignment data, supported by the alignment tax’s packing residual vanishing with model dimension [Young \(2026\)](#), better win-rate–KL trade-offs at scale [Rafailov et al. \(2024\)](#), and diminishing RLHF returns for larger policy models [Hou et al. \(2024\)](#). Fourth, quality and composition dominate quantity: different safety categories scale differently [Tan et al. \(2025\)](#), distributional alignment determines the safety tax [Xie et al. \(2026\)](#), and pretraining composition interacts with post-training alignability [Li et al. \(2025\)](#).

The scaling evidence reviewed above motivates three classes of experiments: joint scaling formulae for alignment data, behavioural scaling laws for safety metrics, and systematic study of pretraining–post-training interaction effects. These are developed as concrete experimental proposals in Section 7.

4 The Multilingual Frontier

The preceding sections have established that alignment is fundamentally geometric (Section 2) and that scaling laws for alignment remain incomplete (Section 3). We now confront a question that cuts across both findings: do these principles hold beyond English?

The answer matters urgently. An estimated 78.5% of safety-relevant datasets are English-only, yet LLMs are deployed to users across every inhabited continent. If the geometric structure of alignment (the safety subspaces, refusal directions, and reward manifolds characterized in Section 2) is language-specific rather than universal, then the entire data-centric alignment agenda must be replicated independently for each language. If, conversely, these structures transfer cross-lingually, then the research agenda shifts to understanding the conditions, limits,

and failure modes of that transfer. As we shall see, the evidence points to a nuanced middle ground: cross-lingual transfer of safety representations is real and mechanistically grounded, but it is shallow, fragile, and systematically weakest for the languages whose speakers are most vulnerable.

4.1 Measuring the Gap

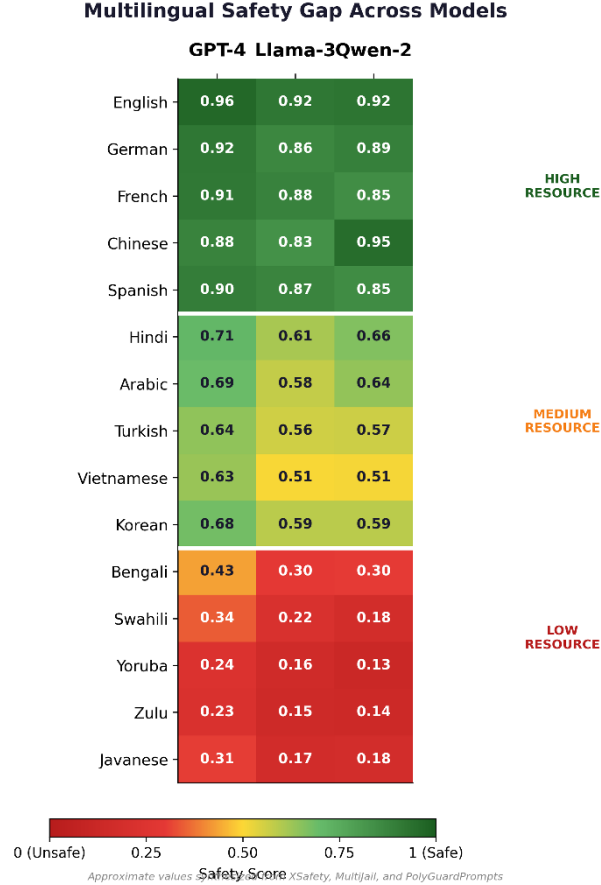


Figure 6: Multilingual safety gap across models. Approximate safety scores for 15 languages across three major LLMs. High-resource languages maintain scores above 0.85, while low-resource languages drop below 0.40. Values synthesised from XSafety, MultiJail, and PolyGuardPrompts.

A consistent finding across more than a dozen benchmarking studies is that all major LLMs produce significantly more unsafe responses when prompted in non-English languages. The first systematic evidence came from XSAFETY Wang et al. (2023), which evaluated 14 safety categories across 10 languages and 28,000 professionally translated instances. Bengali, Hindi, and Japanese (each constituting less than 0.2% of GPT-3’s training corpus) exhibited the highest unsafety rates. A simple intervention (“Please think in English and then generate the response in the original language”) reduced unsafety from 19.1% to 9.7%, foreshadowing the mechanistic findings we discuss below.

Subsequent benchmarks have expanded coverage along multiple dimensions. POLYGLOTOXICITYPROMPTS Jain et al. (2024a) introduced 425,000 prompts across 17 languages, demonstrating that toxicity increases as language resources decrease and, counterintuitively, as model size increases. RTP-LX de Wynter et al. (2024) contributed a human-transcreated toxic prompt corpus in 28 languages, revealing that LLM-based toxicity detectors have low agreement with human judges despite acceptable accuracy, calling into question the reliability of automated

multilingual safety evaluation. Most recently, POLYGUARD Kumar et al. (2025) released POLYGUARDPROMPTS, a 29,000-sample evaluation set for 17 languages created by combining naturally occurring multilingual human-LLM interactions from LMSYS-CHAT-1M and WILDCHAT with human-verified machine translations.

For reward models specifically, M-REWARDBENCH Gureja et al. (2024) established the first multilingual reward model evaluation spanning 23 languages, revealing a maximum 13% performance drop between English and non-English settings, with the gap widening for typologically distant languages. This finding connects directly to the geometric framework of Section 2: if reward models encode preferences as directions in representation space, then the fidelity of those directions degrades as the underlying language representations diverge from English.

Region-specific benchmarks have filled important gaps that translation-based approaches miss. SEA-SAFEGUARBENCH Tasawong et al. (2025) provides the first human-verified Southeast Asian safety benchmark covering 8 languages and 21,600 samples. The Arabic safety dataset of Ashraf et al. (2024) adapts 5,799 questions to Arab socio-cultural contexts. These efforts confirm that multilingual safety evaluation cannot be reduced to translated English benchmarks.

4.2 Cross-Lingual Transfer: Real but Shallow

The central data-centric question is whether safety signals learned from English data transfer to other languages, and if so, through what mechanisms. Recent mechanistic interpretability work provides surprisingly detailed answers that connect directly to the geometric view of alignment developed in Section 2.

Dual multilinguality of MLP layers. Li et al. (2024c) provided the first demonstration that preference tuning for toxicity mitigation generalizes cross-lingually in a zero-shot manner. Using DPO with *only English* training data, they reduced the probability of generating toxic continuations from 46.8% to 3.9% across 17 languages on mGPT-1.3B. The result generalizes to models of different sizes and pretraining compositions, including BLOOM (1.7B and 7.1B), LLaMA2-7B, LLaMA3-8B, and Aya-23-8B. The mechanistic explanation centers on what the authors term *dual multilinguality*: value vectors in MLP layers encode multilingual toxic concepts, and the key vectors that activate them respond consistently to toxic prompts regardless of input language. After DPO, the activations produced by these key vectors are suppressed; because the suppression operates on language-agnostic representations, detoxification propagates across languages.

In the language of Section 2, this finding implies that the “unsafe region” in representation space is not language-specific but rather a shared subspace that MLP value vectors project into across languages. DPO reshapes this subspace once, and the effect is felt multilingually.

Universality of the refusal direction. Wang et al. (2025b) investigated whether the refusal direction (the low-dimensional subspace mediating a model’s tendency to refuse harmful requests) is universal across languages. Using POLYREFUSE, a dataset spanning 14 linguistically diverse languages across 7 language families and writing systems, they demonstrated that refusal vectors extracted from English can bypass refusals in other languages with near-perfect effectiveness, without any fine-tuning. The geometric explanation is striking: refusal directions are approximately *parallel* across languages in the model’s embedding space. This parallelism explains both why English-derived safety interventions transfer and why cross-lingual jailbreaks succeed: if the refusal direction is shared, then ablating it in one language ablates it in all.

Critically, however, the authors found that while the refusal direction is shared, models often fail to adequately *separate* harmful from harmless prompts in non-English languages. The refusal infrastructure exists but is not reliably triggered, because the model’s ability to detect harm degrades for non-English inputs. This distinction between the *existence* of safety geometry and its *activation* is central to understanding why cross-lingual transfer is real but incomplete.

Cross-lingual shared safety neurons. The most fine-grained mechanistic account comes from Wang et al. (2026b), who identified a remarkably sparse ($<0.3\%$) yet critical subset of neurons, termed *cross-lingual shared safety neurons* (SS-Neurons), that jointly regulate safety behavior across languages. The investigation proceeds in two stages. First, *monolingual safety neurons* (MS-Neurons) are identified for each language through contrastive activation analysis. Masking MS-Neurons in Qwen3-8B increases the average attack success rate across all languages by 25.85 percentage points, while masking equivalent random neurons has negligible impact. Second, SS-Neurons are defined as the intersection of MS-Neurons between English and each non-English language: $SS_\ell = MS_\ell \cap MS_{\text{English}}$. This intersection is strikingly small, yet its deactivation causes a disproportionate safety collapse, confirming that these neurons serve as the critical bridge transferring safety from English.

A strong negative correlation exists between SS-Neuron count and attack success rate: languages with fewer shared safety neurons are more vulnerable. This provides a neuron-level explanation for the geometric observation that safety subspaces are less well-defined for low-resource languages.

Reward model transfer. Hong et al. (2024) found that English-trained reward models consistently outperform target-language reward models by 3–4% on average on MULTILINGUAL REWARD BENCH across Spanish, Italian, Korean, and Chinese, with advantages of 12% in Korean and 27% in Chinese on reasoning tasks. The explanation is that English reward model training best preserves the general representations of the multilingual base model, whereas training on other languages induces greater representation collapse. When Qwen2.5-7B-Instruct is fine-tuned with DPO using English reward model labels, win rates improve by up to 11.8 percentage points against GPT-4-Turbo across four non-English languages. English thus functions as a practical *lingua franca* for reward modeling, not merely a convenience but a demonstrably effective strategy.

4.3 The Low-Resource Cliff

Across every benchmark and every method, performance degrades most dramatically for low-resource languages (LRLs). The pattern is consistent and severe.

Deng et al. (2023) introduced MULTI JAIL, the first manually-created multilingual jailbreak dataset, and demonstrated a *threefold* increase in harmful generation rates for LRLs compared to high-resource languages (HRLs), with unsafe rates rising from approximately 11% to 55%. In intentional attack scenarios, ChatGPT exhibits an 80.92% unsafe rate and GPT-4 reaches 40.71%. Wang et al. (2025b) classify Yoruba as a “safety-misaligned language,” with jailbreak compliance rates of 82.9% on LLaMA3.1-8B-Instruct and 74.9% on Qwen2.5-7B. Zhao et al. (2025) demonstrate that the implicit reward gap (the log-likelihood difference between safe and unsafe responses) shrinks nearly to zero for Bengali ($RG=0.04$) and Swahili ($RG=0.05$) on LLaMA-3.1-8B-Instruct, compared to 1.58 for English. When the reward gap vanishes, the model has effectively lost its ability to distinguish safe from unsafe outputs in that language. Wang et al. (2026b) show that non-high-resource languages have drastically fewer shared safety neurons with English, providing the mechanistic substrate for this degradation.

This is not merely a scaling problem. It reflects a structural limitation: the representations for low-resource languages are insufficiently aligned with English to support effective safety transfer. In geometric terms, the safety subspace identified in Section 2 exists primarily in the region of representation space occupied by high-resource languages; low-resource language representations lie partially outside this region, leaving the safety geometry incomplete.

4.4 Code-Switching as Attack Vector

Code-switching, the mixing of two or more languages within a single utterance, has emerged as a particularly potent jailbreak vector that exploits the boundaries between language-specific representations. Yoo et al. (2024) introduced the Code-Switching Red-Teaming (CSRT) framework, which synthesizes intra-sentence code-switched adversarial prompts combining up to 10 languages. CSRT achieves 46.7% more successful attacks than standard English red-teaming and outperforms monolingual multilingual attacks as well. The dataset comprises 315 code-switching queries averaging 7.83 ± 2.19 languages per query.

Two findings are particularly notable. First, *intra-sentence* code-switching is more effective than inter-sentence mixing, suggesting that safety mechanisms operate at a granularity coarser than individual tokens. Second, incorporating more languages and lower-resource languages increases attack success rates. This is concerning because code-switching is a natural phenomenon in multilingual communities (an estimated 2–5% of generation outputs from further-trained LLaMA models involve code-switching), meaning that safety mechanisms must handle it gracefully regardless of adversarial intent.

From the geometric perspective, code-switching creates inputs that do not cleanly map to any single language’s region in representation space, potentially falling outside the safety subspaces defined by monolingual alignment data. The finding that fine-tuning attacks modifying only 20% of parameters can break safety across all languages simultaneously Poppi et al. (2024) further underscores the fragility of the shared safety substrate.

4.5 Cultural Safety Cannot Be Translated

A growing body of work demonstrates that translation, even high-quality human translation, is insufficient for multilingual alignment, because safety itself is culturally situated. Some harms are universal (e.g., instructions for synthesizing explosives); others are meaningful only within specific sociocultural frames.

Shukla et al. (2026) provided direct evidence that translated benchmarks miss culturally specific harms. Studying cross-lingual transfer of composite harms in Indic languages, they found attack success rates exceeding 45% for LLaMA-3-8B under adversarial syntax manipulation. The failures were most pronounced for harms rooted in caste dynamics, regional political tensions, and culturally specific social taboos, none of which are capturable by translating English benchmarks. Banerjee et al. (2026) synthesized findings across the literature to argue that safety guardrails weaken sharply on low-resource and code-mixed inputs characteristic of Global South language use, proposing an agenda centered on participatory workflows involving local communities. Ashraf et al. (2024) demonstrated the need for dual-perspective evaluation frameworks adapted to Arab socio-cultural contexts.

The implication for data-centric alignment is fundamental: the safety manifold is not merely multilingual but *multicultural*. A model that correctly refuses a harmful request in English may comply with the same request reframed in terms of a culture-specific harm category it was never trained on, not because its refusal mechanism failed, but because its harm *detection* was never calibrated for that cultural context.

4.6 Mitigation Strategies

Three families of approaches have emerged for closing the multilingual safety gap, each operating at a different level of the alignment pipeline.

4.6.1 Preference Optimization Across Languages

Standard preference methods (DPO, IPO, KTO) are primarily monolingual. Zhao et al. (2025) observed that the implicit reward gap is strongly correlated with safety performance across

languages and proposed *Multilingual reward gap Optimization* (MPO), which directly minimizes the discrepancy between the dominant language’s reward gap and that of target languages. Using only 100 translated preference pairs per language, MPO consistently outperforms DPO, IPO, KTO, CPO, ORPO, and SimPO on multilingual jailbreak benchmarks (MultiJail, AdvBench-X, and CSRT) across LLaMA-3.1-8B-Instruct, Gemma-2-9B-it, and Qwen2.5-7B-Instruct without degrading general multilingual utility. The key insight is that the reward gap of the dominant language serves as a more reliable supervision signal than noisy multilingual preference data.

Hong et al. (2024) demonstrated that cross-lingual transfer extends to the full RLHF pipeline, with English reward models providing effective supervision for multilingual alignment. Dang et al. (2024) showed that even imperfect multilingual preference data improves safety when combined with cross-lingual transfer.

4.6.2 Representation-Based Alignment

A second family operates on internal representations, exploiting the mechanistic findings above.

Multi-Lingual Consistency (MLC) loss. Bu et al. (2026) proposed a plug-and-play auxiliary loss that promotes representational consistency across languages during safety alignment. The method feeds semantically equivalent prompts in multiple languages and enforces *geometric collinearity* among their representations by manipulating singular values of the cross-lingual representation matrix. By maximizing the dominance of the largest singular value, MLC ensures that multilingual representations of the same query collapse onto a single semantic direction. This connects directly to the geometric framework: MLC explicitly shapes the safety subspace to be language-invariant. The loss integrates into existing SFT or DPO pipelines without modifying the underlying objective and requires no response-level supervision in target languages.

Sparse weight editing. Liang et al. (2026) proposed a training-free approach that formulates cross-lingual alignment as a constrained linear transformation. The method identifies safety neurons through contrastive activation analysis, then computes a closed-form transformation matrix that maps harmful representations of low-resource languages onto safe activation patterns of high-resource languages, subject to a null-space projection constraint that preserves general capabilities. The Jaccard similarity analysis confirms the mechanistic picture: high-resource languages exhibit consistently higher safety-neuron overlap (English-Chinese: 0.39, English-Vietnamese: 0.33), while low-resource languages are more isolated (English-Bengali: 0.11, English-Hebrew: 0.06).

SS-Neuron expansion. The strategy of Wang et al. (2026b) selectively fine-tunes only English MS-Neurons using parallel multilingual safety data, effectively widening the neuronal bridge. This surgical intervention updates less than 0.6% of total parameters while significantly reducing vulnerability in non-high-resource languages, outperforming methods that modify far more parameters.

4.6.3 Multilingual Safety Data Construction

While transfer-based methods reduce the need for multilingual data, they do not eliminate it, particularly for culturally specific harms.

PolyGuardMix. Kumar et al. (2025) released the largest multilingual safety training corpus to date: 1.91 million samples across 17 languages, combining 1.47 million machine translations of WildGuardMix with 0.43 million naturally occurring multilingual human-LLM interactions from LMSYS-CHAT-1M and WILDCHAT. The inclusion of in-the-wild data captures naturally occurring code-switching, dialectal variation, and culturally situated harms that translation alone cannot produce. The resulting POLYGUARD safety classifier outperforms existing state-of-the-art open-weight and commercial classifiers by 5.5%.

CultureGuard. Joshi et al. (2025) addressed the cultural dimension through a four-stage synthetic data generation pipeline: cultural data segregation, cultural data adaptation, machine

translation, and quality filtering via back-translation consistency. The resulting dataset comprises 386,661 samples across 9 languages. A critical design choice is that cultural adaptation is performed on the English source text *before* translation, focusing on ideational and social cultural elements. The fine-tuned model achieves a 30.2% improvement in multilingual scores over the baseline Llama-Nemotron-Safety-Guard-V2.

CREST. Bansal & Mishra (2025) demonstrated that a 0.5B-parameter safety guardrail supporting 100 languages can be trained using data from only 13 high-resource languages via cluster-based cross-lingual transfer, suggesting that principled language clustering can substantially reduce the data requirements for multilingual safety.

4.7 Missing Experiments and Open Questions

Despite rapid progress, several fundamental experiments remain unrun.

Scaling laws for multilingual safety transfer. No systematic study has characterized how cross-lingual safety transfer scales with model size, pretraining data volume, number of languages in the pretraining mix, or the quantity of safety-specific data. The correlational evidence (bilingual retrieval accuracy predicting DPO transfer Li et al. (2024c); SS-Neuron count predicting safety Wang et al. (2026b); reward gap correlating with attack success rate Zhao et al. (2025)) suggests that principled scaling relationships exist. Characterizing them would enable practitioners to predict, before deployment, which languages a given model will be safe in and which require additional intervention. This directly extends the scaling laws agenda of Section 3 to the multilingual setting.

Optimal allocation across the language portfolio. Given a fixed annotation budget, how should safety data be distributed across N languages? Current work suggests that even 100 translated preference pairs per language can substantially improve safety via MPO Zhao et al. (2025), and that 13 high-resource languages suffice to cover 100 via clustering Bansal & Mishra (2025). But the optimal allocation (whether resources should concentrate on languages with the largest safety gaps, distribute uniformly, or follow some criterion based on linguistic typology) remains an open optimization problem.

Corpus-level multilingual safety without English mediation. All current transfer methods use English as the anchor language. Whether direct transfer between non-English language pairs (e.g., Chinese safety data improving Arabic safety) is feasible, and whether this could reduce the single point of failure that English mediation represents, is untested. The SS-Neuron framework suggests this may be difficult, since the shared safety bridge is defined relative to English, but the MLC loss Bu et al. (2026) hints at a path toward language-agnostic safety representations that do not privilege any single language.

Mechanistic studies at frontier scale. The mechanistic picture (shared safety neurons, universal refusal directions, dual multilinguality of MLPs) has been established primarily on models up to 8–14 billion parameters. Whether these mechanisms scale to frontier models with hundreds of billions of parameters, persist across architectural families (mixture-of-experts, state-space models), and can be leveraged for *proactive* safety design rather than post-hoc analysis remains open.

In summary, the multilingual alignment gap is real, large, and consequential. Cross-lingual transfer of safety signals is more effective than early assessments suggested: the geometric structures underlying alignment are substantially shared across languages. But this transfer systematically fails for the languages whose speakers need it most: low-resource languages where the safety subspace is thin, the reward gap vanishes, and fewer than 0.3% of neurons bridge the

safety divide. Closing this gap demands not only better algorithms and more data, but a fundamental shift toward inclusive, community-grounded approaches to defining safety across the world’s languages, and scaling laws that explicitly model the multilingual dimension of alignment.

5 Principled Corpus Composition

The composition of pretraining data (which domains to include, in what proportions, at what quality thresholds, and in what order) is arguably the most consequential design decision in large language model development that still lacks a unified theoretical foundation. As established in Section 2, corpus-level interventions such as upsampling aligned AI discourse (reducing misalignment from 45% to 9%) and filtering harmful knowledge (producing models that resist adversarial fine-tuning for over 10,000 steps) yield alignment effects far more robust than post-hoc methods. Yet these results, like most data mixing decisions in practice, were reached through empirical search rather than principled optimization. This section surveys the rapidly maturing literature on principled corpus composition and argues that while the mathematical machinery for optimizing data mixtures now exists, it has been developed almost exclusively to optimize *capability* metrics. Extending these frameworks to jointly optimize for capability *and* alignment is the central open problem.

This connects directly to the data-centric alignment thesis advanced throughout this paper: if alignment properties are encoded during pretraining rather than grafted on afterward, then corpus composition is not merely an engineering detail but a *first-order alignment decision*.

5.1 Three Generations of Data Mixing

The field has progressed through three discernible generations, each representing a qualitative leap in the formality and effectiveness of composition decisions.

The heuristic era (pre-2023). Mixtures were chosen by intuition, precedent, or limited manual experimentation. The Pile (Gao et al., 2020) allocated 24% to web data, 9% to Wikipedia, and 4% to GitHub based on practitioner judgment, with no formal optimization. PaLM (Chowdhery et al., 2023) and GLaM (Du et al., 2022) tuned domain weights on downstream tasks, but this required training potentially thousands of models at non-trivial scale and risked overfitting to the evaluation suite.

Proxy-model optimization (2023–2024). DoReMi (Xie et al., 2023a), DSIR (Xie et al., 2023b), and RegMix (Liu et al., 2024b) established that small proxy models (often 100–1,000× smaller than the target model) can efficiently determine optimal mixtures. Concurrently, data mixing laws (Ye et al., 2024; Ge et al., 2024) discovered that performance follows predictable functional forms of mixture proportions, enabling optimization before committing to full-scale training. Quality filtering matured through FineWeb (Penedo et al., 2024), DCLM (Li et al., 2024b), and RefinedWeb (Penedo et al., 2023), establishing that well-filtered web data alone can outperform curated multi-source corpora.

Adaptive and multi-objective optimization (2024–present). Dynamic adjustment of mixture proportions during training via online methods (Albalak et al., 2023), reinforcement learning (Ma et al., 2025), and model merging (Wang et al., 2026a) addresses the non-stationarity of optimal data distributions. Quality-aware scaling laws (Subramanyam et al., 2025) add data quality as a formal variable. Topic-level mixing (Peng et al., 2025; Zhu et al., 2025) begins to address the limitations of source-based decomposition.

Across all three generations, a striking pattern persists: every optimization method targets language modeling perplexity or downstream task accuracy. None incorporates alignment metrics as objectives.

5.2 Data Mixing Laws

A decisive advance came with the discovery that model performance follows *predictable functional forms* with respect to mixture proportions, enabling analytical optimization before committing to expensive training runs.

Exponential mixing laws. Ye et al. (2024) discovered that validation loss on domain i can be accurately predicted as an exponential function of the mixture proportions:

$$L_i(\mathbf{r}) = c_i + k_i \exp\left(\sum_{j=1}^M t_{ij} r_j\right), \quad (8)$$

where c_i , k_i , and t_{ij} are fitted parameters, and $\mathbf{r} \in \Delta^{M-1}$ is the mixture distribution over M domains. By nesting this mixing law with classical scaling laws for model size and training steps, the authors constructed a three-stage pipeline that predicts large-model performance on unseen mixtures from small-scale experiments alone. Applied to a 1B-parameter model trained on 100B tokens of RedPajama, the optimized mixture matched the performance of a model trained for 48% more steps on the default mixture, establishing a foundational principle: *mixture optimization can substitute for additional compute*. These results have been validated primarily on models up to 1B parameters trained on 100B tokens; whether the functional forms and fitted parameters extrapolate to frontier-scale models (70B+) remains an open question.

Bivariate scaling laws. Ge et al. (2024) extended the framework by jointly modeling the effects of domain proportions *and* data quantity. Their BiMIX law takes a multiplicative form:

$$\mathcal{L}(s, r) = \left(\frac{\mathcal{A}}{s^\alpha} + \mathcal{C}\right) \frac{\mathcal{B}}{r^\beta}, \quad (9)$$

where s is training steps and r the domain proportion. This formulation achieves mean relative error below 0.2% with $R^2 > 0.97$ when generalizing to unseen mixtures. A key practical contribution is the use of entropy-based proxies (particularly conditional entropy) to estimate BiMIX coefficients *without any training runs*, enabling training-free mixture recommendations.

Mixture optimization as regression. Liu et al. (2024b) reframed the problem as regression: train 512 small proxy models (1M parameters, 1B tokens each) on diverse randomly sampled mixtures, fit a regression model predicting performance from mixture proportions, then simulate millions of candidates to identify the optimum. The estimated FLOPs for all 512 proxy models amounts to roughly 2% of training a single 1B model, yet the predicted best mixture ranks first among 64 candidate 1B models trained for 25B tokens. A counter-intuitive finding is that general web corpora (e.g., CommonCrawl, Pile-CC) exhibit stronger positive correlation with downstream performance than sources traditionally perceived as high-quality such as Wikipedia, underscoring the need for automated optimization over human judgment.

5.3 Domain Reweighting via Proxy Models

Parallel to mixing laws, a complementary line of work developed *algorithmic* approaches to domain reweighting that learn optimal weights through proxy model training rather than fitting explicit functional forms.

DoREMi (Xie et al., 2023a), DSIR (Xie et al., 2023b), and related methods use small proxy models to optimize domain weights without committing to full-scale training. DoReMi’s Group DRO framework minimizes worst-case excess loss across domains:

$$\min_{\theta} \max_{\alpha \in \Delta^k} \sum_{i=1}^k \alpha_i \cdot \frac{1}{\sum_{x \in D_i} |x|} \sum_{x \in D_i} \max\{\ell_{\theta}(x) - \ell_{\text{ref}}(x), 0\}, \quad (10)$$

where ℓ_{θ} and ℓ_{ref} are the token-level log-losses of the proxy and reference models. Together, these methods demonstrate that domain reweighting via proxy models can substitute for substantial additional compute, with DoReMi improving downstream accuracy and matching baselines significantly faster than standard mixtures.

5.4 Quality Filtering: The Compute Multiplier and Its Limits

If domain reweighting determines *how much* of each source to include, quality filtering determines *which documents* within each source survive to the training set.

The compute multiplier effect. FineWeb (Penedo et al., 2024) and DCLM (Li et al., 2024b) established that model-based quality filtering is a large compute multiplier: DCLM-Baseline reached 64% on MMLU at 7B parameters using $6.6\times$ less compute than Llama 3 8B, with model-based filtering as the single most impactful design choice. The practical principle is clear: *for a fixed compute budget, investing in better filtering yields larger returns than additional raw data or parameters.*

Quality-aware scaling laws. Subramanyam et al. (2025) formalized this insight by introducing a dimensionless quality parameter $Q \in (0, 1]$ into the Chinchilla framework:

$$L(N, D, Q) = \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta} Q^{\gamma}} + E, \quad (11)$$

where the *effective sample size* $D_{\text{eff}} = D \cdot Q^{\gamma}$ captures the information-theoretic impact of quality. Under signal-to-noise ratio arguments $\gamma \approx 1$; under symmetric label noise assumptions $\gamma \approx 2$. This quantifies the tradeoff between collecting more data and cleaning existing data, with direct implications for alignment where safety-relevant data is scarce and expensive to curate.

The quality–diversity tension. Nait Saada et al. (2025) issued an important warning: classifier-based quality filtering systematically favors data resembling the classifier’s supervision targets, sacrificing distributional diversity. Documents scoring highly on “educational value” tend to be formal, expository, and encyclopedic, a narrow slice of human language production.

This tension is particularly acute for alignment. Filtering out toxic content is necessary for safety, but models must still *understand* harmful requests to refuse them appropriately. Removing all adversarial or offensive text may leave models unable to recognize such content at inference time. QUADMIX (Liu et al., 2025) attempts to balance quality and diversity jointly, and D4 (Tirumala et al., 2023) shows that excessive deduplication reduces distributional coverage, but the tradeoff remains an open theoretical challenge. As established in Section 2, Safety Pretraining’s approach of rephrasing borderline content rather than removing it, and tagging remaining harmful content with metadata, represents one promising solution, preserving the model’s ability to recognize harmful patterns while removing the patterns themselves.

5.5 Dynamic Mixing and Curriculum Learning

The methods above determine mixture proportions *before* training and hold them fixed. A growing body of work challenges this static assumption.

Online data mixing. Albalak et al. (2023) showed that adjusting domain weights during training based on the model’s evolving loss landscape outperforms static mixtures from DoReMi or The Pile’s defaults. Ma et al. (2025) formalized this as a reinforcement learning problem where the “action” is the data mixture and the “reward” is validation improvement, naturally handling the exploration-exploitation tradeoff between known-useful and underrepresented domains.

Learning rate decay wastes late-training data. Luo et al. (2025) identified a critical interaction between curriculum ordering and optimization: standard learning rate decay schedules cause data presented late in training to receive negligibly small gradient updates, effectively wasting the most valuable data if scheduled last. This has direct implications for alignment: if safety-critical data is positioned at the end of training (common in multi-stage pipelines), its influence on model parameters may be minimal. No existing framework jointly optimizes mixture proportions and temporal scheduling for alignment objectives.

Topic over source. Peng et al. (2025) demonstrated that organizing data by *topic* rather than source leads to more effective mixing. For alignment, this shift is particularly significant: rather than asking “how much web data?” we can ask “how much ethical reasoning, safety discussion, and adversarial content should we include?”, a framing far more directly relevant to alignment outcomes. ToReMi (Zhu et al., 2025) implements this insight by applying DoReMi-style reweighting at the topic level.

5.6 Connecting Composition to Geometry

The connections drawn in this subsection are hypothesis-generating rather than evidence-presenting. No existing work has directly measured how corpus composition changes affect safety subspace geometry; the following analysis identifies testable predictions that the experimental agenda of Section 7 is designed to evaluate.

The corpus composition machinery described above optimizes for capability, but the geometric framework of Section 2 provides a natural bridge to alignment-aware optimization.

A central open question is: *how would DoReMi-optimized mixtures change the safety subspace geometry?* The alignment tax decomposition (Equation 7) shows that the irreducible tax τ_0 depends on the principal angles between safety and capability directions in representation space, angles that are themselves determined by pretraining data composition. If upsampling ethical reasoning content increases the principal angle between the safety direction and the capability subspace (reducing τ_0), then DoReMi’s Group DRO objective could be extended to minimize alignment tax alongside worst-case capability loss. Conversely, if aggressive capability-focused optimization inadvertently collapses the safety subspace (reducing the principal angle), then current DoReMi-optimized mixtures may be actively harmful from an alignment perspective, a testable prediction that no existing work has evaluated.

The data mixing laws of Equation 8 offer another entry point: if alignment metrics (safety scores, refusal rates, value consistency) follow similar exponential functional forms with respect to mixture proportions, then the same three-stage prediction pipeline could be extended to jointly optimize capability and safety. The per-category scaling differences identified by Tan et al. (2025), where explicit, implicit, and mixed harmful data each exhibit distinct scaling trajectories, suggest that alignment mixing laws will require finer-grained domain decomposition than capability mixing laws, but not fundamentally different mathematical machinery.

5.7 The Corpus as Attack Surface

The power of corpus composition cuts both ways. The bidirectional alignment effects demonstrated by Tice et al. (2026) mean that adversarial corpus manipulation is feasible: an adversary

who can influence even a small fraction of the pretraining corpus could shift safety subspace geometry in predictable ways. This threat is compounded because DoReMi-style reweighting may inadvertently *upsample* adversarial content that falls in domains where the proxy model has room to improve, and web-scale corpora are too large for human audit of well-crafted adversarial content. Mitigating corpus-level attacks requires provenance tracking and cryptographic authentication of training data sources, infrastructure discussed in Section 6.

5.8 The Central Gap: Alignment-Aware Optimization

The mathematical infrastructure for principled corpus composition is mature, but optimizes capability exclusively. The machinery *already exists*; it needs alignment-aware extensions:

- **Data mixing laws** (Equation 8) predict performance from mixture proportions, if “performance” is extended to alignment metrics (safety scores, refusal rates, value consistency).
- **DoReMi’s Group DRO** (Equation 10) optimizes worst-case loss across domains, if alignment-relevant content categories (ethical reasoning, adversarial examples, refusal templates) are treated as domains.
- **Quality-aware scaling laws** (Equation 11) model quality’s effect on training, if “quality” includes alignment quality alongside capability quality.
- **Influence functions** (Choe et al., 2024; Chang et al., 2024) trace behaviors to training data, if the behaviors traced include alignment failures.

The challenge is not developing new mathematical tools but extending existing ones with alignment-aware objectives, metrics, and content taxonomies.

Key unresolved tensions. Three fundamental tensions must be addressed by any alignment-aware composition framework. First, *capability versus safety*: maximizing downstream benchmarks may require distributions suboptimal for alignment, though the safety pretraining results of Section 2 suggest the tradeoff is less severe than assumed. Multi-objective optimization representing the Pareto frontier is needed. Second, *quality versus diversity*: aggressive filtering narrows model capabilities, yet alignment requires exposure to harmful content (appropriately contextualized) for recognition and refusal. Third, *static versus dynamic*: the interaction between learning rate decay and data ordering (Luo et al., 2025) makes temporal placement of alignment data a first-order variable that no framework jointly optimizes.

5.9 Missing Experiments

Five experimental programs that would close this gap are developed in full in Section 7: (1) extending mixing law functional forms (Ye et al., 2024; Ge et al., 2024) to predict alignment metrics as functions of mixture proportions; (2) multi-objective DoReMi that treats alignment-relevant content categories as additional domains in the Group DRO objective; (3) defining alignment-aware quality metrics that disentangle capability quality from alignment quality for the parameter Q in Equation 11; (4) determining the optimal temporal placement of safety-relevant data given the learning rate decay interaction (Luo et al., 2025); and (5) using scaled influence functions (Choe et al., 2024; Chang et al., 2024) to identify training documents responsible for alignment failures, creating a closed-loop corpus refinement process.

The tools for principled corpus composition exist. The evidence that corpus composition causally determines alignment is compelling. What remains is to connect these two bodies of work by extending the mature mathematical frameworks for data mixing with alignment-aware objectives, creating a principled foundation for what is currently the most consequential and least theorized decision in building aligned language models.

6 Verification and Governance

The preceding sections have argued that alignment quality is determined by corpus composition: scaling laws govern how much alignment data is needed (Section 3); geometric analysis reveals why post-hoc alignment is fragile (Section 2); multilingual evidence exposes unprotected languages (Section 4); and data mixing laws prescribe how to compose corpora for capability, though not yet for safety (Section 5). If all of this is correct, a single question becomes paramount: *how do we know the corpus is right?*

This section surveys the tools, standards, and regulatory frameworks needed to verify that a training corpus meets defined quality, safety, provenance, and legal requirements. The field’s investment in corpus verification lags far behind its investment in model evaluation and model training.

6.1 A Maturity Model for Corpus Auditing

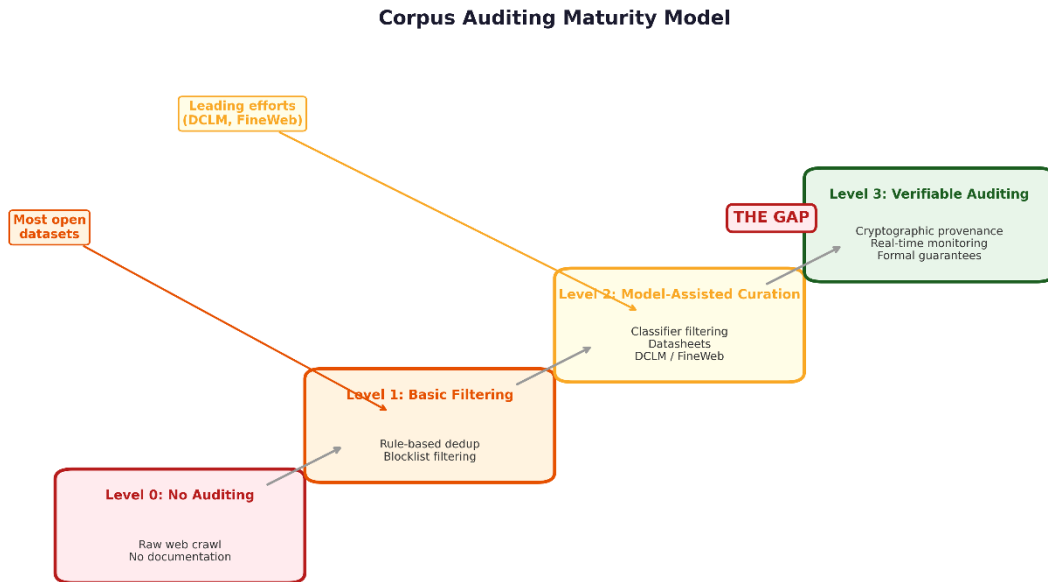


Figure 7: Corpus auditing maturity model. Four ascending levels from no auditing (Level 0) through model-assisted curation (Level 2) to verifiable auditing with cryptographic provenance (Level 3). Most open datasets sit at Level 1; leading efforts reach Level 2.

The literature suggests a natural progression of auditing sophistication that we formalise as a four-level maturity model:

Level 0: No Auditing. Raw web crawl with no filtering, no documentation, and no provenance tracking.

Level 1: Basic Filtering. Rule-based deduplication, heuristic toxicity filtering (blocklists, URL filters), and minimal metadata. Most open datasets, including early C4, sit here (Dodge et al., 2021).

Level 2: Model-Assisted Curation. Classifier-based quality and safety filtering, systematic documentation (datasheets or data cards), and some provenance tracking. Leading efforts currently operate here: DCLM (Li et al., 2024a) achieves a $6.6\times$ compute multiplier through model-based filtering; FineWeb (Penedo et al., 2024) documents ablations across

15 trillion tokens; and Safety Pretraining (Maini et al., 2025) reduces attack success rates from 38.8% to 8.4% through a four-step data-centric pipeline (detailed in Section 5).

Level 3: Verifiable Auditing. Cryptographically verifiable provenance, real-time monitoring, regulatory compliance, and formal safety guarantees. Almost no production system has reached this level. Verifiable Fine-Tuning (Akgul et al., 2025) represents an early prototype.

The gap between Level 2 and Level 3 is the central challenge. The remainder of this section surveys the key obstacles.

6.2 The Documentation Crisis

Despite foundational frameworks proposed years ago (Gebru et al., 2021; Mitchell et al., 2019), documentation adoption remains poor: license omission and error rates are high across thousands of datasets (Longpre et al., 2023), and even venues explicitly committed to dataset quality show systematic underreporting (Bhardwaj et al., 2024; Stalnaker et al., 2025; Yang et al., 2024; Rondina et al., 2025). Two developments point toward solutions: machine-readable formats (Roman et al., 2023; Jain et al., 2024b) that enable automated compliance checking, and LLM-assisted documentation generation (Liu et al., 2024a) that reduces marginal cost to near zero. Both are prerequisites for Level 3 auditing.

6.3 Corpus Security

If alignment is embedded through corpus composition, then adversarial manipulation of the corpus is a direct attack on alignment. Winter Soldier (Bouaziz et al., 2025) demonstrates that indirect data poisoning at pretraining time is feasible with less than 0.005% poisoned tokens. The attack makes a model learn arbitrary secret sequences *absent* from the training corpus, detectable with extremely high confidence ($p < 10^{-55}$) and without performance degradation on standard benchmarks. An adversary who can inject a small number of documents into a web crawl can plant backdoors that survive standard quality filtering.

Even without deliberate manipulation, existing datasets harbour risks. Jahanshahi & Mockus (2025) found that the Stack v2 code dataset contains code blobs vulnerable to 6,947 known CVEs, with license misidentification enabling non-permissively licensed code to enter datasets marked as open.

6.4 The Regulatory Gap

The EU AI Act mandates representative, bias-mitigated training data for high-risk systems, but translating these legal principles into concrete engineering criteria remains unsolved. Practitioner interviews reveal persistent structural challenges: fragmented pipelines, unclear responsibility boundaries, and tooling that prioritises performance over compliance (Wang & Irion, 2026). Existing tools—machine-readable documentation (Roman et al., 2023), verifiable training proofs (Akgul et al., 2025), provenance tracking (Hohensinner et al., 2026)—address components of this gap, but none has been deployed at scale.

6.5 Missing Infrastructure

Individual auditing components exist but lack integrative infrastructure. Four capabilities are missing:

1. **Unified corpus safety assessment:** an integrated pipeline combining documentation verification, quality scoring, safety classification, provenance tracking, and security analysis.

2. **Pre-training safety certificates:** verifiable attestations, analogous to pharmaceutical GMP certificates, that a corpus meets defined safety criteria.
3. **Real-time corpus monitoring:** web-scale pipelines continuously ingest new data whose properties change over time; batch-oriented auditing is insufficient.
4. **Adversarial robustness testing:** standardised red-team exercises that attempt to inject backdoors and characterise the corpus’s vulnerability surface.

The technical foundations for Level 3 exist; what remains is the institutional will to build on them.

7 A Concrete Experimental Agenda

The preceding sections have identified open questions at every level of the data-centric alignment stack. This section consolidates these gaps into a prioritised experimental roadmap.

For each experiment, we specify a hypothesis, method, expected outcome with falsification criteria, and compute estimate. The experiments are organised into three tiers: foundational (executable now), medium-term (requiring new infrastructure), and long-term (defining the frontier). This agenda is the paper’s primary original contribution: not new results, but a concrete blueprint for producing them.

A note on compute constraints. Several of the foundational experiments, particularly Experiment 1, may appear straightforward given that existing measurement tools are publicly available. We have not run them ourselves because even Tier 1 experiments require non-trivial compute: Hessian and Fisher information estimation on 7–8B-parameter models demands hundreds of GPU-hours per model pair, and the controlled comparisons we propose require multiple such pairs at different scales. Our contribution here is to identify the most informative experiments and specify them precisely enough that groups with appropriate compute budgets can execute them efficiently. We view this as complementary to, not a substitute for, the empirical work itself. All GPU-hour estimates assume A100-class hardware and efficient distributed training implementations; actual requirements may vary by a factor of 2–3 $\times$ depending on hardware generation, parallelism efficiency, and implementation details.

7.1 Tier 1: Foundational Experiments

These experiments can be conducted with existing open-weight models, public datasets, and established measurement tools. They require compute on the order of hundreds to low thousands of GPU-hours and could yield publishable results within 6–12 months.

7.1.1 Experiment 1: Safety Basin Geometry of Corpus-Aligned vs. RLHF-Only Models

Hypothesis. Models aligned through deliberate corpus composition (e.g., Safety Pretraining (Maini et al., 2025), Deep Ignorance (O’Brien & Casper, 2025)) occupy wider safety basins with lower curvature in alignment-sensitive subspaces than models aligned purely through post-hoc RLHF/DPO.

Method. Select matched model pairs: (a) a corpus-aligned model and (b) a post-hoc-aligned model of the same architecture and parameter count. Candidate pairs include SmoLLM2 variants from Sam et al. (2026) or Safety Pretraining models vs. their RLHF-only counterparts. For each model, compute:

- SVD of safety vs. utility gradients following [Zhang et al. \(2026\)](#): effective rank of the safety subspace, singular value decay rate, and cosine similarity between safety and utility gradient subspaces.
- Hessian curvature in alignment-sensitive directions following [Springer et al. \(2026\)](#): minimum eigenvalue of the Fisher information matrix restricted to the safety subspace.
- Safety basin width following [Yang et al. \(2025\)](#): maximum perturbation magnitude along the alignment direction d_{aligned} and orthogonal directions $d_{\perp}$ before safety degrades below a threshold.

Expected outcome. Corpus-aligned models should exhibit higher effective rank in their safety subspaces (safety distributed across more dimensions), lower curvature (wider basins), and wider safety basins under both aligned and orthogonal perturbations.

Falsification. If corpus-aligned and post-hoc-aligned models show statistically indistinguishable basin widths and curvature profiles, the geometric depth hypothesis would be refuted.

Compute estimate. $\sim 200\text{--}500$ GPU-hours per model pair (dominated by Hessian estimation); 3–4 pairs needed for a convincing comparison across scales.

7.1.2 Experiment 2: Head-to-Head Stability Under Adversarial Fine-Tuning

Hypothesis. Corpus-aligned models resist adversarial fine-tuning for substantially more gradient steps than post-hoc-aligned models, confirming the depth-robustness gradient identified in Section 2.

Method. Using the same matched model pairs as Experiment 1, apply controlled adversarial fine-tuning following [Zhang et al. \(2026\)](#) (Harm dataset, GSM8K, SQL code, up to 20 epochs) and the “10-shot benign fine-tuning” attack ([Xie et al., 2025b](#)). Measure ASR at each epoch (HarmBench, WildGuard) and the number of steps to reach $\text{ASR} > 0.50$. Repeat at 1B, 3B, 7B scales.

Expected outcome. Corpus-aligned models should maintain low ASR for substantially more steps than post-hoc-aligned models, consistent with the results of Section 2. Under benign fine-tuning, the gap should be smaller but still present.

Falsification. If corpus-aligned models degrade at a comparable rate to post-hoc-aligned models under adversarial fine-tuning, the claim that pretraining-time alignment is structurally deeper would be undermined.

Compute estimate. $\sim 100\text{--}300$ GPU-hours per model (fine-tuning is cheaper than Hessian estimation); scales linearly with number of model pairs and scales tested.

7.1.3 Experiment 3: Alignment Mixing Laws

Hypothesis. The exponential mixing laws discovered by [Ye et al. \(2024\)](#) for capability metrics extend to safety and alignment metrics: alignment benchmark scores can be predicted as a function of the proportion of safety-relevant content in the pretraining mixture.

Method. Using the RegMix framework (Liu et al., 2024b), train 200–500 small proxy models (100M–300M parameters, 1–5B tokens each) on mixtures that systematically vary the proportion of safety-relevant content. Define safety-relevant content using the Safety Pretraining taxonomy (Maini et al., 2025) and the “aligned AI discourse” category from Tice et al. (2026). For each proxy model, evaluate on both capability benchmarks (MMLU, GSM8K) and safety benchmarks (HarmBench, WildGuard, ToxiGen).

Fit the exponential mixing law (Equation 8) and bivariate scaling law (Equation 9) to the safety metrics. Test whether the fitted parameters extrapolate to larger models (1B, 7B).

Expected outcome. Safety metrics should follow exponential mixing laws with distinct parameters from capability metrics. The optimal safety proportion should be scale-dependent, with larger models achieving comparable safety at lower safety-data proportions, consistent with the sublinear scaling hypothesis from Section 3.

Falsification. If safety metrics do not fit any smooth functional form (e.g., exhibiting threshold effects or non-monotonic behaviour)—specifically, if the best-fitting model achieves $R^2 < 0.8$ or mean relative error exceeds 5%—the mixing law framework would need fundamental revision for alignment applications.

Compute estimate. ~500–2,000 GPU-hours total for the proxy model sweep (approximately 2% of training a single 1B model per proxy, following Liu et al. 2024b); validation runs at 1B and 7B add ~1,000–5,000 GPU-hours.

7.1.4 Experiment 4: Multilingual Safety Transfer as a Function of Scale and Language Distance

Hypothesis. Cross-lingual safety transfer follows a predictable function of three variables: model scale, pretraining data volume in the target language, and typological distance from English.

Method. Select a family of multilingual models at multiple scales (e.g., Qwen2.5 at 0.5B, 3B, 7B, 14B) and evaluate safety on 15–20 languages stratified by resource level and typological distance from English. Use existing multilingual safety benchmarks: XSafety (Wang et al., 2023), MultiJail (Deng et al., 2023), PolyGuardPrompts (Kumar et al., 2025), and M-RewardBench (Gureja et al., 2024). For each model–language pair, measure:

- Safety score (refusal rate on harmful prompts, toxicity rate on benign prompts).
- SS-Neuron count (shared safety neurons between the target language and English, following Wang et al. 2026b).
- Reward gap between English and the target language (Zhao et al., 2025).

Fit a regression model predicting safety score from $\log(\text{model size})$, $\log(\text{pretraining tokens in target language})$ and a typological distance metric.

Expected outcome. Safety transfer should degrade log-linearly with typological distance, improve log-linearly with model scale and pretraining data volume, and exhibit a critical threshold below which safety essentially fails.

Falsification. If safety transfer shows no systematic relationship with typological distance (e.g., if language-specific factors dominate over distance), the predictive framework would be insufficient and language-by-language characterisation would be necessary.

Compute estimate. $\sim 50\text{--}200$ GPU-hours (inference-only on existing models); the main cost is benchmark evaluation across many model–language pairs.

7.2 Tier 2: Medium-Term Experiments

These experiments require new datasets, custom training runs, or infrastructure that does not yet exist in standardised form. They are feasible within 1–2 years and require compute on the order of tens of thousands of GPU-hours.

7.2.1 Experiment 5: Multi-Objective DoReMi for Capability and Alignment

Hypothesis. DoReMi’s Group DRO framework (Xie et al., 2023a) can be extended to jointly optimise for capability and alignment by treating alignment-relevant content categories as additional domains, producing mixtures that Pareto-dominate single-objective optimisation on both metrics.

Method. Supplement standard source-based domains with alignment-relevant topic domains (ethical reasoning, social norms, safety demonstrations, adversarial content, aligned AI discourse), classified using the HarmFormer taxonomy (Mendu et al., 2025). Run the DoReMi proxy model pipeline twice: once optimising worst-case excess loss on capability domains only, and once on a joint capability-and-alignment objective. Train 1B and 7B target models on both mixtures.

Expected outcome. The joint-objective mixture should achieve meaningful safety improvements over the capability-only mixture with modest capability degradation, expanding the Pareto frontier.

Falsification. If the joint-objective mixture does not Pareto-dominate the capability-only mixture (i.e., if safety gains require proportional capability losses), it would suggest that the alignment tax at the corpus composition level is less compressible than hypothesised.

Compute estimate. $\sim 5,000\text{--}15,000$ GPU-hours (proxy model sweep plus 1B and 7B target training runs).

7.2.2 Experiment 6: Critical Depth Threshold for Irremovable Alignment

Hypothesis. There exists a critical depth, defined in terms of alignment intervention timing during pretraining and the dimensionality of the resulting safety subspace, beyond which alignment becomes qualitatively harder to remove through fine-tuning.

Method. Extend the intervention timing study of Sam et al. (2026) to more granular timing points (every 5% of pretraining) and larger model scales (1.7B, 3B, 7B). At each timing point and scale, measure: (a) the effective rank of the safety subspace via SVD, (b) the safety basin width under adversarial fine-tuning, and (c) the number of fine-tuning steps to reach $\text{ASR} > 0.50$. Plot these quantities as functions of intervention timing and model scale.

Expected outcome. We expect a transition region: models that receive safety interventions before a critical fraction of pretraining should exhibit qualitatively wider safety basins and higher-dimensional safety subspaces than models intervened upon later. The transition region may shift earlier with increasing model scale (larger models establish representational structure faster).

Falsification criteria.

1. If safety basin width increases *linearly* with earlier intervention timing rather than exhibiting a transition, the “critical depth” framing would be misleading: it would be a continuous gradient, not a phase transition.
2. If the transition point does not shift systematically with model scale, the hypothesised connection between representational maturation speed and alignment depth would be unsupported.
3. If adversarial fine-tuning steps to degradation show no dependence on intervention timing at all, the entire depth-robustness hypothesis would require revision.

Compute estimate. $\sim 20,000\text{--}50,000$ GPU-hours (multiple full pretraining runs at 1.7B–7B scale across timing points).

7.2.3 Experiment 7: Chinchilla for Alignment

Hypothesis. There exists a joint scaling law of the form $\text{Safety}(N, D_{\text{align}}, Q) = f(N, D_{\text{align}}, Q)$ that predicts safety performance as a function of model size N , alignment data volume D_{align} , and alignment data quality Q .

Method. Design a large-scale sweep varying:

- Model size: 100M, 300M, 1B, 3B, 7B parameters.
- Alignment data volume: 0.1%, 0.5%, 1%, 2%, 5%, 10% of pretraining tokens.
- Alignment data quality: low (heuristic filtering only), medium (classifier-based), high (Safety Pretraining four-step pipeline).

For each combination, evaluate safety (HarmBench, WildGuard) and capability (MMLU, GSM8K). Fit candidate functional forms: the exponential plateau (Equation 6), power laws, and the quality-aware scaling law (Equation 11).

Expected outcome. Safety should follow an exponential plateau in D_{align} whose saturation level increases with model size and data quality. The optimal alignment data proportion should decrease with model size (sublinear scaling), and the quality sensitivity should be larger for alignment than for capability.

Falsification. If safety metrics do not exhibit plateau behaviour (e.g., if they scale as a power law without saturation), the exponential plateau model would be inappropriate and the “alignment data budget” framing would need revision.

Compute estimate. $\sim 50,000\text{--}100,000$ GPU-hours (hundreds of training runs across the sweep grid; the largest individual runs are 7B models).

7.2.4 Experiment 8: Optimal Safety Curriculum Accounting for Learning Rate Decay

Hypothesis. The interaction between safety data placement and learning rate schedule has a large effect on alignment durability, and safety data placed during the learning rate’s active phase produces deeper, more robust alignment than safety data placed during the decay phase.

Method. Extend Sam et al. (2026) and incorporate the learning rate decay findings of Luo et al. (2025). Train 1.7B models with identical safety data volumes but five temporal placements: front-loaded (first 20%), uniform, back-loaded (last 20%), LR-matched (proportional to instantaneous learning rate), and cyclic (periodic bursts). Evaluate on safety (ASR under standard and adversarial conditions) and capability benchmarks.

Expected outcome. LR-matched placement should Pareto-dominate other schedules, achieving the deepest alignment for a given volume of safety data. Back-loaded placement should perform worst, confirming that data presented during learning rate decay has minimal impact on final model parameters.

Falsification. If uniform placement matches or exceeds LR-matched placement, the learning rate interaction hypothesis would be unsupported, and safety data scheduling would be less important than total volume.

Compute estimate. ~10,000–20,000 GPU-hours (5–10 full pretraining runs at 1.7B scale).

7.2.5 Experiment 9: Corpus-Level and Post-Hoc Alignment Interaction

Hypothesis. Corpus-level alignment and post-hoc alignment (RLHF/DPO) interact synergistically: models that receive both interventions are more robust than either intervention alone, and the marginal benefit of post-hoc alignment is smaller on corpus-aligned models than on conventionally pretrained models.

Method. Train four matched model variants at 1B and 7B scale: (a) conventional pretraining only, (b) conventional pretraining + RLHF, (c) safety pretraining only, (d) safety pretraining + RLHF. Evaluate all four on safety benchmarks (HarmBench, WildGuard), capability benchmarks (MMLU, GSM8K), and adversarial robustness (fine-tuning attack survival). Measure the interaction term: does the safety benefit of RLHF differ between (a→b) and (c→d)?

Expected outcome. The combination (d) should be the most robust, but the marginal benefit of RLHF should be smaller for corpus-aligned models (c→d) than for conventional models (a→b), consistent with the hypothesis that corpus-level alignment already occupies much of the safety subspace that RLHF would target.

Falsification. If RLHF provides equal marginal benefit regardless of pretraining strategy, corpus-level and post-hoc alignment would be independent mechanisms rather than substitutes, undermining the geometric interpretation.

Compute estimate. ~5,000–15,000 GPU-hours (four model variants at two scales, plus RLHF training and evaluation).

7.3 Tier 3: Long-Term Research Programmes

These are multi-year efforts that require community coordination, new infrastructure, or fundamental methodological advances.

7.3.1 Experiment 10: Unified Corpus Safety Assessment Framework

Hypothesis. A comprehensive corpus auditing system (combining documentation verification, quality scoring, toxicity classification, bias detection, provenance tracking, and security analysis) can reduce downstream safety failures by an order of magnitude compared to any single component in isolation.

Method. Build the “food safety inspection” system for training corpora: accept a candidate corpus and produce a structured safety certificate. Integrate existing tools: HarmFormer for safety classification (Mendu et al., 2025), DCLM-style quality filters (Li et al., 2024a), Open Datasheets for documentation verification (Roman et al., 2023), and Winter Soldier-style canary detection for security (Bouaziz et al., 2025). Validate by training matched model pairs on certified vs. uncertified corpora.

Expected outcome. Models trained on certified corpora should achieve low ASR without post-training alignment, matching or exceeding Safety Pretraining’s results.

Falsification. If certified corpora produce models no safer than uncertified ones (controlling for data volume), the additive value of integrated auditing over individual components would be questionable.

Compute estimate. Primarily engineering effort; validation training runs $\sim 10,000$ – $30,000$ GPU-hours.

7.3.2 Experiment 11: Verifiable Pretraining

Hypothesis. The Verifiable Fine-Tuning framework (Akgul et al., 2025) can be extended from fine-tuning to pretraining, producing cryptographic proofs that a pretrained model was obtained from a declared corpus under declared training policies.

Method. Develop amortised proof generation (per-checkpoint rather than per-step), explore probabilistic verification (random auditing of a subset of training steps with statistical guarantees), and design policy languages that express alignment-relevant constraints and compile them to verifiable circuits.

Expected outcome. A prototype system producing end-to-end verifiable certificates for models up to 1B parameters trained on up to 100B tokens, with manageable proof generation overhead.

Falsification. If proof generation overhead exceeds training cost by more than an order of magnitude even with amortisation, the approach may not be practical for pretraining-scale runs within current hardware constraints.

Compute estimate. Research and development dominated; prototype validation $\sim 10,000$ – $50,000$ GPU-hours.

7.3.3 Experiment 12: Culturally-Grounded Multilingual Safety Without English Mediation

Hypothesis. Multilingual safety alignment that incorporates natively-collected, culturally-grounded safety data produces qualitatively better outcomes for low-resource languages than translated English data, particularly for culturally-specific harms.

Method. Select 5–8 languages representing different resource levels and cultural contexts. For each, collect native safety data through participatory workshops (Banerjee et al., 2026), construct matched training sets (translated English, natively-collected, and combined), and fine-tune multilingual models on each. Evaluate on both universal and culturally-specific safety benchmarks.

Expected outcome. Native data should substantially outperform translated data on culturally-specific harms while performing comparably on universal harms. The combination should Pareto-dominate either alone, with larger effect sizes for languages more culturally distant from English.

Falsification. If translated data performs comparably to native data even on culturally-specific benchmarks, the case for expensive native data collection would be weakened.

Compute estimate. $\sim 5,000$ – $10,000$ GPU-hours for model training; the primary cost is the social infrastructure for community-based data collection.

7.3.4 Experiment 13: Alignment-Aware Quality Metrics

Hypothesis. The quality parameter Q in quality-aware scaling laws (Subramanyam et al., 2025) can be decomposed into a capability component Q_C and an alignment component Q_A , where these components are partially independent.

Method. Develop a multidimensional quality scoring system assigning each document both a capability score Q_C and an alignment score Q_A . Train models on corpora filtered to different (Q_C, Q_A) thresholds and measure both capability and safety outcomes. Fit the quality-aware scaling law with separate quality exponents.

Expected outcome. The two quality dimensions should be weakly correlated, confirming that high capability quality does not imply high alignment quality. The alignment quality exponent should be larger than the capability exponent.

Falsification. If Q_C and Q_A are strongly correlated ($r > 0.7$), the decomposition would be unnecessary; a single quality metric would suffice for both objectives.

Compute estimate. $\sim 10,000$ – $30,000$ GPU-hours (training runs across the (Q_C, Q_A) grid at multiple model scales).

7.4 Summary

Table 1 summarises the experimental agenda. Tier 1 experiments would ground the data-centric alignment thesis in quantitative evidence; Tier 2 experiments would build infrastructure for alignment-aware corpus design; and Tier 3 experiments would establish the long-term programmes needed for mature engineering practice.

8 Conclusion

When the whitepaper that preceded this work (Müller, 2026) was written, the proposition that alignment might be fundamentally a data problem was a minority position. The dominant paradigm treated pretraining as a capability-building phase and alignment as a behavioural correction applied afterward. The evidence assembled in this paper suggests that this paradigm is

#	Experiment	Key Metric	Section	Tier	GPU-hrs
1	Safety basin comparison	Basin width, curvature	2	1	~1k
2	Adversarial FT stability	Steps to ASR > 0.50	2	1	~1k
3	Alignment mixing laws	Mixing law fit quality	5	1	~2k
4	Multilingual safety scaling	Safety vs. distance	4	1	~0.2k
5	Multi-objective DoReMi	Pareto frontier expansion	5	2	~10k
6	Critical depth threshold	Phase transition point	2	2	~35k
7	Chinchilla for alignment	Scaling law fit quality	3	2	~75k
8	Optimal safety curriculum	Alignment durability	5	2	~15k
9	Corpus/post-hoc interaction	Interaction term	2	2	~10k
10	Unified safety framework	ASR of certified models	6	3	~20k
11	Verifiable pretraining	Proof overhead	6	3	~30k
12	Culturally-grounded safety	Cultural harm coverage	4	3	~8k
13	Alignment quality metrics	Q_A vs. Q_C correlation	5	3	~20k

Table 1: Summary of the proposed experimental agenda, organised by priority tier. The “Section” column indicates which section of this paper motivates each experiment. GPU-hour estimates are order-of-magnitude and assume A100-class hardware.

incomplete, and that the field is moving from “alignment might be a data problem” to “alignment is *substantially* a data problem.”

The case rests on convergent evidence from multiple independent research threads. Post-hoc alignment modifies a low-rank, high-curvature subspace that can be ablated with a single direction (Arditi et al., 2024), disrupted by fine-tuning on 10 benign examples (Xie et al., 2025b), and degraded at a quartic lower bound by curvature coupling (Springer et al., 2026). Corpus-level alignment, by contrast, exhibits the robustness documented in Section 2—plausibly because it shapes feature geometry during maximum representational plasticity, though the precise mechanism remains to be established by the experiments of Section 7.

The mathematical infrastructure for principled corpus composition exists in prototype—mixing laws (Ye et al., 2024; Ge et al., 2024), domain reweighting (Xie et al., 2023a), quality-aware scaling (Subramanyam et al., 2025), and verifiable training (Akgul et al., 2025)—but none of these tools has been extended to alignment objectives.

What does not yet exist is the framework that connects these tools into an alignment-aware corpus design system. No data mixing law predicts safety metrics. No domain reweighting algorithm jointly optimises for capability and alignment. No scaling law relates model size to the optimal volume and quality of alignment data. No verification system certifies that a pretraining corpus meets defined safety standards. These are the gaps this paper has identified, and the experimental agenda of Section 7 charts a concrete path toward closing them.

Three findings deserve emphasis.

First, *the alignment tax may be largely compressible, but it is not zero*. The alignment tax decomposes into an irreducible component τ_0 (genuine safety–capability conflicts) and a packing residual that vanishes as $O(m'/d)$ with model dimension (Young, 2026). Current results suggest τ_0 is small, but it is data-dependent and may be substantial for more demanding safety specifications or smaller models. Better corpus design can dramatically reduce the tax, but should not be expected to eliminate it entirely.

Second, *multilingual safety requires native data, not translation*. Cross-lingual transfer of safety representations is real but shallow: it correlates with typological distance, fails for low-resource languages, and misses culturally-specific harms (Wang et al., 2025b, 2026b; Shukla et al., 2026). Addressing this requires community-grounded data collection in the languages where safety gaps are largest, a social and institutional challenge as much as a technical one.

Third, *the verification gap is among the most pressing constraints*. The field has sophisticated benchmarks for model outputs and increasingly powerful tools for corpus curation, but no

systematic way to verify that a corpus meets safety standards before training begins. Until the infrastructure for verifiable auditing exists, claims about corpus-level alignment remain difficult to independently validate.

Scope and limitations. This paper is an agenda paper with theoretical grounding, not a pure survey. We have synthesised evidence from over 370 papers to argue for a specific research direction, but the experimental agenda we propose remains unexecuted. The strength of our claims is proportional to the evidence we cite, not to the experiments we envision. Moreover, the corpus-centric approach we advocate faces practical constraints that we have not fully addressed: the cost of pretraining-scale experiments limits who can conduct this research; access to pretraining data is restricted for most of the field; and the alignment quality metrics we call for (Experiment 13) do not yet exist. A further methodological limitation is that roughly half of the empirical claims in this paper cite 2025–2026 preprints that have not yet undergone formal peer review. While we have verified claims against source documents where possible, the evidentiary foundation will strengthen or weaken as these results are replicated and reviewed. We present this agenda in the hope that it will focus community effort on the experiments most likely to advance the science of alignment, while acknowledging that the path from agenda to validated theory is long and uncertain.

The community has invested enormous resources in making models safer after training. The evidence reviewed here suggests that comparable investment in making the data right before training begins would yield substantial returns in robustness, efficiency, and verifiability. Whether those returns are sufficient to justify the shift is, ultimately, an empirical question, one that the experiments proposed here are designed to answer.

References

- Ma, J., Dang, C., & Liao, M. (2025). Actor-Critic Based Online Data Mixing For Language Model Pre-Training. *arXiv:2505.23878*.
- Stalnaker, T., Wintersgill, N., Chaparro, O., Heymann, L. A., Di Penta, M., German, D. M., & Poshyvanyk, D. (2025). The ML Supply Chain in the Era of Software 2.0: Lessons Learned from Hugging Face. *arXiv:2502.04484*.
- Akgul, H., Borg, D., Berisha, A., Rahimova, A., Novak, A., & Petrov, M. (2025). Verifiable Fine-Tuning for LLMs: Zero-Knowledge Training Proofs Bound to Data Provenance and Policy. *arXiv:2510.16830*.
- Albalak, A., Pan, L., Raffel, C., & Wang, W. Y. (2023). Efficient Online Data Mixing For Language Model Pre-Training. *Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models at NeurIPS 2023*. *arXiv:2312.02406*.
- Tice, C., Radmard, P., Ratnam, S., Kim, A., Africa, D. D., & O’Brien, K. (2026). Alignment Pretraining: AI Discourse Causes Self-Fulfilling (Mis)alignment. *arXiv:2601.10160*.
- Roy, A., Patel, J., Chadha, A., Jain, V., & Das, A. (2025). AlignMerge: Alignment-Preserving Large Language Model Merging via Fisher-Guided Geometric Constraints. *arXiv:2512.16245*.
- Alssum, L., Itani, H., Hammoud, H. A. A. K., Torr, P., Bibi, A., & Ghanem, B. (2025). Unforgotten Safety: Preserving Safety Alignment of Large Language Models with Continual Learning. *arXiv:2512.10150*.
- Arditi, A., Obeso, O., Syed, A., Paleka, D., Rimsky, N., Gurnee, W., & Nanda, N. (2024). Refusal in Language Models Is Mediated by a Single Direction. *arXiv:2406.11717*.

- Ashraf, Y., Wang, Y., Gu, B., Nakov, P., & Baldwin, T. (2024). Arabic Dataset for LLM Safeguard Evaluation. *arXiv:2410.17040*.
- Liu, J., Li, W., Jin, Z., & Diab, M. (2024). Automatic Generation of Model and Data Cards: A Step Towards Responsible AI. *arXiv:2405.06258*.
- Banerjee, S., Hazra, R., & Mukherjee, A. (2026). Bridging the Multilingual Safety Divide: Efficient, Culturally-Aware Alignment for Global South Languages. *AAAI 2026*. *arXiv:2602.13867*.
- Bansal, L. & Mishra, N. (2025). CREST: Universal Safety Guardrails Through Cluster-Guided Cross-Lingual Transfer. *arXiv:2512.02711*.
- Barez, F., Fu, T., Prabhu, A., Casper, S., Sanyal, A., Bibi, A., O’Gara, A., Kirk, R., Bucknall, B., Fist, T., Ong, L., Torr, P., Lam, K.-Y., Trager, R., Krueger, D., Mindermann, S., Hernandez-Orallo, J., Geva, M., & Gal, Y. (2025). Open Problems in Machine Unlearning for AI Safety. *arXiv:2501.04952*.
- Bereska, L. & Gavves, E. (2024). Mechanistic Interpretability for AI Safety — A Review. *arXiv:2404.14082*.
- Bhardwaj, E., Gujral, H., Wu, S., Zogheib, C., Maharaj, T., & Becker, C. (2024). The State of Data Curation at NeurIPS: An Assessment of Dataset Development Practices in the Datasets and Benchmarks Track. *NeurIPS 2024 Datasets and Benchmarks Track*. *arXiv:2410.22473*.
- Bouaziz, W., Videau, M., Usunier, N., & El Mhamdi, E. M. (2025). Winter Soldier: Backdooring Language Models at Pre-Training with Indirect Data Poisoning. *arXiv:2506.14913*.
- Bu, Y., Liu, X., Ren, Z., Yang, Y., & Dai, J. (2026). Align Once, Benefit Multilingually: Enforcing Multilingual Consistency for LLM Safety Alignment. *ICLR 2026*. *arXiv:2602.16660*.
- Chang, T. A., Rajagopal, D., Bolukbasi, T., Dixon, L., & Tenney, I. (2024). Scalable Influence and Fact Tracing for Large Language Model Pretraining. *arXiv:2410.17413*.
- Choe, S. K., Ahn, H., Bae, J., Zhao, K., Kang, M., Chung, Y., Pratapa, A., Neiswanger, W., Strubell, E., Mitamura, T., Schneider, J., Hovy, E., Grosse, R., & Xing, E. (2024). What is Your Data Worth to GPT? LLM-Scale Data Valuation with Influence Functions. *arXiv:2405.13954*.
- Chen, J., Wang, X., Yao, Z., Bai, Y., Hou, L., & Li, J. (2024). Finding Safety Neurons in Large Language Models. *arXiv:2406.14144*.
- Chen, M., Chen, Y., Sun, W., & Zhang, X. (2025a). Avoiding $\exp(R_{\max})$ Scaling in RLHF through Preference-based Exploration. *arXiv:2502.00666*.
- Abu Shairah, H., Hammoud, H. A. A. K., Ghanem, B., & Turkiyyah, G. (2025). An Embarrassingly Simple Defense Against LLM Abliteration Attacks. *arXiv:2505.19056*.
- Subramanyam, A., Chen, Y., & Grossman, R. L. (2025). Scaling Laws Revisited: Modeling the Role of Data Quality in Language Model Pretraining. *arXiv:2510.03313*.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2023). PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, 24(240), 1–113. *arXiv:2204.02311*.
- Rondina, M., Vetrò, A., & De Martin, J. C. (2025). Completeness of Datasets Documentation on ML/AI Repositories: an Empirical Investigation. *Progress in Artificial Intelligence. EPIA 2023. Lecture Notes in Computer Science, vol 14115. Springer*. *arXiv:2503.13463*.

- Joshi, R., Paul, R., Singla, K., Kamath, A., Evans, M., Luna, K., Ghosh, S., Vaidya, U., Long, E., Chauhan, S. S., & Wartikar, N. (2025). CultureGuard: Towards Culturally-Aware Dataset and Guard Model for Multilingual Safety Applications. *arXiv:2508.01710*.
- Dang, J., Ahmadian, A., Marchisio, K., Kreutzer, J., Üstün, A., & Hooker, S. (2024). RLHF Can Speak Many Languages: Unlocking Multilingual Preference Optimization for LLMs. *arXiv:2407.02552*.
- Das, A., Jain, V., & Chadha, A. (2025). Tracing the Drift: Attributing Alignment Failures to Training-Time Belief Sources in LLMs. *arXiv:2508.02063*.
- Nait Saada, T., Bethune, L., Klein, M., Grangier, D., Cuturi, M., & Ablin, P. (2025). The Data-Quality Illusion: Rethinking Classifier-Based Quality Filtering for LLM Pretraining. *arXiv:2510.00866*.
- Jiao, C., Pan, Y., Xiao, E., Sheng, D., Jain, N., Zhao, H., Dasgupta, I., Ma, J. W., & Xiong, C. (2025). DATE-LM: Benchmarking Data Attribution Evaluation for Large Language Models. *arXiv:2507.09424*.
- Deng, Y., Zhang, W., Pan, S. J., & Bing, L. (2023). Multilingual Jailbreak Challenges in Large Language Models. *arXiv:2310.06474*.
- Djuhera, A., Kadhe, S. R., Ahmed, F., Zawad, S., & Boche, H. (2025). SafeMERGE: Preserving Safety Alignment in Fine-Tuned Large Language Models via Selective Layer-Wise Model Merging. *arXiv:2503.17239*.
- Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., & Gardner, M. (2021). Documenting the English Colossal Clean Crawled Corpus. *EMNLP 2021*. *arXiv:2104.08758*.
- Du, N., Huang, Y., Dai, A. M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A. W., Firat, O., Zoph, B., Fedus, L., Bosma, M., Zhou, Z., Wang, T., Wang, Y. E., Webster, K., Pellat, M., Robinson, K., et al. (2022). GLaM: Efficient Scaling of Language Models with Mixture-of-Experts. *ICML 2022*. *arXiv:2112.06905*.
- Fadli, S. (2025). Entropy-Based Measurement of Value Drift and Alignment Work in Large Language Models. *arXiv:2512.03047*.
- Fan, S., Pagliardini, M., & Jaggi, M. (2023). DoGE: Domain Reweighting with Generalization Estimation. *arXiv:2310.15393*.
- Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., & Leahy, C. (2020). The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv:2101.00027*.
- Gao, L., Schulman, J., & Hilton, J. (2022). Scaling Laws for Reward Model Overoptimization. *arXiv:2210.10760*.
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for Datasets. *Communications of the ACM*, 64(12), 86–92. *arXiv:1803.09010*.
- Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022). Training Compute-Optimal Large Language Models. *arXiv:2203.15556*.
- Hohensinner, R., Mutlu, B., Mendoza Estrada, I. G., Vukovic, M., Kopeinik, S., & Kern, R. (2026). Tracing the Data Trail: A Survey of Data Provenance, Transparency and Traceability in LLMs. *arXiv:2601.14311*.

- Cintron Roman, A., Vaughan, J. W., See, V., Ballard, S., Schifano, N., Torres, J., Robinson, C., & Lavista Ferres, J. M. (2023). Open Datasheets: Machine-readable Documentation for Open Datasets and Responsible AI Assessments. *arXiv:2312.06153*.
- Hong, J., Lee, N., Martínez-Castaño, R., Rodríguez, C., & Thorne, J. (2024). Cross-lingual Transfer of Reward Models in Multilingual Alignment. *arXiv:2410.18027*.
- Hossain, S., Tseng, T., Pandey, P. S., Vajpayee, S., Kowal, M., Nonta, N., Simko, S., Casper, S., Jin, Z., Pelrine, K., & Rambhatla, S. (2026). TamperBench: Systematically Stress-Testing LLM Safety Under Fine-Tuning and Tampering. *arXiv:2602.06911*.
- Hou, Z., Du, P., Niu, Y., Du, Z., Zeng, A., Liu, X., Huang, M., Wang, H., Tang, J., & Dong, Y. (2024). Does RLHF Scale? Exploring the Impacts from Data, Model, and Method. *arXiv:2412.06000*.
- Hsiung, L., Pang, T., Tang, Y.-C., Song, L., Ho, T.-Y., & Chen, P.-Y. (2025). Why LLM Safety Guardrails Collapse After Fine-tuning: A Similarity Analysis Between Alignment and Fine-tuning Datasets. *arXiv:2506.05346*.
- Huang, T., Hu, S., & Liu, L. (2024). Vaccine: Perturbation-aware Alignment for Large Language Model. *arXiv:2402.01109*.
- Huang, T., Hu, S., Ilhan, F., Tekin, S. F., Yahn, Z., Xu, Y., & Liu, L. (2025). Safety Tax: Safety Alignment Makes Your Large Reasoning Models Less Reasonable. *arXiv:2503.00555*.
- Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. *arXiv:2401.05566*.
- Isik, B., Ponomareva, N., Hazimeh, H., Paparas, D., Vassilvitskii, S., & Koyejo, S. (2024). Scaling Laws for Downstream Task Performance of Large Language Models. *arXiv:2402.04177*.
- Jahanshahi, M. & Mockus, A. (2025). Cracks in The Stack: Hidden Vulnerabilities and Licensing Risks in LLM Pre-Training Datasets. *arXiv:2501.02628*.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling Laws for Neural Language Models. *arXiv:2001.08361*.
- Khera, A., Ghosh, R., & Dutta, D. (2024). Efficient Alignment of Large Language Models via Data Sampling. *4th NeurIPS Workshop on Efficient Natural Language and Speech Processing (ENLSP-IV 2024)*.
- Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K., Bishop, C., Hall, E., Carbune, V., Rastogi, A., & Prakash, S. (2023). RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. *arXiv:2309.00267*.
- Li, J., Fang, A., Smyrnis, G., et al. (2024a). DataComp-LM: In Search of the Next Generation of Training Sets for Language Models. *arXiv:2406.11794*.
- Li, J., Fang, A., Smyrnis, G., et al. (2024b). DataComp-LM: In Search of the Next Generation of Training Sets for Language Models. *arXiv:2406.11794*.
- Li, X., Yong, Z.-X., & Bach, S. H. (2024). Preference Tuning For Toxicity Mitigation Generalizes Across Languages. *arXiv:2406.16235*.
- Li, J. & Kim, J.-E. (2024). Superficial Safety Alignment Hypothesis. *arXiv:2410.10862*.

- Li, K., Chen, Y., Viégas, F., & Wattenberg, M. (2025). When Bad Data Leads to Good Models. *arXiv:2505.04741*.
- Wollschläger, T., Elstner, J., Geisler, S., Cohen-Addad, V., Günnemann, S., & Gasteiger, J. (2025). The Geometry of Refusal in Large Language Models: Concept Cones and Representational Independence. *arXiv:2502.17420*.
- Ge, C., Ma, Z., Chen, D., Li, Y., & Ding, B. (2024). Data Mixing Made Efficient: A Bivariate Scaling Law for Language Model Pretraining. *arXiv:2405.14908*.
- Liang, J., Wang, Z., & Wang, H. (2026). Multilingual Safety Alignment Via Sparse Weight Editing. *arXiv:2602.22554*.
- Liu, Q., Zheng, X., Muennighoff, N., Zeng, G., Dou, L., Pang, T., Jiang, J., & Lin, M. (2024). RegMix: Data Mixture as Regression for Language Model Pre-training. *arXiv:2407.01492*.
- Longpre, S., Mahari, R., Chen, A., et al. (2023). The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. *arXiv:2310.16787*.
- Luo, K., Sun, Z., Wen, H., Shi, X., Cui, J., Dang, C., Lyu, K., & Chen, W. (2025). How Learning Rate Decay Wastes Your Best Data in Curriculum-Based LLM Pretraining. *arXiv:2511.18903*.
- Lucki, J., Wei, B., Huang, Y., Henderson, P., Tramèr, F., & Rando, J. (2024). An Adversarial Perspective on Machine Unlearning for AI Safety. *arXiv:2409.18025*.
- Maini, P., Goyal, S., Sam, D., Robey, A., Savani, Y., Jiang, Y., Zou, A., Fredrikson, M., Lipton, Z. C., & Kolter, J. Z. (2025). Safety Pretraining: Toward the Next Generation of Safe AI. *arXiv:2504.16980*.
- McKenzie, I. R., Lyzhov, A., Pieler, M., Parrish, A., Mueller, A., Prabhu, A., McLean, E., et al. (2023). Inverse Scaling: When Bigger Isn't Better. *arXiv:2306.09479*.
- Mendu, S. K., Yenala, H., Gulati, A., Kumar, S., & Agrawal, P. (2025). Towards Safer Pretraining: Analyzing and Filtering Harmful Content in Webscale Datasets for Responsible LLMs. *IJCAI 2025*. *arXiv:2505.02009*.
- Wang, J., Tian, C., Chen, K., Liu, Z., Mao, J., Zhao, W. X., Zhang, Z., & Zhou, J. (2026). MergeMix: Optimizing Mid-Training Data Mixtures via Learnable Model Merging. *arXiv:2601.17858*.
- Müller, O. (2026). Data-Centric Alignment of Large Language Models: A Survey of Corpus-Level Safety, Filtering, and Synthetic Data Approaches. *arXiv:26XX.XXXXX*.
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. *Proceedings of FAT* 2019*. *arXiv:1810.03993*.
- Mou, Y., Zhou, X., Luo, Y., Zhang, S., & Ye, W. (2025a). Decoupling Safety into Orthogonal Subspace: Cost-Efficient and Performance-Preserving Alignment for Large Language Models. *arXiv:2510.09004*.
- Liu, F., Zhou, W., Liu, B., Yu, Z., Zhang, Y., Lin, H., Yu, Y., Zhou, X., Wang, T., & Cao, Y. (2025). QuaDMix: Quality-Diversity Balanced Data Selection for Efficient LLM Pretraining. *arXiv:2504.16511*.
- Muennighoff, N., Rush, A. M., Barak, B., Le Scao, T., Piktus, A., Tazi, N., Pyysalo, S., Wolf, T., & Raffel, C. (2023). Scaling Data-Constrained Language Models. *arXiv:2305.16264*.

- O’Brien, K., Casper, S., Anthony, Q., Korbak, T., Kirk, R., Davies, X., Mishra, I., Irving, G., Gal, Y., & Biderman, S. (2025). Deep Ignorance: Filtering Pretraining Data Builds Tamper-Resistant Safeguards into Open-Weight LLMs. *ICLR 2026*. arXiv:2508.06601.
- Pan, W., Liu, Z., Chen, Q., Zhou, X., Yu, H., & Jia, X. (2025). The Hidden Dimensions of LLM Alignment: A Multi-Dimensional Safety Analysis. *arXiv:2502.09674*.
- Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., & Launay, J. (2023). The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. *arXiv:2306.01116*.
- Penedo, G., Kydlíček, H., Ben allal, L., Lozhkov, A., Mitchell, M., Raffel, C., Von Werra, L., & Wolf, T. (2024). The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. *NeurIPS 2024 (Datasets and Benchmarks Track)*. arXiv:2406.17557.
- Jain, D., Kumar, P., Gehman, S., Zhou, X., Hartvigsen, T., & Sap, M. (2024). PolygloToxicityPrompts: Multilingual Evaluation of Neural Toxic Degeneration in Large Language Models. *arXiv:2405.09373*.
- Kumar, P., Jain, D., Yerukola, A., Jiang, L., Beniwal, H., Hartvigsen, T., & Sap, M. (2025). PolyGuard: A Multilingual Safety Moderation Tool for 17 Languages. *arXiv:2504.04377*.
- Poppi, S., Yong, Z.-X., He, Y., Chern, B., Zhao, H., Yang, A., & Chi, J. (2024). Towards Understanding the Fragility of Multilingual LLMs against Fine-Tuning Attacks. *arXiv:2410.18210*.
- Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., & Henderson, P. (2023). Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! *arXiv:2310.03693*.
- Qin, Z., Dong, Q., Zhang, X., Dong, L., Huang, X., Yang, Z., Khademi, M., Zhang, D., Awadalla, H. H., Chen, W., Cheng, M., & Wei, F. (2025). Scaling Laws of Synthetic Data for Language Models. *arXiv:2503.19551*.
- Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *NeurIPS 2023*. arXiv:2305.18290.
- Rafailov, R., Chittepudi, Y., Park, R., Sikchi, H., Hejna, J., Knox, W. B., Finn, C., & Niekum, S. (2024). Scaling Laws for Reward Model Overoptimization in Direct Alignment Algorithms. *arXiv:2406.02900*.
- Rosati, D., Wehner, J., Williams, K., Bartoszcze, Ł., Batzner, J., Sajjad, H., & Rudzicz, F. (2024). Immunization against Harmful Fine-Tuning Attacks. *arXiv:2402.16382*.
- de Wynter, A., Watts, I., Altintoprak, N. E., Wongsangaroonsri, T., Zhang, M., Farra, N., et al. (2024). RTP-LX: Can LLMs Evaluate Toxicity in Multilingual Scenarios? *arXiv:2404.14397*.
- Sam, D., Goyal, S., Maini, P., Robey, A., & Kolter, J. Z. (2026). When Should We Introduce Safety Interventions During Pretraining? *arXiv:2601.07087*.
- Tasawong, P., Ngui, J. G., Aji, A. F., Cohn, T., & Limkonchotiawat, P. (2025). SEA-SafeguardBench: Evaluating AI Safety in SEA Languages and Cultures. *arXiv:2512.05501*.
- Shen, W., Liu, G., Wu, Z., Zhu, R., Yang, Q., Xin, C., Yue, Y., & Yan, L. (2025). Exploring Data Scaling Trends and Effects in Reinforcement Learning from Human Feedback. *arXiv:2503.22230*.

- Sheshadri, A., Ewart, A., Guo, P., Lynch, A., Wu, C., Hebbar, V., Sleight, H., Stickland, A. C., Perez, E., Hadfield-Menell, D., & Casper, S. (2024). Targeted Latent Adversarial Training Improves Robustness to Persistent Harmful Behaviors in LLMs. *arXiv:2407.15549*.
- Shi, Z., Zhou, Y., Li, J., Jin, Y., Li, Y., He, D., Liu, F., Alharbi, S., Yu, J., & Zhang, M. (2025). Safety Alignment via Constrained Knowledge Unlearning. *arXiv:2505.18588*.
- Shukla, V., Sharma, H., Reganti, A. N., Wasmatkar, S., Kumar, B., & Singh, V. (2026). Lost in Translation? A Comparative Study on the Cross-Lingual Transfer of Composite Harms. *AAAI 2026*. *arXiv:2602.07963*.
- Shukor, M., Bethune, L., Busbridge, D., Grangier, D., Fini, E., El-Nouby, A., & Ablin, P. (2025). Scaling Laws for Optimal Data Mixtures. *arXiv:2507.09404*.
- Simko, S., Sachan, M., Schölkopf, B., & Jin, Z. (2025). Improving Large Language Model Safety with Contrastive Representation Learning. *arXiv:2506.11938*.
- Gureja, S., Miranda, L. J., Islam, S. B., Maheshwary, R., Sharma, D., Winata, G., Lambert, N., Ruder, S., Hooker, S., & Fadaee, M. (2024). M-RewardBench: Evaluating Reward Models in Multilingual Settings. *arXiv:2410.15522*.
- Springer, M., Lee, C. P., Metevier, B., Castleman, J., Turbal, B., Jung, H., Shen, Z., & Korolova, A. (2026). The Geometry of Alignment Collapse: When Fine-Tuning Breaks Safety. *arXiv:2602.15799*.
- Jain, N., Akhtar, M., Giner-Miguel, J., Shinde, R., Vanschoren, J., Vogler, S., Goswami, S., Rao, Y., Santos, T., Oala, L., Karamousadakis, M., Maskey, M., Marcenac, P., Conforti, C., Kuchnik, M., Aroyo, L., Benjelloun, O., & Simperl, E. (2024). A Standardized Machine-readable Dataset Documentation Format for Responsible AI. *AAAI 2024*. *arXiv:2407.16883*.
- Tamirisa, R., Bharathi, B., Phan, L., Zhou, A., Gatti, A., Suresh, T., Lin, M., Wang, J., Wang, R., Arel, R., Zou, A., Song, D., Li, B., Hendrycks, D., & Mazeika, M. (2024). Tamper-Resistant Safeguards for Open-Weight LLMs. *arXiv:2408.00761*.
- Tan, Y., Jiang, Y., Li, Y., Liu, J., Bu, X., Su, W., Yue, X., Zhu, X., & Zheng, B. (2025). Equilibrate RLHF: Towards Balancing Helpfulness-Safety Trade-off in Large Language Models. *arXiv:2502.11555*.
- Tirumala, K., Simig, D., Aghajanyan, A., & Morcos, A. S. (2023). D4: Improving LLM Pre-training via Document De-Duplication and Diversification. *arXiv:2308.12284*.
- Peng, J., Zhuang, X., Qiu, J., Ma, R., Yu, J., Bai, T., & He, C. (2025). Unsupervised Topic Models are Data Mixers for Pre-training Language Models. *arXiv:2502.16802*.
- Zhu, X., Gu, Z., Zheng, S., Wang, T., Li, T., Feng, H., & Xiao, Y. (2025). ToReMi: Topic-Aware Data Reweighting for Dynamic Pre-Training Data Selection. *arXiv:2504.00695*.
- Wang, W., Tu, Z., Chen, C., Yuan, Y., Huang, J., Jiao, W., & Lyu, M. R. (2023). All Languages Matter: On the Multilingual Safety of Large Language Models. *arXiv:2310.00905*.
- Wang, Y., Huang, T., Shen, L., Yao, H., Luo, H., Liu, R., Tan, N., Huang, J., & Tao, D. (2025). Panacea: Mitigating Harmful Fine-tuning for Large Language Models via Post-fine-tuning Perturbation. *arXiv:2501.18100*.
- Wang, X., Wang, M., Liu, Y., Schütze, H., & Plank, B. (2025). Refusal Direction is Universal Across Safety-Aligned Languages. *arXiv:2505.17306*.

- Wang, D., Ma, Q., Shang, Y., Lu, Z., Ning, L., Xu, Z., Wu, H., & He, Z. (2025). Interpretable Safety Alignment via SAE-Constructed Low-Rank Subspace Adaptation. *arXiv:2512.23260*.
- Wang, Y., Irion, K., Groth, P., & Harmouch, H. (2026). “Detective Work We Shouldn’t Have to Do”: Practitioner Challenges in Regulatory-Aligned Data Quality in Machine Learning Systems. *ACM Conference on Fairness, Accountability, and Transparency*. *arXiv:2602.05944*.
- Xie, W., Tonin, F., & Cevher, V. (2025). Chameleon: A Flexible Data-mixing Framework for Language Model Pretraining and Finetuning. *ICML 2025*. *arXiv:2505.24844*.
- Wu, C., Zhang, Z., Wei, Z., Zhang, Y., & Sun, M. (2025). Mitigating Fine-tuning Risks in LLMs via Safety-Aware Probing Optimization. *arXiv:2505.16737*.
- Xie, S. M., Pham, H., Dong, X., Du, N., Liu, H., Lu, Y., Liang, P., Le, Q. V., Ma, T., & Yu, A. W. (2023). DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining. *NeurIPS 2023*. *arXiv:2305.10429*.
- Xie, S. M., Santurkar, S., Ma, T., & Liang, P. (2023). Data Selection for Language Models via Importance Resampling. *NeurIPS 2023*. *arXiv:2302.03169*.
- Xie, Z., Song, X., & Luo, J. (2025). Attack via Overfitting: 10-shot Benign Fine-tuning to Jailbreak LLMs. *NeurIPS 2025*. *arXiv:2510.02833*.
- Xie, Y., Huang, T., Yang, E., Min, R., Lu, W., Cao, X., Tan, N., & Shen, L. (2026). Mitigating Safety Tax via Distribution-Grounded Refinement in Large Reasoning Models. *arXiv:2602.02136*.
- Xu, X., Yue, X., Liu, Y., Ye, Q., Hu, H., & Du, M. (2025). Unlearning Isn’t Deletion: Investigating Reversibility of Machine Unlearning in LLMs. *arXiv:2505.16831*.
- Yang, X., Wang, X., Zhang, Q., Petzold, L., Wang, W. Y., Zhao, X., & Lin, D. (2023). Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models. *arXiv:2310.02949*.
- Yang, X., Liang, W., & Zou, J. (2024). Navigating Dataset Documentations in AI: A Large-Scale Analysis of Dataset Cards on Hugging Face. *ICLR 2024*. *arXiv:2401.13822*.
- Yang, S., Zhang, Q., Liu, Y., Huang, Y., Jia, X., Ning, K., Yao, J., Wang, J., Dai, H., Song, Y., & Yuan, L. (2025). AsFT: Anchoring Safety During LLM Fine-Tuning Within Narrow Safety Basin. *arXiv:2506.08473*.
- Ye, J., Liu, P., Sun, T., Zhou, Y., Zhan, J., & Qiu, X. (2024). Data Mixing Laws: Optimizing Data Mixtures by Predicting Language Modeling Performance. *arXiv:2403.16952*.
- Yi, X., Zheng, S., Wang, L., de Melo, G., Wang, X., & He, L. (2024). NLSR: Neuron-Level Safety Realignment of Large Language Models Against Harmful Fine-Tuning. *arXiv:2412.12497*.
- Yi, B., Li, J., Zhang, B., Nie, L., Li, T., Huang, T., & Liu, Z. (2025). SafeGrad: Gradient Surgery for Safe LLM Fine-Tuning. *arXiv:2508.07172*.
- Yoo, H., Yang, Y., & Lee, H. (2024). CSRT: Evaluation and Analysis of LLMs using Code-Switching Red-Teaming Dataset. *arXiv:2406.15481*.
- Young, R. (2026). What Is the Geometry of the Alignment Tax? *arXiv:2603.00047*.
- Zhang, B., Yang, Y., Renzhe, Guo, D., Gu, J., Torr, P., & Ghanem, B. (2025). A Guardrail for Safety Preservation: When Safety-Sensitive Subspace Meets Harmful-Resistant Null-Space. *arXiv:2510.14301*.

- Wang, Z., Liang, J., Zhu, F., Zhao, W., Fang, J., Ji, J., Wang, H., & Chua, T.-S. (2026). SafeNeuron: Neuron-Level Safety Alignment for Large Language Models. *arXiv:2602.12158*.
- Zhang, J., Hu, Y., Chen, K., He, L., Ma, J., Lou, J., Li, D., Liu, J., Yang, X., & Jia, R. (2026). Understanding and Preserving Safety in Fine-Tuned LLMs. *arXiv:2601.10141*.
- Zhao, W., Hu, Y., Deng, Y., Wu, T., Zhang, W., Guo, J., Zhang, A., Zhao, Y., Qin, B., Chua, T.-S., & Liu, T. (2025). MPO: Multilingual Safety Alignment via Reward Gap Optimization. *arXiv:2505.16869*.
- Zhou, Z., Liu, J., Dong, Z., Liu, J., Yang, C., Ouyang, W., & Qiao, Y. (2024). Emulated Disalignment: Safety Alignment for Large Language Models May Backfire! *arXiv:2402.12343*.

Acknowledgements

This paper builds directly on [Müller \(2026\)](#), which surveyed 248 papers and introduced the taxonomy of data-centric alignment interventions that structures the present work. We are grateful to the authors of the additional 370+ papers reviewed here, whose open publication made this synthesis possible.