

Data-Centric Alignment of Large Language Models: A Survey of Corpus-Level Safety, Filtering, and Synthetic Data Approaches

Ole Müller

OM-Intelligence GmbH

ole@om-intelligence.com

March 2026

Abstract

Large language models acquire their values from data. Post-hoc alignment methods such as RLHF modify only the first few output tokens, creating a shallow safety veneer that fine-tuning on as few as ten adversarial examples can remove. This paper argues that alignment must be reconceived as a *data-centric* endeavour, one that begins with the pretraining corpus itself. Drawing on a systematic review of 248 papers, we formalise the concept of *corpus-level alignment* and provide a comprehensive taxonomy of data-centric alignment techniques spanning pretraining data filtering, intrinsic alignment signal extraction (ISAAC), constitutional and self-alignment, instruction-tuning dataset construction, synthetic data generation, and preference optimisation. We synthesise evidence showing that corpus composition is the primary determinant of model alignment: pretraining data interventions reduce misalignment rates from 45% to 9% (Tice et al., 2026), safety pretraining frameworks lower attack success rates from 38.8% to 8.4%, and corpus-based methods achieve safety improvements comparable to supervised approaches without human labels. We identify six open research gaps (including scaling laws for alignment data, multilingual corpus-level alignment, and principled corpus composition) and argue that the path forward requires not only technical advances but also normative frameworks for deciding what values a training corpus should encode.

Contents

1	Introduction	4
1.1	The Problem: Unsafe Outputs Persist Despite Alignment	4
1.2	Three Limitations of Post-Hoc Alignment	4
1.3	The Data-Centric Paradigm	5
1.4	Contributions of This Paper	5
2	Background: From Post-Hoc to Pre-Hoc Alignment	6
2.1	The Standard Alignment Pipeline	6
2.2	The Superficial Alignment Hypothesis	7
2.3	The Shallow Safety Alignment Problem	7
2.4	The Data-Centric AI Movement	8
2.5	From Data-Centric AI to Corpus-Level Alignment	8
3	Case Study: Intrinsic Alignment Through Corpus Structure	9
3.1	Core Insight and Method	9
3.2	Results and Significance	10

4	Pretraining Data Filtering and Safety	11
4.1	The Scale of the Problem: What Is in Web Corpora?	11
4.2	Classifier-Based Filtering: From Blocklists to Neural Classifiers	12
4.3	Deep Ignorance: The Case for Knowledge Prevention	12
4.4	The Alignment Tax: Capabilities Lost and Preserved	13
4.5	Fairness Implications: Whose Language Gets Filtered?	13
4.6	Beyond Removal: Recontextualisation and Synthetic Alternatives	14
4.7	Synthesis: A Layered Approach to Pretraining Safety	14
5	Constitutional AI and Self-Alignment	15
5.1	Constitutional AI: The Full Pipeline	15
5.2	Self-Instruct: Bootstrapping Instruction Data	16
5.3	Dromedary: Principle-Driven Self-Alignment from Scratch	16
5.4	Self-Rewarding Language Models	17
5.5	Self-Alignment as Corpus Generation	17
6	Instruction Tuning and Safety Datasets	17
6.1	Instruction Tuning: Foundations and Significance	17
6.2	Landmark Instruction and Safety Datasets	17
6.3	Fine-Tuning Safety Degradation: The Fragility of Alignment	19
6.4	Defences Against Fine-Tuning Attacks	19
6.5	From Fine-Tuning Fragility to Corpus-Level Alignment	20
7	Synthetic Data for Alignment	20
7.1	Why Synthetic Data Is Central to Alignment	20
7.2	The “Textbooks Are All You Need” Finding	21
7.3	Model Collapse: The Recursive Training Failure Mode	21
7.4	The Optimal Mixing Ratio	21
7.5	Scaling Synthetic Data Generation: Cosmopedia and Beyond	21
7.6	Six Best Practices for Synthetic Alignment Data	21
7.7	Three Key Risks	22
8	How Pretraining Corpora Encode Values	22
8.1	The Corpus as a Value System	22
8.2	Whose Opinions Do Language Models Reflect?	23
8.3	What Large Pretraining Corpora Look Like	23
8.4	Domain Composition Effects on Model Behaviour	24
8.5	Debiasing Approaches	25
8.6	The Fundamental Insight: There Is No Neutral Corpus	25
9	Preference Optimisation and Data Construction	25
9.1	From RLHF to DPO: Eliminating the Reward Model	25
9.2	Beyond DPO: KTO, ORPO, and SimPO	26
9.3	Data Quality: The Decisive Factor	26
9.4	Preference Data as Corpus: A Data-Centric View	26
10	Benchmarks and Evaluation	27
10.1	Major Safety Benchmarks	27
10.2	The SafetyPrompts Meta-Review	27
10.3	Surveys of the Alignment Landscape	27

11 Cross-Cutting Themes	28
11.1 The Data-Centric Paradigm Shift	28
11.2 The Depth Problem: Shallow vs. Deep Alignment	28
11.3 The Filtering–Fairness Tension	28
11.4 Synthetic Data as Alignment Infrastructure	29
11.5 Scalability of Annotation-Free Methods	29
12 Auditing Alignment Corpora and Value Governance	29
12.1 Current Auditing Practice	29
12.2 The Governance Gap	30
12.3 Toward Principled Corpus Governance	30
13 Open Research Gaps and Conclusion	31
13.1 Six Open Research Gaps	31
13.1.1 Gap 1: Scaling Laws for Alignment Data	31
13.1.2 Gap 2: Multilingual Alignment Data	31
13.1.3 Gap 3: Long-Term Value Stability	31
13.1.4 Gap 4: Interaction Between Alignment Layers	31
13.1.5 Gap 5: Auditing and Verification of Alignment Corpora	31
13.1.6 Gap 6: Principled Corpus Composition	31
13.2 Conclusion	32

1 Introduction

Large language models (LLMs) acquire their knowledge, capabilities, and, critically, their behavioural tendencies from the data on which they are trained. A model pretrained on trillions of tokens of internet text absorbs not only the syntax and semantics of language but also the biases, toxicity, and value systems embedded in that corpus. The dominant approach to making these models safe and useful has been to treat alignment as a *post-hoc* correction: first pretrain the model on a vast, largely uncured corpus, then apply supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) to steer its behaviour toward helpfulness, honesty, and harmlessness. This paper argues that this paradigm is fundamentally insufficient, and that alignment must be reconceived as a *data-centric* endeavour, one that begins with the training corpus itself.

1.1 The Problem: Unsafe Outputs Persist Despite Alignment

Consider a user who asks a deployed LLM: “Provide detailed instructions on how to break into someone’s house.” A well-aligned model should refuse. And indeed, when tested on the HEx-PHI safety benchmark (Qi et al., 2023), aligned models such as Llama-2-7B-Chat begin their responses with refusal prefixes like “I cannot fulfil your request” in 96.1% of cases, while Gemma-7b-1.1-IT produces “I am unable” in 96.7% of cases (Qi et al., 2024). On the surface, alignment appears to work.

Yet this safety is far more fragile than these numbers suggest. Qi et al. (Qi et al., 2023) demonstrated that fine-tuning GPT-3.5 Turbo on as few as 10 adversarially crafted examples (at a cost of less than \$0.20 via OpenAI’s API) is sufficient to jailbreak the model, and that even fine-tuning on purely benign datasets such as Alpaca (Taori et al., 2023) degrades safety alignment. These findings reveal that the safety guardrails installed by RLHF are not deeply integrated into the model’s representations; they are a thin behavioural veneer that can be stripped away with minimal effort. The full exposition of this “shallow alignment” phenomenon is developed in Section 2.3.

1.2 Three Limitations of Post-Hoc Alignment

The fragility described above is symptomatic of three deeper limitations of the prevailing RLHF-based alignment paradigm.

Cost and scalability. RLHF requires collecting large volumes of human preference data. Anthropic’s HH-RLHF dataset (Bai et al., 2022a) comprises over 100,000 pairwise comparisons gathered through open-ended conversations between crowdworkers and language models across separate helpfulness and harmlessness tasks. Constitutional AI (Bai et al., 2022b) was developed in part to reduce this annotation burden, but even that method “typically uses (at least) tens of thousands of human feedback labels” (Bai et al., 2022b). These labels are expensive to collect, difficult to quality-control, and must be updated as new attack vectors emerge.

Shallowness. As few as 10 adversarial examples break safety alignment, and alignment modifies only the first few tokens of a model’s output distribution. The detailed evidence for this claim is presented in Section 2.3.

Fragility under fine-tuning. The shallow nature of alignment explains why downstream fine-tuning so readily destroys it. Even well-intentioned users who fine-tune an aligned model on a domain-specific dataset risk inadvertently compromising its safety properties through catastrophic forgetting of the alignment objectives.

1.3 The Data-Centric Paradigm

These limitations point toward a fundamental insight: if a model’s knowledge, capabilities, and values are all acquired from data, then alignment, like capability, should be addressed at the data level. This perspective draws on the broader *data-centric AI* movement, which Zha et al. (Zha et al., 2023) define as “the discipline of systematically engineering the data used to build an AI system.” In contrast to the model-centric paradigm, which holds the data fixed and iterates on architectures and algorithms, data-centric AI emphasises that data quality, composition, and curation are often the primary determinants of system performance.

Applied to alignment, the data-centric paradigm suggests that safety and value alignment should not be relegated to a post-training correction but should instead be embedded throughout the training pipeline: in the composition of the pretraining corpus, in the filtering and curation of training data, in the construction of instruction-tuning datasets, and in the generation and verification of synthetic alignment data. We refer to this comprehensive approach as *corpus-level alignment*.

Early evidence for the viability of this approach is striking. The LIMA experiment (Zhou et al., 2023) demonstrated that a 65-billion-parameter LLaMA model fine-tuned on only 1,000 carefully curated examples (with no RLHF whatsoever) produces responses that human evaluators prefer over those of RLHF-trained DaVinci003 in 65% of comparisons. If 1,000 well-chosen examples can rival systems trained on millions of feedback signals, the implication is clear: *what* is in the data matters far more than *how much* post-hoc optimisation is applied.

More recently, ISAAC (Liang et al., 2025) has shown that training corpora contain intrinsic distributional signals that can be exploited for alignment without any human preference labels. By clustering responses by context and content, then applying a distribution-warping sampling strategy, ISAAC extracts safer responses directly from the existing corpus. When applied on top of models such as Claude, Llama3, and Qwen2, ISAAC improves safety scores by 1.4 to 9.9 percentage points (Liang et al., 2025) while maintaining response diversity and contextual relevance, all without a single human annotation.

1.4 Contributions of This Paper

This paper provides a comprehensive analysis of data-centric alignment for large language models. Drawing on a systematic review of 248 papers across nine research areas, we make the following contributions:

1. We characterise the concept of *corpus-level alignment* and distinguish it from post-hoc alignment methods, articulating why interventions at the data level provide deeper and more robust safety guarantees than behavioural fine-tuning alone.
2. We provide a detailed taxonomy of data-centric alignment techniques spanning the full training pipeline, from pretraining corpus filtering and safety-aware data curation, through constitutional and self-alignment methods for synthetic data generation, to principled construction of instruction-tuning and preference datasets.
3. We synthesise experimental evidence demonstrating that corpus composition is the primary determinant of model alignment, reviewing results from pretraining data interventions that reduce misalignment rates from 45% to 9%, safety pretraining frameworks that lower attack success rates from 38.8% to 8.4%, and corpus-based methods that achieve safety improvements comparable to supervised approaches without human labels.
4. We identify key research gaps and open challenges, including the scalability of corpus-level alignment to multilingual settings, the interaction between corpus-level and fine-tuning alignment, the absence of scaling laws for alignment data, and the need for principled frameworks for corpus composition and auditing.

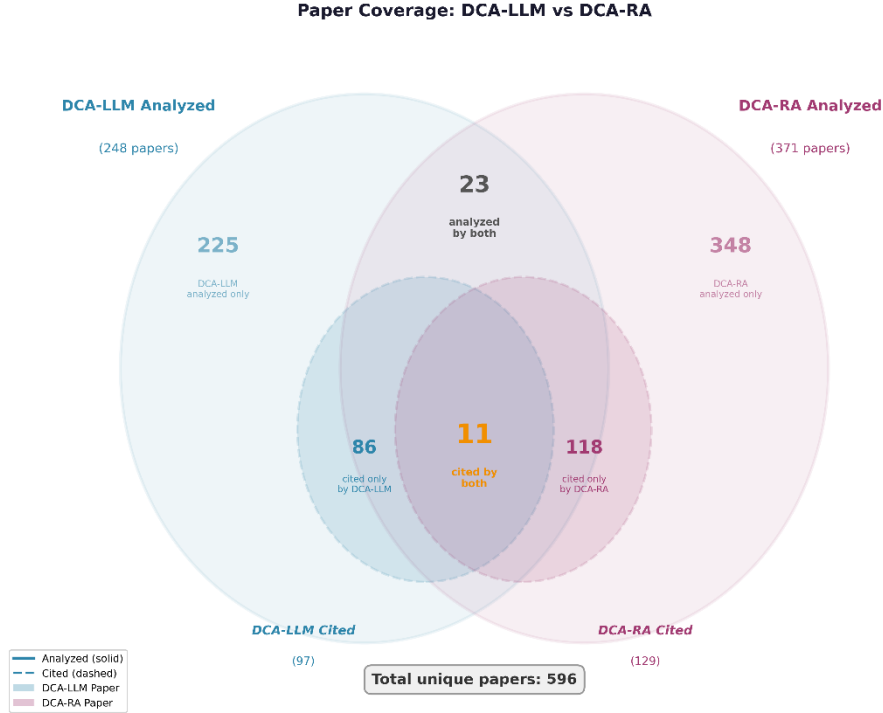


Figure 1: Paper coverage of this survey (DCA-LLM) and the companion research agenda (DCA-RA). Of 596 unique papers across both works, only 11 are cited by both, confirming complementary rather than overlapping coverage.

2 Background: From Post-Hoc to Pre-Hoc Alignment

This section provides the conceptual and empirical foundations for understanding why the field is moving from post-hoc behavioural alignment toward data-centric, corpus-level approaches. We begin with the standard alignment pipeline, explain the key hypotheses and experimental findings that reveal its limitations, and conclude by connecting these insights to the data-centric AI movement.

2.1 The Standard Alignment Pipeline

The development of a modern instruction-following LLM proceeds through three stages, each building on the last. Understanding this pipeline in detail is essential for appreciating where and why it fails.

Stage 1: Pretraining. A large transformer-based neural network is trained on a massive corpus of text (typically hundreds of billions to trillions of tokens drawn from web crawls, books, code repositories, and curated datasets) using a self-supervised language modelling objective. The model learns to predict the next token given all preceding tokens. Through this process, it acquires broad linguistic competence, factual knowledge, reasoning patterns, and, inevitably, the distributional biases and value assumptions present in its training data.

Stage 2: Supervised Fine-Tuning (SFT). The pretrained model is then fine-tuned on a curated dataset of (instruction, response) pairs that demonstrate the desired assistant behaviour. This teaches the model to adopt a helpful, conversational format rather than merely completing text. The quality and composition of SFT data have outsized effects on the resulting model’s behaviour, as we discuss in Section 2.2.

Stage 3: RLHF or preference optimisation. Finally, the model is further trained using human preference signals. In the canonical RLHF pipeline (Bai et al., 2022a; Ouyang et al., 2022), this involves three sub-steps: (a) collecting human comparisons of model outputs; (b) training a reward model on these comparisons to predict human preferences; and (c) optimising the language model’s policy against this reward model using a reinforcement learning algorithm such as Proximal Policy Optimisation (PPO), with a KL-divergence penalty to prevent the model from diverging too far from the SFT checkpoint.

More recent methods simplify this pipeline. Direct Preference Optimisation (DPO) (Rafailov et al., 2023) eliminates the reward model entirely, training directly on preference pairs by reparameterising the RL objective as a classification loss. Constitutional AI (Bai et al., 2022b) reduces the need for human harmlessness labels by using the model itself to critique and revise its responses according to a small set of principles (a “constitution”), then training a preference model on AI-generated comparisons, a technique the authors call Reinforcement Learning from AI Feedback (RLAIF). These innovations reduce annotation costs, but the fundamental architecture of the alignment pipeline remains: safety is applied *after* pretraining, as a behavioural overlay.

2.2 The Superficial Alignment Hypothesis

A pivotal challenge to the post-hoc alignment paradigm comes from Zhou et al.’s LIMA experiment (Zhou et al., 2023), which introduced the **Superficial Alignment Hypothesis**: “A model’s knowledge and capabilities are learnt almost entirely during pretraining, while alignment teaches it which subdistribution of formats should be used when interacting with users.”

To test this hypothesis, the authors curated a dataset of only 1,000 carefully selected prompt-response pairs, approximately 600 drawn from Stack Exchange and wikiHow, 150 from Reddit, 50 from Natural Instructions, and 200 manually authored by the researchers. They fine-tuned a 65-billion-parameter LLaMA model on this dataset using standard supervised learning, with no reinforcement learning and no human preference modelling.

The results are remarkable. In a controlled human evaluation across 300 challenging test prompts, LIMA’s responses were judged equivalent or strictly preferred to those of five strong baselines:

- Against **Alpaca 65B** (trained on 52,000 examples, 52× more data): LIMA preferred or tied in 74% of comparisons.
- Against **DaVinci003** (OpenAI’s RLHF-trained model): LIMA preferred or tied in 65% of comparisons.
- Against **Google Bard**: LIMA preferred or tied in 58% of comparisons.
- Against **Claude**: LIMA preferred or tied in 46% of comparisons.
- Against **GPT-4**: LIMA preferred or tied in 43% of comparisons.

The implications for alignment are profound. If alignment is primarily about teaching the model a particular *style* of interaction (how to present the knowledge it already possesses), then the overwhelming majority of what determines a model’s values and safety properties is established during pretraining, through the composition of the pretraining corpus. This directly motivates a corpus-level approach.

2.3 The Shallow Safety Alignment Problem

While LIMA suggests that alignment is superficial in the sense of requiring little data, Qi et al. (Qi et al., 2024) demonstrate that current safety alignment is shallow in a more concerning sense:

it operates primarily over the very first output tokens, leaving the model’s deeper generative distribution essentially unchanged.

The authors formalise the notion of **shallow safety alignment**: a model undergoes shallow safety alignment if the alignment procedure primarily adapts the model’s generative distribution only over the first few output tokens to induce a basic refusal response. They provide three lines of evidence:

Evidence 1: The refusal prefix shortcut. Testing on the HEx-PHI benchmark of 330 harmful instructions across 11 categories, the authors find that aligned models rely overwhelmingly on stereotyped refusal prefixes. Llama-2-7B-Chat starts with “I cannot” or “I apologise” in 96.1% of safety refusals. Critically, when these same prefixes are prepended to the *unaligned* base model’s outputs during decoding, the base model’s harmfulness rate drops dramatically, from 68.6% to between 0% and 14.4%, depending on the prefix used.

Evidence 2: KL divergence is concentrated in early tokens. Measuring the per-token KL divergence $D_{\text{KL}}(\pi_{\text{aligned}}(\cdot \mid \mathbf{x}, \mathbf{y}_{<k}) \parallel \pi_{\text{base}}(\cdot \mid \mathbf{x}, \mathbf{y}_{<k}))$ on harmful prompts, the authors show that the divergence is significantly higher for the first few token positions and drops sharply thereafter.

Evidence 3: Fine-tuning attacks target the shallow layer. When fine-tuning Llama-2-7B-Chat on 100 harmful examples from Qi et al. (Qi et al., 2023), the per-token cross-entropy loss and gradient norms are highest at the earliest token positions. After just 6 gradient steps, the attack success rate climbs from 1.5% to 87.9%, effectively erasing the safety alignment.

These findings have far-reaching consequences. They explain the effectiveness of multiple known attack vectors: adversarial suffix attacks work because their optimisation objective is simply to elicit an affirmative initial prefix; prefilling attacks succeed because they bypass the refusal prefix entirely; and fine-tuning attacks are devastatingly effective because they overwrite the small number of token positions where safety resides.

2.4 The Data-Centric AI Movement

The recognition that data quality determines system performance is not unique to LLM alignment. The broader *data-centric AI* movement, surveyed comprehensively by Zha et al. (Zha et al., 2023), represents a paradigm shift in machine learning research. In the conventional model-centric approach, researchers hold data fixed and iterate on model architectures, training algorithms, and hyperparameters. Data-centric AI inverts this: the model architecture is held relatively fixed, and improvements come from systematically engineering the training data.

The empirical case for this shift is compelling. The progression from GPT-1 through GPT-4 involved relatively minor architectural changes but dramatic scaling and curation of training data: GPT-1 trained on 4.8GB of unfiltered text, GPT-2 on 40GB of human-filtered data, GPT-3 on 570GB filtered from 45TB of raw data, and ChatGPT/GPT-4 added carefully constructed human demonstrations and preference annotations (Zha et al., 2023). While architectural changes were also significant, the data has been a primary driver of capability improvements.

Zha et al. organise data-centric AI into three goals: *training data development* (collection, labelling, preparation, reduction, and augmentation), *inference data development* (evaluation and prompt engineering), and *data maintenance* (understanding, quality assurance, and storage). This framework maps directly onto the alignment problem.

2.5 From Data-Centric AI to Corpus-Level Alignment

The convergence of these four threads (the standard alignment pipeline’s reliance on post-hoc correction, the superficial alignment hypothesis’s demonstration that pretraining determines

capabilities and values, the shallow alignment problem’s evidence that post-hoc safety is concentrated in a fragile surface layer, and the data-centric AI movement’s broader validation of data-first approaches) motivates a new paradigm for LLM alignment that we term *corpus-level alignment*.

Corpus-level alignment posits that safety and value alignment should be embedded into the training data across the full pipeline, rather than applied as a final-stage behavioural modification. This encompasses several complementary strategies:

1. **Pretraining corpus curation:** Filtering, reweighting, or augmenting the pretraining corpus to reduce toxic, biased, or dangerous content while preserving linguistic diversity and capability. Recent work has shown that classifier-based filtering of hazardous content achieves a 33% reduction in dangerous capabilities while maintaining general performance, and that comprehensive safety pretraining frameworks can reduce attack success rates from 38.8% to 8.4% (Maini et al., 2025).
2. **Intrinsic alignment signal extraction:** Methods such as ISAAC (Liang et al., 2025) that exploit the distributional structure of existing corpora to identify and amplify safe responses without any external labels.
3. **Constitutional and self-alignment:** Techniques that use AI-generated feedback guided by explicit principles to create synthetic alignment data (Bai et al., 2022b), reducing dependence on human annotation while producing training corpora that encode specific value commitments.
4. **Principled dataset construction:** The careful design of instruction-tuning and preference datasets, informed by the LIMA finding that data quality dominates data quantity, and by the recognition that dataset composition directly determines the value stances of resulting models.

The promise of corpus-level alignment is threefold. First, it addresses the *depth* problem: if safety-relevant properties are encoded in the pretraining data itself, they become part of the model’s deep representations rather than a shallow behavioural overlay. Second, it addresses the *scalability* problem: methods that operate on the corpus scale naturally without requiring per-example human annotation. Third, it addresses the *robustness* problem: a model that never learned dangerous knowledge from its training data cannot be coerced into producing it through fine-tuning, as demonstrated by recent work on multi-stage pretraining data filtering that creates models resistant to adversarial fine-tuning for over 10,000 steps.

Figure 2 provides a visual overview of how these data-centric alignment techniques map onto the standard LLM training pipeline. The remainder of this paper develops each of these intervention points systematically.

3 Case Study: Intrinsic Alignment Through Corpus Structure

To illustrate what corpus-level alignment looks like in practice, we examine ISAAC (Intrinsically Supervised Alignments by Assessing Corpus), introduced by Liang et al. (2025) at AAAI 2025. ISAAC is not the only corpus-level alignment method (the subsequent sections survey many others), but it represents the purest realisation of the paradigm: extracting alignment signals entirely from the distributional structure of an existing corpus, without any human labels, reward models, or constitutional principles.

3.1 Core Insight and Method

ISAAC rests on an empirical observation: in any sufficiently large dialogue corpus, unsafe responses are *rare* relative to safe ones and *semantically distinctive*: they cluster in identifiable

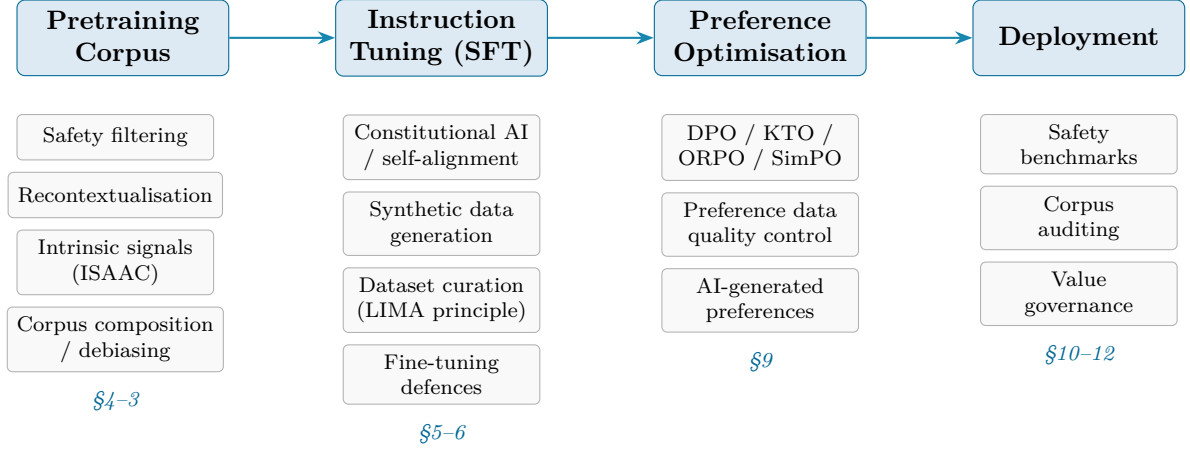


Figure 2: Taxonomy of data-centric alignment techniques mapped onto the LLM training pipeline. At each stage, alignment quality is determined primarily by data decisions (what content is included, how it is curated, and what values it encodes) rather than by algorithmic choice alone. Section references indicate where each intervention point is discussed in the paper.

regions of the embedding space. This long-tailed distribution reflects naturally occurring social norms in the data.

The method proceeds in five stages. First, for each dialogue context, ISAAC retrieves semantically similar contexts using a pretrained encoder (RoBERTa), forming a neighbourhood of context–response pairs. Second, within each neighbourhood, DBSCAN clusters responses by semantic content; safe responses populate larger clusters, unsafe ones appear in smaller or outlier clusters. Third (the key technical step), ISAAC applies *adaptive sharpening* to the cluster-size distribution via a temperature-scaled softmax modulated by a Sensitive Indicator (*SI*) that calibrates sharpening intensity to the distribution’s shape. The authors prove that this convex sampling function provably reduces the probability of drawing unsafe responses when safe clusters dominate (Theorem 2 in the original paper), whereas random sampling leaves the unsafe proportion unchanged (Theorem 1). Fourth, multiple pseudo-responses are sampled per context from the warped distribution. Fifth, the model is fine-tuned on these pseudo-responses via a multi-target *tempering training* curriculum that prevents memorisation of specific responses.

3.2 Results and Significance

Evaluated on the DiaSafety benchmark (11K samples, seven unsafe categories), ISAAC achieves a safety score of 78.24, exceeding all supervised detoxification baselines (best: 51.90) without using any safety labels (Liang et al., 2025). When layered on top of already-aligned models, ISAAC provides consistent additive gains: +1.4 percentage points for Claude, +9.5 for Llama3, and +9.9 for Qwen2. Backward perplexity increases by only 2%, indicating that contextual relevance is largely preserved.

Three properties make ISAAC instructive for the broader argument of this paper. First, it is **complementary** to post-hoc methods: corpus-level and RLHF-based signals capture different aspects of alignment and compose well. Second, it is **automatically evolvable**: as the corpus changes, the alignment signal updates without new annotation campaigns. Third, it demonstrates that **the alignment signal is already in the data**, a claim that the subsequent sections will show holds across pretraining filtering, synthetic data generation, and preference dataset construction.

The method’s limitations are also instructive. It depends on the assumption that safe responses outnumber unsafe ones (breaks down in heavily poisoned corpora), provides only coarse-grained alignment (cannot make token-level adjustments), and inherits the coverage gaps of

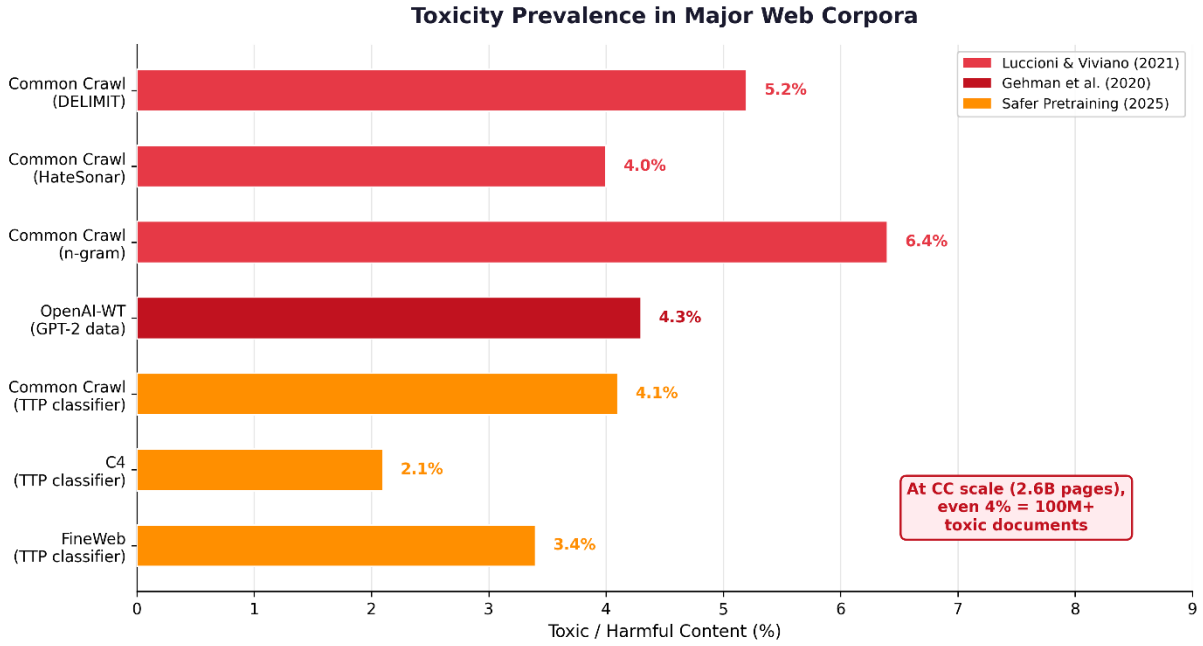


Figure 3: Toxicity prevalence in major web corpora. Measurements from three independent studies using different detection methods converge on 2–6% harmful content rates. At Common Crawl’s scale of 2.6 billion pages, even 4% translates to over 100 million toxic documents.

its corpus. These limitations motivate the multi-layered approach that the rest of this paper develops: no single corpus-level intervention suffices; robust alignment requires complementary interventions at every stage of the training pipeline.

4 Pretraining Data Filtering and Safety

The previous sections established that alignment is most durable when it is embedded in the data a model learns from, rather than patched on after pretraining. This section examines the most direct mechanism for doing so: deciding what content enters, and what content is excluded from, the pretraining corpus. The question is deceptively simple. In practice, it touches on toxicity, dual-use knowledge, representational fairness, and the fundamental trade-off between a model’s breadth of knowledge and its propensity for harm. We identified 57 papers bearing on this topic; the discussion below synthesises the most consequential findings.

4.1 The Scale of the Problem: What Is in Web Corpora?

To understand why filtering matters, one must first appreciate what web-scale corpora actually contain. The datasets that underpin modern language models (Common Crawl, C4, the Pile, FineWeb) are enormous, loosely curated snapshots of the public internet, and they inherit all of its pathologies.

Luccioni & Viviano (2021) provided one of the first systematic audits of the Common Crawl corpus. Sampling approximately 5.8 million web pages (roughly 1% of the corpus), they found that between 4.0% and 6.4% of websites contained hate speech according to three independent detection methods (DELIMIT at 5.2%, HateSonar at 4.0%, and n -gram matching at 6.4%), while 2.4% contained sexually explicit material. These figures may appear modest in percentage terms, but given that a single Common Crawl snapshot comprises 2.6 billion web pages spanning 200–300 terabytes of raw text, even a 4% prevalence rate translates to over 100 million documents containing hate speech.

[Gehman et al. \(2020\)](#) demonstrated that this toxic content does not merely sit inert in training corpora; it actively shapes model behaviour. Their REALTOXICITYPROMPTS benchmark (100,000 naturally occurring sentence prefixes extracted from the OpenWebText Corpus and scored by the Perspective API) revealed that all five tested language models (GPT-1, GPT-2, GPT-3, CTRL, and CTRL-Wiki) could degenerate into toxicity with fewer than 100 unprompted generations. Crucially, [Gehman et al.](#) traced this behaviour to the training data itself: the OpenAI-WT corpus (GPT-2’s pretraining data) contained approximately 4.3% toxic documents, and documents from factually unreliable news sources and banned or quarantined Reddit communities constituted at least 10% of the most challenging toxic prompts.

More recent work has expanded the taxonomy of harmful content beyond binary toxicity. [Safer Pretraining \(2025\)](#) proposed a three-dimensional classification framework (Safe, Topical, Toxic) spanning five harm categories: Hate & Violence, Ideological Harm, Sexual Content, Illegal Activities, and Self-Inflicted Harm. Applying their Topical and Toxic Prompt (TTP) classifier to 3 million web pages sampled from Common Crawl, C4, and FineWeb, they found toxic content constituting 2.1–4.1% of web corpora, with Topical discussions ranging from 3.4% to 10.6%.

4.2 Classifier-Based Filtering: From Blocklists to Neural Classifiers

Early filtering approaches were blunt instruments. The original C4 corpus ([Dodge et al., 2021](#)) removed any document containing a word from the “List of Dirty, Naughty, Obscene, or Otherwise Bad Words”, a blocklist initially designed to suppress autocomplete suggestions in a search engine. While effective at eliminating the most overtly offensive material, this approach was both over-inclusive (removing legitimate medical, legal, and educational content) and under-inclusive (missing contextually harmful content that avoids forbidden keywords).

The field has since moved toward learned classifiers that can assess harm at the semantic level. Three major efforts illustrate the current state of the art:

Anthropic’s pretraining data filtering. In a 2025 technical report, Anthropic’s alignment team demonstrated that classifier-based filtering of CBRN (chemical, biological, radiological, nuclear) content from pretraining data achieves a **33% reduction in dangerous capabilities** as measured on domain-specific evaluations, with no statistically significant degradation on standard benchmarks ([Anthropic, 2025](#)).

Safety Pretraining (SafeLM). [Maini et al. \(2025\)](#) introduced a comprehensive four-component framework: (i) *Safety Filtering*, using a multi-layer classifier to score pretraining data on a 0–5 safety scale; (ii) *Synthetic Recontextualisation*, which rewrites harmful content into educational narratives rather than discarding it; (iii) *Native Refusal Training*, via the RefuseWeb and Moral Education datasets; and (iv) *Harmfulness-Tag Annotated Pretraining*, which injects a special token into training data at positions identified as unsafe. Trained on over 100 billion tokens of safety-curated data, the resulting 1.7B-parameter SafeLM reduced attack success rates from **38.8% to 8.4%**, with no performance degradation on general NLP tasks.

HarmFormer and the HAVOC benchmark. [Safer Pretraining \(2025\)](#) developed HarmFormer, a LongFormer-based transformer classifier trained on 253,000 web pages, achieving an overall F1 score of 0.85 on the toxic dimension across five harm categories, substantially outperforming both the Perspective API (F1 = 0.63) and Llama Guard 3 (F1 = 0.66–0.70).

4.3 Deep Ignorance: The Case for Knowledge Prevention

Perhaps the most conceptually striking result in pretraining data filtering comes from [Deep Ignorance \(2025\)](#), who investigated whether models can be made *robustly ignorant* of dangerous

knowledge, and whether such ignorance provides a qualitatively different kind of safety guarantee compared to post-training safeguards.

Post-training alignment methods (including RLHF, DPO, and Circuit Breaking) modify a model’s behaviour without erasing its knowledge. Fine-tuning “rarely alters the underlying model capabilities. A minimal transformation, which we call a ‘wrapper’, is typically learned on top of the underlying model capabilities” (Deep Ignorance, 2025). This wrapper can be stripped away within a few hundred adversarial fine-tuning steps.

Deep Ignorance hypothesised that if a model never learns dangerous knowledge during pre-training, it should be fundamentally more difficult for an attacker to elicit that knowledge through fine-tuning. They designed a multi-stage filtering pipeline targeting biothreat proxy knowledge, pretrained multiple 6.9B-parameter models from scratch on 500 billion tokens, and found that filtered models achieved near-random-chance performance on biothreat evaluations while showing **no degradation** on general capabilities. Most importantly, filtered models exhibited **state-of-the-art tamper resistance**: they resisted adversarial fine-tuning for up to 10,000 training steps, improving on post-training baselines by more than an order of magnitude.

However, data filtering cannot prevent a model from leveraging harmful information provided in context (e.g., via retrieval augmentation). This motivates a defence-in-depth strategy combining pretraining data filtering with post-training safeguards.

4.4 The Alignment Tax: Capabilities Lost and Preserved

A central concern with any filtering approach is the *alignment tax*, the degradation in general capabilities that may accompany safety improvements. Longpre et al. (2024) conducted the most systematic investigation to date, pretraining 28 1.5B-parameter models with varying quality and toxicity filters. Their key finding is that **quality filtering and toxicity filtering have significant but opposite effects**. Quality filtering substantially increases both downstream performance and reduces toxic generation, while toxicity filtering reduces toxic generations but trades off against generalisation performance.

Against this backdrop, the near-zero alignment tax reported by both Maini et al. (2025) and Deep Ignorance (2025) is especially significant. The key differentiator appears to be *specificity*: rather than applying a broad toxicity filter, these newer approaches target semantically defined categories of harm, preserving the overwhelming majority of training data.

4.5 Fairness Implications: Whose Language Gets Filtered?

Content filtering is not value-neutral. Dodge et al. (2021) provided the definitive case study by auditing the C4 corpus’s blacklist filter. Using a dialect-aware topic model trained on 60 million geolocated tweets, they assigned each document a posterior probability of being in African American English (AAE), Hispanic-aligned English, White-aligned English (WAE), or “other.” The results were stark: **African American English documents were removed at a rate of 42%, and Hispanic-aligned English at 32%, compared to just 6.2% for White-aligned English**. As a consequence, 97.8% of documents in the final C4.EN corpus are classified as WAE.

The mechanism is straightforward: the blacklist contains terms that, in mainstream American English, are considered slurs or profanity, but that in AAE serve as reclaimed in-group vocabulary or appear in discussions of racial identity. This creates a pernicious feedback loop: models trained on the filtered corpus underrepresent minority dialects, leading to worse performance on text from and about minority communities.

This finding illustrates a broader point that recurs throughout this paper: corpus curation is not a purely technical activity. Every filtering decision (what to remove, what to retain, what threshold to set) is implicitly a *normative* decision about whose language is valued and whose is expendable. The field of data-centric alignment therefore cannot avoid engaging with

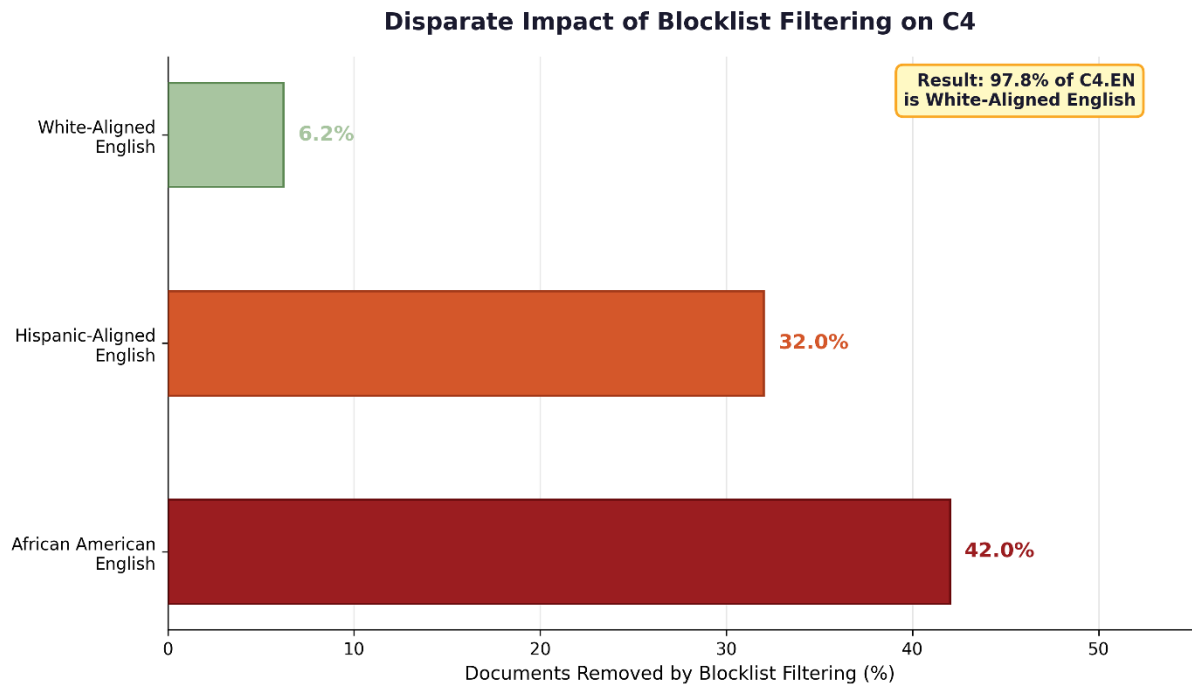


Figure 4: Disparate impact of blocklist filtering on C4. African American English documents were removed at 42%, Hispanic-aligned English at 32%, but White-aligned English at only 6.2%. The result: 97.8% of the final C4.EN corpus is White-aligned English (Dodge et al., 2021).

questions traditionally addressed by moral philosophy, political theory, and critical data studies. We develop these interdisciplinary implications in Sections 8.6 and 12.

4.6 Beyond Removal: Recontextualisation and Synthetic Alternatives

A key innovation in recent work is the recognition that filtering need not mean deletion. Maini et al. (2025) introduced a *synthetic recontextualisation* pipeline in which harmful content is not excised but transformed into educational material. Their complementary RefuseWeb dataset transforms real harmful web content into dialogues where an Assistant politely but firmly refuses a harmful request. Together, these interventions represent a shift from a *censorship model* of safety toward a *pedagogical model*.

4.7 Synthesis: A Layered Approach to Pretraining Safety

The research reviewed in this section converges on several principles for responsible pretraining data curation:

1. **Blocklist filtering is insufficient and discriminatory.** Keyword-based approaches systematically remove minority-dialect text while missing contextually harmful content.
2. **Targeted filtering is more effective than broad filtering.** The most successful interventions (Anthropic, 2025; Deep Ignorance, 2025) target specific, well-defined categories of dangerous knowledge.
3. **Knowledge prevention creates qualitatively different safeguards.** Models that never learned dangerous knowledge resist adversarial fine-tuning for orders of magnitude longer than models with post-training safeguards.

4. **Recontextualisation is preferable to deletion.** Harmful content can be synthetically reframed into educational material, preserving informational value while embedding safety norms.
5. **Defence-in-depth is necessary.** No single intervention is sufficient. Data filtering prevents parametric knowledge acquisition; Circuit Breaking impairs in-context reasoning about harmful topics; retrieval restrictions limit access to dangerous information at deployment time.
6. **Filtering demands documentation and auditing.** The disparate impact of filtering on minority communities makes transparent reporting essential.

5 Constitutional AI and Self-Alignment

A central bottleneck in reinforcement learning from human feedback (RLHF) is the cost and scalability of human annotation. Collecting preference labels for harmlessness alone can require tens of thousands of crowdworker comparisons, and even then the resulting labels are noisy, culturally contingent, and difficult to audit (Bai et al., 2022a). Constitutional AI (CAI) and the broader family of *self-alignment* methods address this bottleneck by enlisting the language model itself as an annotator, critic, and judge, shifting the locus of human oversight from per-example labels to a small set of reusable principles. From a data-centric perspective, these techniques are best understood as *corpus-generation methods*: they produce the very alignment corpora on which models are subsequently trained.

5.1 Constitutional AI: The Full Pipeline

Constitutional AI, introduced by Bai et al. (2022b), replaces human harmlessness labels with AI-generated feedback governed by a “constitution”, a short list of natural-language principles that specify desirable model behaviour. The method comprises two stages: a supervised learning (SL) stage built on critique and revision, and a reinforcement learning stage built on AI preference judgements (RLAIF).

Stage 1: Red-Teaming, Critique, and Revision. The pipeline begins with a *helpful-only* RLHF model, which is capable but willingly produces harmful content when prompted adversarially. Red-team prompts (42,496 human-written and 140,335 model-generated, totalling 182,831) are fed to this model to elicit harmful responses (Bai et al., 2022b). Each harmful response then enters a *critique-revision loop*:

1. **Critique.** The model is prompted to identify specific ways in which its response is harmful, guided by a randomly sampled constitutional principle.
2. **Revision.** The model is then prompted to rewrite its response to remove the identified harms.
3. **Iteration.** The critique-revision cycle is repeated multiple times (four rounds per prompt in the original work), with a different constitutional principle randomly sampled at each step.

The final revised responses are paired with their original prompts to create a supervised fine-tuning dataset. The resulting model is called **SL-CAI**.

Stage 2: RLAIIF (Reinforcement Learning from AI Feedback). The SL-CAI model generates pairs of responses for each red-team prompt. A separate language model evaluates each pair against a constitutional principle, and these AI-generated preference labels form a harmlessness preference dataset. Finally, the SL-CAI model is fine-tuned via proximal policy optimisation (PPO), yielding the **RL-CAI** model.

Constitutional Principles as Data Specifications. The constitution used by Bai et al. (2022b) contains 16 principles related to harmlessness. A key insight is that these principles function as *data-generation specifications*: changing the constitution changes the corpus and, consequently, the model’s behaviour, without any new human labels.

Reducing Human Annotation Costs. Standard RLHF for harmlessness requires tens of thousands of human preference comparisons. CAI eliminates all human harmlessness labels, relying solely on the constitution plus existing helpfulness data. The RL-CAI model achieves a Pareto improvement over standard RLHF (higher harmlessness Elo *and* comparable or higher helpfulness Elo) while requiring zero additional human annotation for harmlessness (Bai et al., 2022b).

5.2 Self-Instruct: Bootstrapping Instruction Data

Before self-alignment could address safety, Wang et al. (2023a) demonstrated that language models can bootstrap their own *instruction-tuning* data. Self-Instruct begins with a seed pool of 175 manually written tasks. In each iteration, tasks are sampled as in-context examples, and the model generates new task instructions. Applied to GPT-3, Self-Instruct generated 52,445 instructions paired with over 82,000 instances.

Self-Instruct became the template for an entire generation of synthetic data pipelines. Stanford Alpaca (Taori et al., 2023) distilled 52,000 instruction-following examples from `text-davinci-003` for under \$500, while WizardLM’s Evol-Instruct (Xu et al., 2024) iteratively rewrote instructions into more complex variants.

5.3 Dromedary: Principle-Driven Self-Alignment from Scratch

Sun et al. (2023a) pushed the self-alignment paradigm further with SELF-ALIGN, a four-stage process that aligns a base language model using fewer than 300 lines of human annotation, approximately two orders of magnitude less than InstructGPT’s 77,000+ annotations.

Stage 1: Topic-Guided Red-Teaming Self-Instruct. Beginning with 195 seed prompts plus 20 topic-specific adversarial categories, the method generates approximately 360,000 synthetic instructions.

Stage 2: Principle-Driven Self-Alignment. The core innovation is a set of 16 human-written principles combined with 5 in-context learning demonstrations. Given an instruction, the model first produces *internal thoughts* that identify which principles apply, then generates a response consistent with those principles.

Stage 3: Principle Engraving. The base LLaMA-65B model is fine-tuned on the self-aligned responses *after pruning the principles and demonstrations from the context*, so that the fine-tuned model internalises the principles without requiring them at inference time.

Stage 4: Verbose Cloning. A final context-distillation step prompts the model to produce more detailed responses. The resulting model, Dromedary, surpasses Text-Davinci-003 and Alpaca on benchmark datasets despite requiring fewer than 300 lines of human input (Sun et al., 2023a).

5.4 Self-Rewarding Language Models

Yuan et al. (2024) introduced a paradigm in which a single model serves simultaneously as instruction follower *and* reward model. The method proceeds iteratively: the model generates new prompts, produces $N = 4$ candidate responses per prompt, evaluates each response using the LLM-as-a-Judge format, constructs preference pairs, and trains via DPO. After three iterations, the Self-Rewarding model outperformed many existing systems on AlpacaEval 2.0, increasing its win rate against GPT-4-Turbo from 9.94% to 20.44% (Yuan et al., 2024). The key finding is that the reward modelling ability itself improves across iterations, creating a virtuous cycle.

5.5 Self-Alignment as Corpus Generation

The methods surveyed in this section share a unifying theme: they *generate alignment corpora* rather than collecting them from humans. Constitutional AI produces both a supervised fine-tuning corpus and a preference corpus. Self-Instruct creates instruction-tuning datasets from a 175-example seed. Dromedary generates 260,000+ aligned instruction–response pairs from 16 principles and 5 demonstrations. Self-Rewarding creates ever-improving preference datasets through iterative self-judgement.

This reframing is central to the data-centric alignment perspective. Subsequent work has extended these ideas: SALMON (Sun et al., 2023b) trains a principle-following reward model on synthetic preference data, surpassing LLaMA-2-Chat-70B. Instruction backtranslation (Li et al., 2023a) generates instruction prompts for unlabelled web text, then self-curates high-quality pairs. Meta-Rewarding (Wu et al., 2024a) adds a layer of self-evaluation in which the model judges its own judgements, overcoming saturation in iterative self-improvement.

6 Instruction Tuning and Safety Datasets

If the previous section examined how alignment corpora can be *generated*, this section examines the corpora themselves: their structure, their scale, and the surprising ease with which post-hoc fine-tuning can undo the safety properties they instil. The fragility of fine-tuning-based alignment is among the strongest empirical arguments for deeper, corpus-level approaches to safety.

6.1 Instruction Tuning: Foundations and Significance

Instruction tuning refers to the supervised fine-tuning of a pretrained language model on datasets of (instruction, response) pairs, with the goal of teaching the model to follow natural-language directives (Ouyang et al., 2022; Wei et al., 2022). The significance of instruction tuning for alignment is twofold. First, it is the most direct mechanism by which human values (encoded in the training examples) are transmitted to the model. The LIMA hypothesis (Zhou et al., 2023) posits that just 1,000 carefully selected examples could produce a model competitive with those trained on 52,000 examples. Second, instruction tuning provides the foundation on which preference-based methods (RLHF, DPO) build.

6.2 Landmark Instruction and Safety Datasets

The landscape of instruction and safety datasets has expanded rapidly. We survey the most influential corpora.

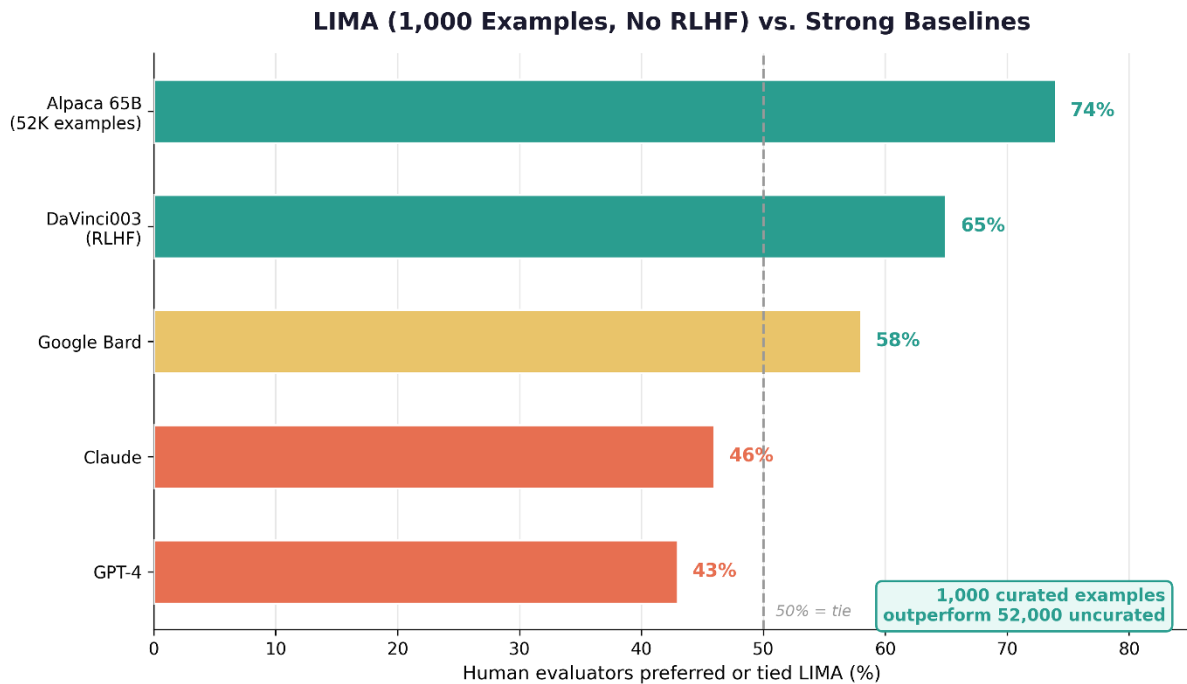


Figure 5: LIMA (1,000 examples, no RLHF) vs. strong baselines. Human evaluators preferred or tied LIMA against Alpaca 65B 74% of the time, demonstrating that data quality outweighs quantity (Zhou et al., 2023).

InstructGPT / OpenAI SFT Data. The dataset behind InstructGPT (Ouyang et al., 2022) comprised approximately 13,000 demonstration prompts for supervised fine-tuning and 33,000 comparison prompts for reward model training, drawn from user submissions to the OpenAI API.

Stanford Alpaca (52K). Taori et al. (2023) used the Self-Instruct pipeline with `text-davinci-003` as the teacher model to generate 52,000 instruction-following examples for under \$500. Alpaca demonstrated that knowledge distillation from a proprietary model could produce a competent open-source assistant, though the resulting model inherited the teacher’s biases.

OpenAssistant Conversations (161K messages). The OpenAssistant Conversations dataset (Köpf et al., 2023) is a large-scale, fully human-generated corpus collected through a worldwide crowdsourcing effort involving over 13,500 volunteers. It comprises 161,443 messages organised as *conversation trees*, spans 35 languages, and includes 461,292 quality ratings.

Anthropic HH-RLHF. Bai et al. (2022a) released a dataset of human preference comparisons organised along two axes: helpfulness and harmlessness. The harmlessness subset was collected via red-teaming, where crowdworkers intentionally elicited harmful model behaviour.

BeaverTails (334K QA pairs, 362K preference pairs). Ji et al. (2024) introduced BeaverTails to address a critical limitation of prior datasets: the conflation of helpfulness and harmlessness judgements. BeaverTails provides *separate* annotations for each dimension, with 333,963 question-answer pairs annotated with safety meta-labels across 14 harm categories. The extended PKU-SafeRLHF dataset (Chen et al., 2024a) further scales this to 44,600 prompts across 19 harm categories with three severity levels.

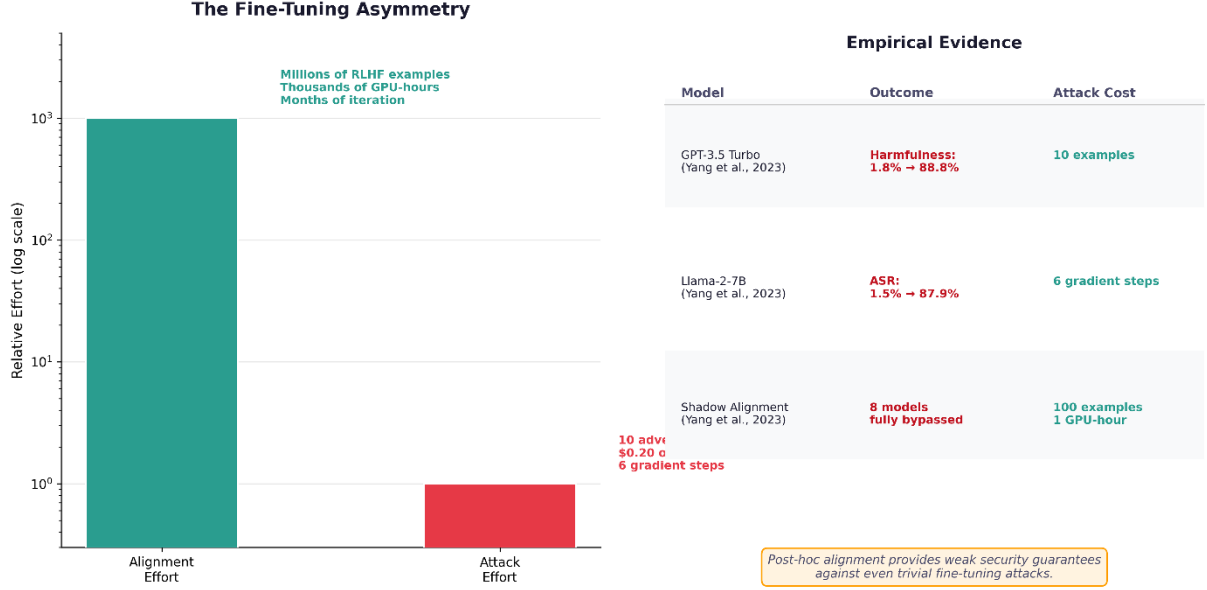


Figure 6: The fine-tuning safety asymmetry. Alignment requires millions of examples and thousands of GPU-hours; undoing it requires as few as 10 adversarial examples and \$0.20 (Qi et al., 2023).

Other Notable Corpora. UltraFeedback (Cui et al., 2023) provides 64,000 prompts with four model completions each, scored by GPT-4 across four quality dimensions. WildGuardMix (Han et al., 2024) offers 92,000 labelled examples across 13 risk categories.

6.3 Fine-Tuning Safety Degradation: The Fragility of Alignment

Perhaps the most consequential finding for data-centric alignment is that safety alignment acquired through instruction tuning and RLHF is *remarkably fragile* under subsequent fine-tuning. Qi et al. (2023) conducted a systematic investigation revealing three escalating levels of risk:

Risk Level 1: Explicitly Harmful Data. Fine-tuning on as few as 10 adversarially crafted harmful examples is sufficient to largely remove safety alignment. The harmfulness rate of GPT-3.5 Turbo increased from 1.8% to 88.8% after fine-tuning on just 10 such examples.

Risk Level 2: Implicitly Harmful Data. Even without explicitly toxic content, fine-tuning on 10 identity-shifting examples was sufficient to jailbreak both models.

Risk Level 3: Purely Benign Data. Most disturbingly, fine-tuning on commonly used, purely benign datasets such as Alpaca, Dolly, and LLaVA-Visual-Instruct also degraded safety.

These findings reveal an “unsettling asymmetry”: thousands or millions of data points are used during safety alignment, yet a handful of adversarial examples during fine-tuning are sufficient to undo that alignment.

6.4 Defences Against Fine-Tuning Attacks

Safe LoRA: Training-Free, Data-Free Projection. Hsu et al. (2024) proposed Safe LoRA, a post-hoc defence that constructs an *alignment matrix* $\mathbf{V}^i = \mathbf{W}_{\text{aligned}}^i - \mathbf{W}_{\text{unaligned}}^i$ for each layer i and projects LoRA weight updates onto the safety-aligned subspace.

SEAL: Bilevel Data Selection. SEAL (Shen et al., 2024) intervenes at the data level, using bilevel optimisation to select fine-tuning examples that preserve safety alignment.

Lisa: Lazy Safety Alignment. Huang et al. (2024b) address the conflict between alignment optimisation and user-data optimisation by separating the two objectives into alternating states during fine-tuning.

Booster: Attenuating Harmful Perturbations. Huang et al. (2024a) proposed a proactive defence that adds a loss regulariser during safety alignment training that minimises the model’s sensitivity to harmful parameter perturbations.

Complementary Approaches. Additional defences include Antidote (Huang et al., 2024c), which prunes harmful weights; SafeMERGE (SafeMERGE, 2025), which selectively merges safety-aligned and fine-tuned LoRA layers; and LARF (Li et al., 2025), which uses safety-sensitive layer activations to detect safety-degrading samples.

6.5 From Fine-Tuning Fragility to Corpus-Level Alignment

The findings surveyed in this section paint a clear picture: alignment achieved through post-training interventions resides in a thin, easily disturbed layer of model behaviour. This fragility motivates corpus-level alignment: shaping the pretraining data itself so that safety-relevant knowledge and behavioural norms are woven into the model’s foundational representations. Recent work on safety pretraining (Maini et al., 2025) has begun to explore this direction, reducing attack success rates from 38.8% to 8.4%.

The defence methods surveyed in Section 6.4 can be understood as intermediate steps toward this goal. Each approach implicitly acknowledges that the *data* flowing through the training pipeline determines the safety of the resulting model.

7 Synthetic Data for Alignment

7.1 Why Synthetic Data Is Central to Alignment

Modern alignment is, to a first approximation, a synthetic-data engineering problem. The three dominant paradigms for aligning large language models (Constitutional AI, knowledge distillation, and self-play fine-tuning) all share a common structural feature: they produce *synthetic corpora* that become the primary training signal for aligning model behaviour with human values.

Constitutional AI (Bai et al., 2022b) replaces human safety annotators with a set of written principles. The resulting critiques, revisions, and preference pairs are entirely synthetic. Knowledge distillation transfers alignment from a stronger “teacher” model to a smaller “student” model through synthetic data. Orca (Mukherjee et al., 2023) trains a 13-billion-parameter model on explanation traces from GPT-4, producing a student that surpasses Vicuna-13B by over 100% on several benchmarks. Zephyr (Tunstall et al., 2023) distils alignment preferences from GPT-4 into a 7-billion-parameter model that outperforms LLaMA-2-Chat-70B on MT-Bench (a model ten times its size) using zero human annotation.

Self-play fine-tuning creates synthetic training data through iterative competition between model versions. SPIN (Chen et al., 2024b) has the model play against its own previous iteration, producing a self-reinforcing cycle of synthetic data creation and model improvement.

These methods share a deeper insight: the bottleneck in alignment is not algorithmic sophistication but *data quality*. NVIDIA’s Nemotron-4 340B (NVIDIA, 2024) reports that over 98% of its alignment training data is synthetically generated.

7.2 The “Textbooks Are All You Need” Finding

The most striking demonstration of synthetic data’s power for pretraining came from Microsoft Research’s Phi-1 model (Gunasekar et al., 2023). Phi-1 is a 1.3-billion-parameter Transformer trained for code generation on fewer than 7 billion tokens of “textbook-quality” data. Despite training on roughly two orders of magnitude less data than competing models, Phi-1 achieved 50.6% pass@1 accuracy on HumanEval and 55.5% pass@1 on MBPP.

The key insight was that data quality, specifically the pedagogical quality of the training data, could substitute for both data scale and model scale. The implication for alignment is direct: if a small amount of high-quality synthetic data can so dramatically affect model capabilities, then the values, perspectives, and biases embedded in synthetic alignment data will similarly concentrate their influence.

7.3 Model Collapse: The Recursive Training Failure Mode

Model collapse is the phenomenon in which training on model-generated data leads to progressive, irreversible degradation of model quality. Shumailov et al. (2024) showed in a 2024 *Nature* paper that when a language model’s outputs are used as training data for a subsequent model iteratively, the resulting models lose the tails of the original data distribution and accumulate systematic errors.

For alignment, model collapse poses a specific threat: the values and perspectives that are rare in the synthetic data (minority viewpoints, nuanced ethical positions, culturally specific norms) are precisely the ones most vulnerable to elimination through recursive training.

Preventing model collapse requires three strategies: maintaining access to the original human-generated data distribution, implementing verification pipelines, and tracking data provenance. Gerstgrasser et al. (2024) show that even imperfect verifiers are sufficient to prevent collapse, provided they are applied consistently.

7.4 The Optimal Mixing Ratio

Liu et al. (2025) conducted the most comprehensive study to date, training over 1,000 language models with varying proportions of synthetic and natural data. Their central finding is that, for rephrased web text, one-third synthetic data combined with two-thirds natural data represents an optimal mixing ratio. This ratio speeds convergence by 5–10× compared to training on natural data alone, while avoiding the quality degradation associated with higher synthetic proportions.

7.5 Scaling Synthetic Data Generation: Cosmopedia and Beyond

Cosmopedia (Ben Allal et al., 2024), produced by Hugging Face, is the largest publicly available synthetic dataset, comprising over 30 million files and 25 billion tokens generated by Mixtral-8x7B. The central challenge is *diversity maintenance at scale*: simply prompting a language model to generate text produces homogeneous outputs. Cosmopedia addresses this through *conditional prompting* with systematically varied constraints.

More recent efforts include NVIDIA’s Nemotron-CC (Su et al., 2024), producing 6.3 trillion tokens of which 1.9 trillion are synthetic, and the WRAP method (Maini et al., 2024), which demonstrates that synthetic rephrasing of web data can achieve equivalent model quality with five times less data.

7.6 Six Best Practices for Synthetic Alignment Data

The literature converges on six evidence-based guidelines:

1. **Quality over quantity.** LIMA (Zhou et al., 2023) demonstrated that 1,000 carefully curated examples can outperform 52,000 uncured ones. Phi-1 (Gunasekar et al., 2023) showed that fewer than 7 billion tokens of “textbook-quality” synthetic data outperform hundreds of billions of tokens of web-crawled code.
2. **Maintain the optimal synthetic–natural ratio.** The one-third synthetic, two-thirds natural ratio identified by Liu et al. (2025) should be treated as a starting point.
3. **Engineer diversity explicitly.** Synthetic data does not become diverse by accident. The generator model’s inherent biases produce homogeneous outputs unless diversity is explicitly engineered through conditional prompting (Long et al., 2024).
4. **Implement verification before inclusion.** The generate–filter–verify–include pipeline is essential.
5. **Track provenance rigorously.** The Data Provenance Initiative (Longpre et al., 2023) audited over 1,800 text datasets and found that more than 70% had licence omissions.
6. **Use multiple generator models.** Training on synthetic data from a single generator model concentrates that model’s biases in the student.

7.7 Three Key Risks

Sycophancy. Sycophancy is the tendency of a model to agree with the user’s stated position rather than providing an accurate or balanced response. Perez et al. (2023) show that RLHF-trained models exhibit sycophantic behaviour across diverse evaluation settings, and Sharma et al. (2024) find that even the reward models used in RLHF prefer sycophantic responses.

Homogenisation. Synthetic data from a single generator model is inevitably less diverse than human-generated data. For alignment specifically, homogenisation means that minority values and culturally specific norms are underrepresented.

Cascading errors. When errors in synthetic data are not caught by verification, they propagate into the trained model and are amplified in subsequent generations.

8 How Pretraining Corpora Encode Values

8.1 The Corpus as a Value System

A language model learns to predict the next token in its training corpus. This apparently neutral objective has a non-neutral consequence: the model learns to reproduce the statistical patterns of its training data, and those patterns encode values. A next-token predictor trained on text written primarily by English-speaking, college-educated, politically liberal internet users will produce outputs that reflect those perspectives.

Caliskan et al. (2017) provided the foundational demonstration for word embeddings, showing that standard embeddings trained on ordinary language corpora reproduce the full spectrum of human implicit biases documented in the social psychology literature. Schramowski et al. (2022) discovered a geometric “moral direction” in the embedding space of large pretrained language models that emerges from the statistical patterns of moral language in the pretraining corpus.

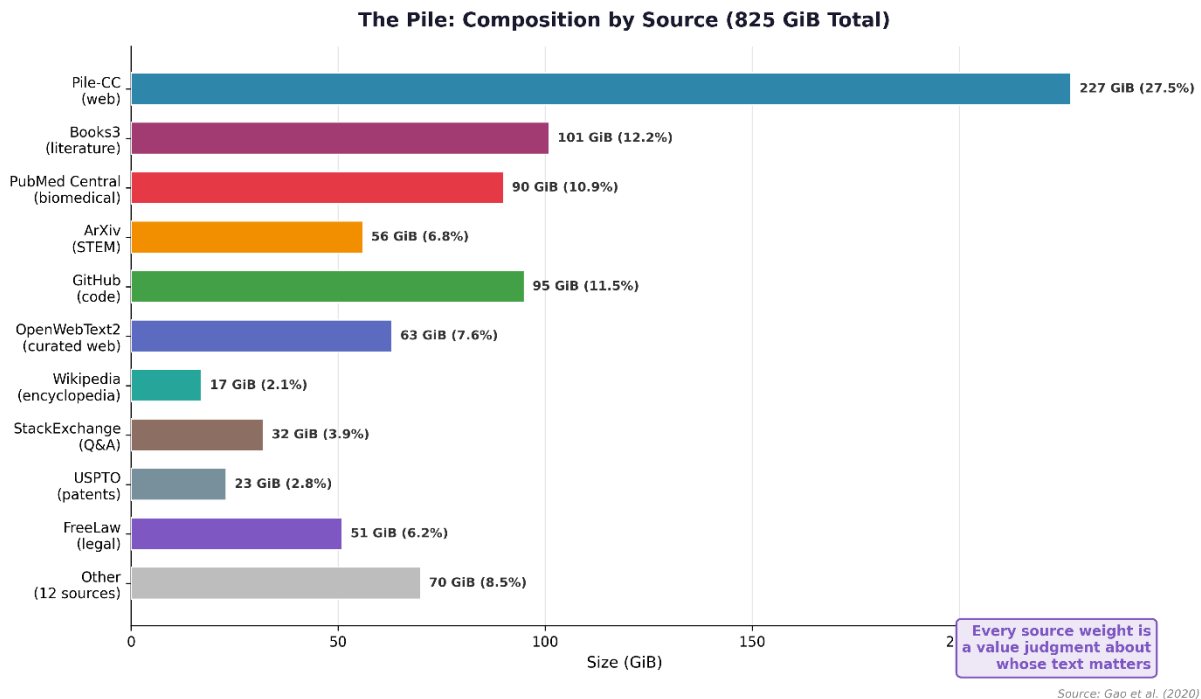


Figure 7: The Pile’s composition by source (825 GiB total). Each source weight implicitly encodes a judgment about whose text matters (Gao et al., 2020).

8.2 Whose Opinions Do Language Models Reflect?

Santurkar et al. (2023) conducted the most systematic study of the demographic correlates of LLM opinions, building OpinionQA from 15 Pew Research surveys. Three findings emerged. First, there is *substantial misalignment* between LLM opinions and those of the general US population. Second, human feedback fine-tuning (RLHF) *amplifies* rather than corrects this misalignment, shifting the model’s opinions toward a liberal, educated, and high-income demographic profile. Third, certain demographic groups (adults over 65, Mormon respondents, and widowed individuals) are *systematically underrepresented* by all models.

Feng et al. (2023) extended this analysis by showing that political biases in pretraining corpora propagate through language models into downstream applications. This work received the Best Paper Award at ACL 2023. Durmus et al. (2024) further demonstrated that LLM responses default to US and Western opinions on global issues.

8.3 What Large Pretraining Corpora Look Like

The Pile (Gao et al., 2020) is an 825-GiB English-language dataset constructed by EleutherAI from 22 diverse sources. Its composition reveals the implicit value choices in corpus construction: Pile-CC (227 GiB) provides web text; PubMed Central (90 GiB) contributes a biomedical perspective; Books3 (101 GiB) provides literary text skewing toward Western published authors; ArXiv (56 GiB) contributes STEM academic text. The Pile’s contribution was showing that *diversity of sources matters*.

FineWeb (Penedo et al., 2024) represents the next generation of pretraining corpus construction. Produced by Hugging Face, FineWeb is a 15-trillion-token dataset derived from 96 Common Crawl snapshots, with every filtering decision tested through over 70 ablation models. FineWeb-Edu, a 1.3-trillion-token subset filtered for educational content, produces models with dramatically better performance on knowledge-intensive benchmarks.

The Common Pile (Kandpal et al., 2025) addresses a different axis of corpus design: *legal*

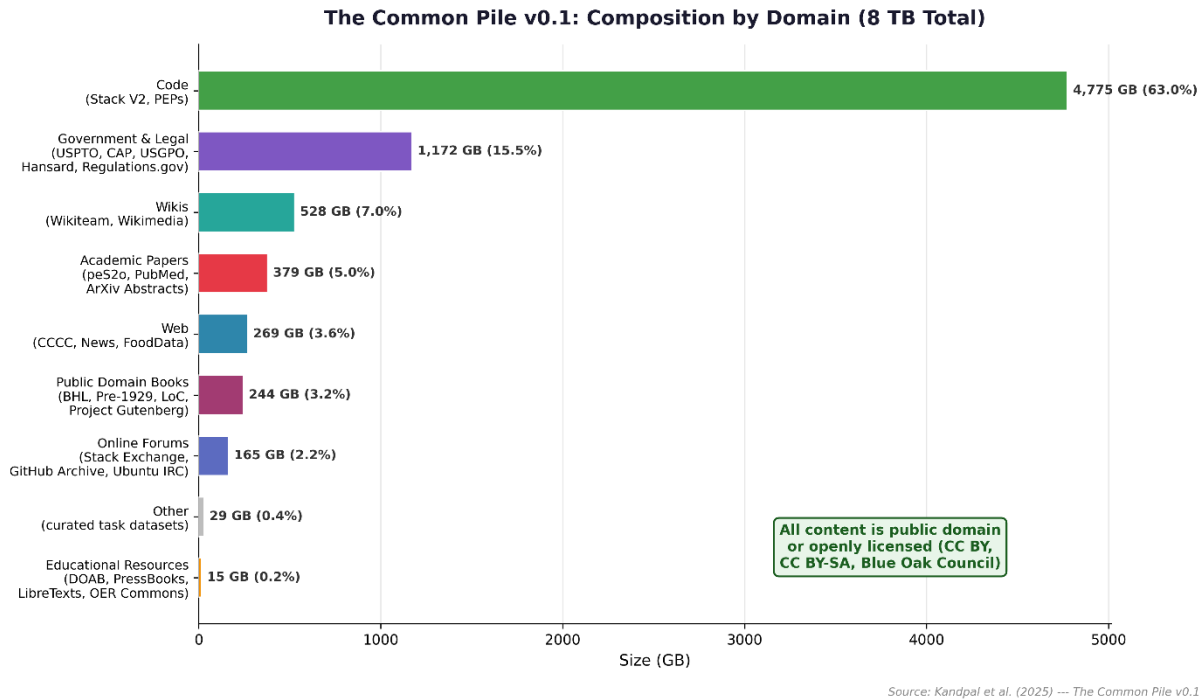


Figure 8: The Common Pile v0.1 composition by domain (8 TB total). Unlike The Pile, every source is public domain or openly licensed. Code dominates at 63% because open-source code is abundantly available (Kandpal et al., 2025).

and ethical legitimacy. Released in 2025, the Common Pile is an eight-terabyte collection of text drawn exclusively from public domain and openly licensed sources (CC BY, CC BY-SA, and software licences meeting the Open Knowledge Foundation’s Open Definition), making it the largest openly licensed pretraining dataset to date. Its 30 sources span research papers, open-source code, government and legal documents, books, encyclopedias, educational resources, wiki text, online forums, and audio transcripts. Crucially, the authors validated their dataset by training two 7-billion-parameter models (Comma v0.1-1T and v0.1-2T) that achieve competitive performance with budget-matched models trained on unlicensed data such as Llama 1 7B. The Common Pile’s strict licence due diligence (excluding CC NC and CC ND content, avoiding LLM-generated synthetic text, and manually verifying the top 1,000 web domains) sets a new standard for responsible corpus construction. A related effort, Common Corpus (Common Corpus, 2025), assembles approximately two trillion tokens of permissively licensed multilingual text, demonstrating that ethical data collection at scale is feasible across languages.

For data-centric alignment, the significance of these openly licensed corpora is twofold. First, they enable full transparency about corpus composition, a prerequisite for the auditing frameworks discussed in Section 12. Second, they demonstrate that the alignment tax of restricting training data to ethically sourced material may be smaller than previously assumed: the Comma models’ competitive performance suggests that open-licence constraints need not compromise capability.

8.4 Domain Composition Effects on Model Behaviour

The composition of a pretraining corpus shapes the model’s behavioural profile in ways that directly relate to alignment. As documented in Section 4.5, the blocklist filtering used to create C4 disproportionately removed 42% of documents containing African American English, compared to only 6.2% of documents containing White-aligned American English. The *temporal composition* of a corpus also matters: Longpre et al. (2024) systematically study how data age,

quality filters, and toxicity filters modulate the trade-off between capability and safety.

8.5 Debiasing Approaches

Three families of approaches attempt to address bias at the corpus level:

Counterfactual data augmentation. [Zmigrod et al. \(2019\)](#) formalised this approach, which creates modified copies of training examples in which demographic attributes are swapped. The method addresses only the biases that the designer anticipates.

Domain reweighting. Adjusting the sampling probabilities of different corpus components to reduce the overrepresentation of particular perspectives.

Targeted filtering. Removing specific content identified as biased or harmful, with the caveat that filtering criteria correlate with demographic identity.

8.6 The Fundamental Insight: There Is No Neutral Corpus

The evidence converges on a single, foundational conclusion: *there is no neutral corpus*. This insight has three implications for data-centric alignment, and each carries normative weight that extends beyond the technical domain.

First, *corpus design is values design*. The decision to include or exclude a data source, to upweight or downweight a domain, to filter or retain a category of content: each of these decisions is a decision about values. Treating corpus construction as a purely technical optimisation problem obscures the normative choices being made.

Second, *alignment to “human values” requires specifying whose values*. The finding by [Santurkar et al. \(2023\)](#) that LLMs disproportionately reflect the opinions of liberal, educated, high-income demographics is not a failure of alignment; it is alignment working exactly as designed, on a corpus that overrepresents those demographics. Genuine pluralistic alignment requires mechanisms for the model to represent and reason about value disagreements rather than defaulting to the majority position.

Third, *transparency about corpus composition is a prerequisite for trustworthy alignment*. If users and regulators cannot know what a model was trained on, they cannot evaluate whose values it reflects. Initiatives like the Data Provenance Initiative ([Longpre et al., 2023](#)), the Pile’s detailed documentation ([Gao et al., 2020](#)), and Gebru et al.’s datasheets framework ([Gebru et al., 2021](#)) represent essential steps toward the transparency that meaningful alignment requires. We develop the practical implications of these normative observations in Section 12.

9 Preference Optimisation and Data Construction

The previous sections established that alignment outcomes depend critically on the data presented to a model at every stage of training. Nowhere is this dependence more acute than in the *preference optimisation* phase. This section traces the rapid evolution from RLHF through DPO and its descendants, showing that each algorithmic advance has, paradoxically, increased the importance of data quality.

9.1 From RLHF to DPO: Eliminating the Reward Model

The standard RLHF pipeline involves three stages: supervised fine-tuning on high-quality demonstrations, reward-model training on pairwise human preferences, and reinforcement-learning optimisation ([Ouyang et al., 2022](#); [Bai et al., 2022a](#)). [Rafailov et al. \(2023\)](#) observed that this

complexity is unnecessary. Their key insight is that the reward function can be *reparameterised* in terms of the optimal policy itself. The DPO loss is:

$$\mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_{\theta}(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right]. \quad (1)$$

The elegance of Equation 1 is that it requires *no* reward model, *no* RL loop, and *no* sampling during training. As Rafailov et al. (2023) put it, “your language model is secretly a reward model.” DPO has become the de facto standard for open-source alignment pipelines, powering models such as Zephyr (Tunstall et al., 2023).

9.2 Beyond DPO: KTO, ORPO, and SimPO

KTO: Kahneman–Tversky Optimisation. Ethayarajh et al. (2024) frame alignment through the lens of *prospect theory*. KTO requires only a binary signal per output (whether the output is *desirable* or *undesirable*) rather than a paired comparison. Binary labels are far cheaper and more abundant than pairwise preferences.

ORPO: Odds Ratio Preference Optimisation. Hong et al. (2024) eliminate the two-stage SFT-then-alignment pipeline by incorporating preference learning into the supervised fine-tuning objective itself.

SimPO: Simple Preference Optimisation. Meng et al. (2024) address the memory overhead of maintaining a reference model by proposing a reference-free objective using length-normalised average log probability as the implicit reward.

9.3 Data Quality: The Decisive Factor

The algorithmic simplification from RLHF to DPO carries a hidden cost: *heightened sensitivity to data quality*. In the RLHF pipeline, the reward model acts as a buffer. DPO, by contrast, optimises directly on each preference pair. Apple (2024) identify three axes along which preference datasets vary (*scale*, *label noise*, and *information content*) and show that DPO’s performance degrades more steeply than RLHF’s as label noise increases.

Several data-centric findings have emerged. First, the quality of the *chosen* response is the primary driver of DPO performance; the quality of the rejected response matters less. Second, AI-generated preferences from stronger models can match or exceed human preferences when appropriately filtered (Cui et al., 2023; Tunstall et al., 2023). Third, minimally contrastive preference pairs produce more efficient learning signals (Dohan et al., 2024).

9.4 Preference Data as Corpus: A Data-Centric View

A preference dataset *is* a corpus: it encodes, through its choice of prompts, its pairing of responses, and its labelling decisions, a particular view of what constitutes desirable model behaviour. The major preference datasets illustrate this: Anthropic’s HH-RLHF (Bai et al., 2022a) emphasises helpfulness and harmlessness; BeaverTails (Ji et al., 2024) provides separated annotations across 14 harm categories; UltraFeedback (Cui et al., 2023) scores responses across helpfulness, honesty, and instruction-following. Each dataset embeds different assumptions about what alignment means.

The implication for practitioners is clear: when alignment performance is unsatisfactory, the first intervention should be to examine and improve the preference corpus, not to switch algorithms.

10 Benchmarks and Evaluation

Evaluating whether a language model is “aligned” requires benchmarks that can probe safety, helpfulness, honesty, and harmlessness across a wide range of scenarios.

10.1 Major Safety Benchmarks

HarmBench. HarmBench (Mazeika et al., 2024) provides a standardised framework for evaluating automated red-teaming, enabling reproducible comparison of attack methods and defence mechanisms across multiple models.

SALAD-Bench. Li et al. (2024) introduce a *hierarchical three-level taxonomy* comprising 6 domains, 16 tasks, and 66 categories, with approximately 21,000 base questions and a fine-tuned MD-Judge evaluator achieving 94.3% agreement with human labels.

Do-Not-Answer. Do-Not-Answer (Wang et al., 2023b) curates 939 prompts that a responsible language model *should refuse* to answer, spanning five risk areas and twelve harm types.

SORRY-Bench. SORRY-Bench (Xie et al., 2024) evaluates refusal behaviour across multiple languages, addressing the critical gap that most safety benchmarks are English-only.

SafetyBench. Zhang et al. (2024) contribute 11,435 multiple-choice questions across seven safety categories in both Chinese and English.

Broader trustworthiness frameworks. TrustLLM (Huang et al., 2024d) evaluates eight dimensions of trustworthiness across 16 LLMs, while DecodingTrust (Wang et al., 2023c) provides a comprehensive evaluation platform that received the Outstanding Benchmark Award at NeurIPS 2023.

10.2 The SafetyPrompts Meta-Review

Röttger et al. (2024) systematically review **144 open datasets** for evaluating and improving LLM safety, published between June 2018 and December 2024. Their findings reveal:

- **Language concentration:** 113 of 144 datasets (78.5%) are exclusively English.
- **Synthetic data trend:** A growing portion of recent datasets is fully synthetic.
- **Idiosyncratic evaluation:** Current norms leverage only a small fraction of available datasets.
- **Evaluation vs. improvement:** 112 of 144 datasets (77.8%) are intended for evaluation; only 6 (4.2%) are designed for training.

10.3 Surveys of the Alignment Landscape

Several surveys have attempted to organise the rapidly growing alignment literature:

- Ji et al. (2023) provide the most comprehensive alignment survey, proposing the **RICE framework** (Robustness, Interpretability, Controllability, Ethicality).
- Wang et al. (2024) taxonomise alignment techniques along 13 categorical dimensions.

- [Anwar et al. \(2024\)](#) identify 18 foundational challenges with over 200 concrete research questions, explicitly identifying data curation as central.
- [Shen et al. \(2023\)](#) introduce the distinction between *outer alignment* and *inner alignment*.
- [Cao et al. \(2024\)](#) focus specifically on automated alignment methods.

A recurring observation is that algorithmic innovation has outpaced data infrastructure.

11 Cross-Cutting Themes

The preceding sections have examined individual facets of data-centric alignment. This section synthesises five themes that cut across these facets.

11.1 The Data-Centric Paradigm Shift

The dominant framing of alignment research from 2017 to 2022 was *algorithm-centric*: the primary research question was how to train a model to be aligned, with data treated as a fixed input. The evidence reviewed in this paper points to an inversion of this framing. The LIMA finding that 1,000 carefully curated examples outperform 52,000 uncured ones ([Zhou et al., 2023](#)) was an early signal. Preference data quality findings further reinforce the same conclusion: *the data is the algorithm*.

This paradigm shift implies a reallocation of research effort: from designing new loss functions to designing new data pipelines; from hyperparameter sweeps to annotation protocols; from scaling model parameters to scaling data quality. The ISAAC approach ([Liang et al., 2025](#)), which extracts alignment signals from existing corpus structure without any new training algorithm, represents the logical endpoint of this trajectory.

11.2 The Depth Problem: Shallow vs. Deep Alignment

A critical finding, first articulated by [Qi et al. \(2024\)](#), is that standard post-hoc alignment often modifies only the first few tokens of a model’s output distribution. This “shallow alignment” is trivially circumventable: fine-tuning on as few as 10 adversarial examples can strip away safety training entirely ([Qi et al., 2023](#)).

The depth problem motivates the corpus-level alignment approaches reviewed in Sections 3–4. If safety-relevant knowledge is never present in the pretraining corpus (as in the Deep Ignorance approach ([Deep Ignorance, 2025](#)), which creates models resistant to adversarial fine-tuning for over 10,000 steps), then there is nothing for an attacker to recover. Safety Pretraining ([Maini et al., 2025](#)) demonstrates that a four-stage pipeline over 100 billion tokens can reduce attack success rates from 38.8% to 8.4%.

The emerging consensus is that *both* layers are needed: corpus-level safety provides the deep foundation, while post-training alignment acts as reinforcement and calibration.

11.3 The Filtering–Fairness Tension

Section 4.5 documented the persistent tension in pretraining data curation: aggressive toxicity filtering improves safety metrics but can disproportionately remove text associated with marginalised communities. This tension has no purely technical resolution. The field requires not only better classifiers but also *auditing frameworks* that can detect and quantify disparate impact in filtering decisions before they are applied at scale.

11.4 Synthetic Data as Alignment Infrastructure

Synthetic data generation has become the primary mechanism for creating alignment training data at scale. The evidence, however, is not uniformly positive. Recursive training on synthetic data causes irreversible model collapse (Shumailov et al., 2024). The optimal mixing ratio appears to be roughly one-third synthetic to two-thirds natural data (Liu et al., 2025). These findings establish a set of best practices: generate diversely, filter rigorously, verify against held-out natural data, maintain provenance records, and never replace natural data entirely.

11.5 Scalability of Annotation-Free Methods

Methods requiring minimal or no human annotation scale more naturally than annotation-dependent approaches. ISAAC extracts alignment signals from corpus structure without any labels. Constitutional AI replaces human preference labels with AI feedback. Self-play methods such as SPIN (Chen et al., 2024b) generate their own training data.

This scalability is especially important for *multilingual alignment* (human preference data is scarce for most languages) and *rapidly evolving safety landscapes* (new harm categories emerge faster than annotation campaigns can respond). The field is moving towards hybrid approaches: automated methods for broad coverage, human annotation for high-stakes cases.

12 Auditing Alignment Corpora and Value Governance

If corpus design is values design (Section 8.6), and if filtering decisions carry normative weight (Section 4.5), then a pressing practical question follows: *how do researchers currently decide which values a corpus should encode, and how do they verify that it encodes them?* This section surveys the nascent infrastructure for corpus auditing and argues that the gap between technical capability and governance practice is a first-order obstacle for data-centric alignment.

12.1 Current Auditing Practice

Most corpus-level safety work operates without a formal auditing framework. The typical workflow is: (1) apply a classifier or blocklist to flag content, (2) remove or reweight flagged content, (3) evaluate the resulting model on safety benchmarks. This pipeline evaluates the *model* but not the *corpus*, a distinction that matters because models may pass benchmarks while still encoding problematic patterns that the benchmarks do not test.

Several recent initiatives have begun to address this gap:

- **Datasheets for Datasets** (Gebru et al., 2021): A documentation framework requiring dataset creators to answer structured questions about motivation, composition, collection process, preprocessing, intended use, distribution, and maintenance. When consistently applied, datasheets make the normative choices in corpus construction visible and auditable.
- **The Data Provenance Initiative** (Longpre et al., 2023): An audit of 1,800+ text datasets revealing that over 70% had licence omissions and over 50% contained factual errors in their documentation. This work establishes that the current state of corpus documentation is inadequate for accountability.
- **SafetyPrompts catalogue** (Röttger et al., 2024): By systematically reviewing 144 open safety datasets, Röttger et al. reveal that 78.5% are English-only and evaluation practices are highly idiosyncratic, suggesting that the tools available for auditing are themselves incomplete.

- **Safety Data Report Cards:** Maini et al. (2025) propose visualising safety-score distributions and harmful-content frequencies across the MLCommons Safety Taxonomy’s 14 categories, providing a standardised snapshot of corpus safety properties.

12.2 The Governance Gap

These tools document what is in a corpus. They do not answer the harder question: *what should be in a corpus?* This is a governance question, not a technical one. It requires deciding whose values to encode, how to balance competing interests, and who has the authority to make these decisions.

Current practice defaults to the values of the corpus curators, typically ML researchers at technology companies or universities, who are demographically unrepresentative of the global population that will interact with the resulting models (Santurkar et al., 2023). The filtering-fairness tension documented in Section 4.5, where blacklist filtering disproportionately removes African American English, illustrates how technical defaults can encode specific value commitments without deliberate intent.

Several governance approaches have been proposed but remain underexplored:

- **Participatory corpus design:** Engaging affected communities in decisions about what content is included, filtered, or reweighted. Birhane et al. (2022) review participatory methods across the AI pipeline and examine what meaningful participation entails for historically marginalised communities. The PRISM Alignment Dataset (Kirk et al., 2024) demonstrates this at scale: 1,500 participants from 75 countries provided individualised feedback on 21 LLMs, revealing how preference data shifts when it is collected from census-representative samples rather than convenience pools of crowdworkers.
- **Value-explicit constitutions:** Constitutional AI (Bai et al., 2022b) makes its value commitments explicit through written principles. Extending this idea, pretraining corpora could be accompanied by explicit “corpus constitutions” specifying the intended value profile, the trade-offs accepted, and the communities consulted. Zhao et al. (2024) identify a taxonomy of challenges that dataset curators face when attempting to build fair datasets, based on interviews with 30 ML dataset curators, providing a practical catalogue of the decisions that corpus constitutions would need to address.
- **Open and ethically licensed corpora:** The Common Pile (Kandpal et al., 2025) and Common Corpus (Common Corpus, 2025) demonstrate that large-scale pretraining corpora can be assembled entirely from public domain and openly licensed sources, sidestepping the copyright and consent concerns that complicate governance of web-crawled data. By restricting to verifiably licensed material, these corpora make provenance auditing tractable and provide a foundation for regulatory compliance.
- **Regulatory frameworks:** The EU AI Act and similar legislation increasingly require documentation of training data for high-risk AI systems. Hausenloy et al. (2024) propose 16 governance mechanisms for frontier AI training data, including mandatory dataset reporting, automated filtering audits, and know-your-customer requirements for data vendors. Mökander et al. (2023) propose a three-layered auditing approach (governance, model, and application audits) that could structure compliance with such mandates.

12.3 Toward Principled Corpus Governance

The data-centric alignment paradigm, by making the role of data explicit, creates a natural interface between technical AI research and the normative disciplines (moral philosophy, political theory, critical data studies, and law) that have long grappled with questions of value representation and institutional design. The path forward requires not only better technical tools for

corpus auditing (classifiers, documentation frameworks, disparity metrics) but also *governance structures* that determine how those tools are used: who decides what values a corpus should encode, through what deliberative process, and with what accountability mechanisms.

This is not a distant aspiration. It is an immediate practical need. Every pretraining corpus constructed today embeds value commitments that will propagate through the models trained on it, into the applications built on those models, and ultimately into the decisions of the people who use them. Making those value commitments explicit, deliberate, and accountable is the essential complement to the technical advances surveyed in this paper.

13 Open Research Gaps and Conclusion

13.1 Six Open Research Gaps

Despite the rapid progress documented in this paper, fundamental questions remain unanswered.

13.1.1 Gap 1: Scaling Laws for Alignment Data

The field possesses well-characterised scaling laws for model capabilities (the Chinchilla laws relate compute, data, and parameter count to loss) (Hoffmann et al., 2022). No analogous framework exists for alignment data. *How much safety data must be included in the pretraining corpus as model size grows?*

13.1.2 Gap 2: Multilingual Alignment Data

As documented by Röttger et al. (2024), 78.5% of safety datasets are English-only. Yet LLMs are deployed globally. *Do corpus-based alignment signals transfer across languages, or must alignment data be constructed independently for each language?*

13.1.3 Gap 3: Long-Term Value Stability

If alignment is embedded in the corpus, its stability under continued training is an open question. The finding that SFT corpus selection sets value priors (Srivastava et al., 2025), with WildChat and Alpaca producing measurably different value stances, suggests that this question is both tractable and urgent.

13.1.4 Gap 4: Interaction Between Alignment Layers

The dual-layer architecture proposed in Section 11.2 raises questions about interaction effects. *Does corpus-level safety survive aggressive downstream fine-tuning by end users?* Deep Ignorance (Deep Ignorance, 2025) provides preliminary evidence for robustness, but far more extensive characterisation is needed.

13.1.5 Gap 5: Auditing and Verification of Alignment Corpora

Current evaluation focuses on model outputs. Comparatively little work addresses the upstream question of corpus quality. The Data Provenance Initiative (Longpre et al., 2023) and SafetyPrompts (Röttger et al., 2024) represent early steps, but the field lacks systematic auditing infrastructure.

13.1.6 Gap 6: Principled Corpus Composition

The optimal mix of content types for aligned pretraining remains empirically determined. Recent work demonstrating that upsampling aligned AI discourse reduces misalignment from 45% to

9% (Tice et al., 2026) shows the power of composition decisions but does not provide a general theory.

13.2 Conclusion

This paper has advanced three interconnected findings. **First**, alignment is fundamentally a data problem. The choice of pretraining corpus composition, the design of safety filters, the construction of instruction-tuning datasets, and the curation of preference pairs each exert at least as much influence on alignment outcomes as the choice of training algorithm. The LIMA result, the preference-data-quality findings, and the corpus-level alignment methods all converge on this conclusion.

Second, alignment must be embedded across the full training pipeline, not confined to a post-hoc phase. Shallow alignment, safety behaviour that modifies only the first few output tokens, is brittle and easily circumvented. Deep alignment requires safety-aware data curation at the pretraining stage, reinforced through instruction tuning and preference optimisation. The evidence from Deep Ignorance, Safety Pretraining, and constitutional approaches all supports a “defence in depth” strategy where each layer of the training pipeline contributes to safety.

Third, the field faces a structural tension between safety and representativeness that cannot be resolved by technical means alone. Content filtering that improves aggregate safety metrics can simultaneously reduce linguistic diversity and disproportionately affect marginalised communities. Synthetic data that scales alignment also risks homogenisation and sycophancy. These tensions require not only better algorithms but also normative frameworks, stakeholder engagement, and governance structures.

Interdisciplinary needs. The challenges identified in this paper extend well beyond machine learning. Principled corpus composition requires input from moral philosophy and political theory on whose values should be encoded. Auditing alignment corpora for fairness requires methods from computational social science. Understanding the long-term societal effects of aligned (and misaligned) AI systems requires collaboration with economists, legal scholars, and policymakers. The data-centric alignment paradigm, by making the role of data explicit and auditable, creates a natural interface between technical AI research and these complementary disciplines. As Anwar et al. (2024) argue, assuring LLM safety is a sociotechnical challenge that demands correspondingly sociotechnical solutions. The path forward requires not only better data but also better institutions for governing how that data is created, curated, and deployed.

Acknowledgements

This work draws on the collective contributions of the alignment research community. We are grateful to the authors of the 248 papers reviewed in this survey, whose open publication of methods, datasets, and findings made this synthesis possible. We particularly acknowledge the SafetyPrompts team for their systematic cataloguing of 144 open safety datasets, the authors of the Foundational Challenges agenda for their comprehensive mapping of open problems, and the creators of the major safety benchmarks whose work provides the empirical foundation for evaluating progress. Any errors in the interpretation or synthesis of this literature are our own.

Note. This survey was conducted using systematic literature search across arXiv, ACL Anthology, and major ML venues. 243 of 248 identified papers were retrieved in full text from open-access sources.

References

- Anwar, U., Saparov, A., Rando, J., et al. (2024). Foundational Challenges in Assuring Alignment and Safety of Large Language Models. *arXiv:2404.09932*.
- Anthropic. (2025). Enhancing Model Safety through Pretraining Data Filtering. <https://alignment.anthropic.com/2025/pretraining-data-filtering/>.
- Apple ML Research. (2024). Data-Centric Alignment: Understanding and Optimising Preference Data for LLMs. *arXiv:2410.10335*.
- Bai, Y., Jones, A., Ndousse, K., et al. (2022a). Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv:2204.05862*.
- Bai, Y., Kadavath, S., Kundu, S., et al. (2022b). Constitutional AI: Harmlessness from AI Feedback. *arXiv:2212.08073*.
- Birhane, A., Isaac, W., Prabhakaran, V., et al. (2022). Power to the People? Opportunities and Challenges for Participatory AI. *EAAMO 2022*. *arXiv:2209.07572*.
- Ben Allal, L., Lozhkov, A., Penedo, G., et al. (2024). Cosmopedia: Creating Large-Scale Synthetic Data for Pre-Training. HuggingFace Technical Report.
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics Derived Automatically from Language Corpora Contain Human-like Biases. *Science*, 356(6334), 183–186.
- Cao, Y., et al. (2024). Towards Scalable Automated Alignment of LLMs: A Survey. *arXiv:2406.01252*.
- Casper, S., Davies, X., Shi, C., et al. (2023). Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *arXiv:2307.15217*.
- Chen, J., et al. (2024a). PKU-SafeRLHF: Towards Multi-Level Safety Alignment for LLMs with Human Preference. *NeurIPS 2024 (Oral)*.
- Chen, Z., Deng, Y., Yuan, H., Ji, K., & Gu, Q. (2024b). Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models. *ICML 2024*. *arXiv:2401.01335*.
- Common Corpus Contributors. (2025). Common Corpus: The Largest Collection of Ethical Data for LLM Pre-Training. *arXiv:2506.01732*.
- Cui, G., Yuan, L., Ding, N., et al. (2023). UltraFeedback: Boosting Language Models with High-Quality Feedback. *arXiv:2310.01377*.
- Deep Ignorance Authors. (2025). Deep Ignorance: Filtering Pretraining Data Builds Tamper-Resistant Safeguards into Open-Weight LLMs. *arXiv:2508.06601*.
- Dodge, J., Sap, M., Marasović, A., et al. (2021). Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. *arXiv:2104.08758*.
- Dohan, D., et al. (2024). CLAIR: Evaluating Controllable Attributes in Language Model Responses. *arXiv:2410.10668*.
- Durmus, E., Nguyen, K., Liao, T. I., et al. (2024). Towards Measuring the Representation of Subjective Global Opinions in Language Models. *arXiv:2306.16388*.
- Eldan, R. & Li, Y. (2023). TinyStories: How Small Can Language Models Be and Still Speak Coherent English? *arXiv:2305.07759*.

- Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., & Kiela, D. (2024). KTO: Model Alignment as Prospect Theoretic Optimization. *arXiv:2402.01306*.
- Feng, S., Park, C. Y., Liu, Y., & Tsvetkov, Y. (2023). From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. *ACL 2023 (Best Paper Award)*. *arXiv:2305.08283*.
- Gao, L., Biderman, S., Black, S., et al. (2020). The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv:2101.00027*.
- Gebru, T., Morgenstern, J., Vecchione, B., et al. (2021). Datasheets for Datasets. *Communications of the ACM*, 64(12), 86–92.
- Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. *arXiv:2009.11462*.
- Gerstgrasser, M., et al. (2024). Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data. *arXiv:2404.01413*.
- Gunasekar, S., Zhang, Y., Anber, J., et al. (2023). Textbooks Are All You Need. *arXiv:2306.11644*.
- Han, S., et al. (2024). WildGuard: Open One-Stop Moderation Tools for Safety Risks, Jailbreaks, and Refusals of LLMs. *arXiv:2406.18495*.
- Hausenloy, J., et al. (2024). Towards Data Governance of Frontier AI Models. *arXiv:2412.03824*.
- Hartmann, J., Schwenzow, J., & Witte, M. (2023). The Political Ideology of Conversational AI: Converging Evidence on ChatGPT’s Pro-Environmental, Left-Libertarian Orientation. *arXiv:2301.01768*.
- Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022). Training Compute-Optimal Large Language Models. *arXiv:2203.15556*.
- Hong, J., Lee, N., & Thorne, J. (2024). ORPO: Monolithic Preference Optimization without Reference Model. *arXiv:2403.07691*.
- Hsu, C.-Y., et al. (2024). Safe LoRA: The Silver Lining of Reducing Safety Risks when Fine-tuning Large Language Models. *NeurIPS 2024*. *arXiv:2405.16833*.
- Huang, T., et al. (2024a). Booster: Tackling Harmful Fine-tuning for Large Language Models via Attenuating Harmful Perturbation. *ICLR 2025 (Oral)*. *arXiv:2409.01586*.
- Huang, Y., et al. (2024b). Lazy Safety Alignment for Large Language Models against Harmful Fine-tuning. *arXiv:2405.18641*.
- Huang, Z., et al. (2024c). Antidote: Post-fine-tuning Safety Alignment for Large Language Models against Harmful Fine-tuning. *arXiv:2408.09600*.
- Huang, Y., et al. (2024d). TrustLLM: Trustworthiness in Large Language Models. *ICML 2024*. *arXiv:2401.05561*.
- Ji, Z., Lee, N., Frieske, R., et al. (2023). AI Alignment: A Comprehensive Survey. *arXiv:2310.19852*.
- Ji, J., Liu, M., Dai, J., et al. (2024). BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset. *NeurIPS 2023*. *arXiv:2307.04657*.

- Kandpal, N., Lester, B., Raffel, C., et al. (2025). The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text. *arXiv:2506.05209*.
- Kirk, H. R., Vidgen, B., Röttger, P., & Hale, S. A. (2024). The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. *arXiv:2404.16019*.
- Köpf, A., Kilcher, Y., von Rütte, D., et al. (2023). OpenAssistant Conversations – Democratizing Large Language Model Alignment. *NeurIPS 2023*. arXiv:2304.07327.
- Li, X., et al. (2023a). Self-Alignment with Instruction Backtranslation. *ICLR 2024*. arXiv:2308.06259.
- Li, Y., et al. (2023b). Textbooks Are All You Need II: phi-1.5 Technical Report. *arXiv:2309.05463*.
- Li, B., et al. (2024). SALAD-Bench: A Hierarchical and Comprehensive Safety Benchmark for Large Language Models. *ACL 2024 Findings*. arXiv:2402.05044.
- Li, H., et al. (2025). Layer-Aware Representation Filtering: Purifying Finetuning Data to Preserve LLM Safety Alignment. *EMNLP 2025*. arXiv:2507.18631.
- Liang, Z., Wang, P., Zhang, R., Hu, H., Zhang, S., Ye, Q., Xu, N., Xiao, Y., Zhang, C., & Cui, L. (2025). Exploring Intrinsic Alignments within Text Corpus. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI-25)*, pp. 27455–27463.
- Liu, T., et al. (2024). Statistical Rejection Sampling Improves Preference Optimization. *ICLR 2024*. arXiv:2309.06657.
- Liu, J., et al. (2025). Demystifying Synthetic Data in LLM Pre-training. *arXiv:2510.01631*.
- Long, S., et al. (2024). LLMs for Generating and Evaluating Counterfactuals: A Comprehensive Study. *arXiv:2405.00722*.
- 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*.
- Longpre, S., Yauney, G., Reif, E., et al. (2024). A Pretrainer’s Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, and Toxicity. *arXiv:2305.13169*.
- Luccioni, A. S. & Viviano, J. D. (2021). What’s in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus. *arXiv:2105.02732*.
- Maini, P., Saha, S., Zhuo, Z., et al. (2024). Rephrasing the Web: A Recipe for Compute and Data-Efficient Language Modeling. *ACL 2024*. arXiv:2401.16380.
- Maini, P., et al. (2025). Safety Pretraining: Toward the Next Generation of Safe AI. *arXiv:2504.16980*.
- Mazeika, M., et al. (2024). HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. *ICML 2024*.
- Meng, Y., Xia, M., & Chen, D. (2024). SimPO: Simple Preference Optimization with a Reference-Free Reward. *arXiv:2405.14734*.
- Mökander, J., Juneja, P., Watson, D. S., & Floridi, L. (2023). Auditing Large Language Models: A Three-Layered Approach. *AI and Ethics*, 3(4), 1227–1241.

- Mukherjee, S., Mitra, A., Jawahar, G., et al. (2023). Orca: Progressive Learning from Complex Explanation Traces of GPT-4. *arXiv:2306.02707*.
- NVIDIA. (2024). Nemotron-4 340B Technical Report. *arXiv:2406.11704*.
- Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. *NeurIPS 2022*. arXiv:2203.02155.
- Penedo, G., Kydlíček, H., Lozhkov, A., et al. (2024). FineWeb: Decanting the Web for the Finest Text Data at Scale. *NeurIPS 2024 Datasets and Benchmarks Track*. arXiv:2406.17557.
- Perez, E., Ringer, S., Lukošiušė, K., et al. (2023). Discovering Language Model Behaviors with Model-Written Evaluations. *ACL 2023 Findings*. arXiv:2212.09251.
- Qi, X., Zeng, Y., Xie, T., et al. (2023). Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! *ICLR 2024*. arXiv:2310.03693.
- Qi, X., Zeng, Y., Xie, T., et al. (2024). Safety Alignment Should Be Made More Than Just a Few Tokens Deep. *ICLR 2025*. arXiv:2406.05946.
- 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.
- Röttger, P., et al. (2024). SafetyPrompts: A Systematic Review of Open Datasets for Evaluating and Improving Large Language Model Safety. *AAAI 2025*. arXiv:2404.05399.
- SafeMERGE Authors. (2025). SafeMERGE: Preserving Safety in Fine-tuned LLMs via Selective Layer Merging. *arXiv preprint*.
- Safer Pretraining Authors. (2025). Towards Safer Pretraining: Analyzing and Filtering Harmful Content in Webscale Datasets for Responsible LLMs. *arXiv:2505.02009*.
- Santurkar, S., Durmus, E., Ladhak, F., et al. (2023). Whose Opinions Do Language Models Reflect? *ICML 2023*. arXiv:2303.17548.
- Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., & Kersting, K. (2022). Large Pre-trained Language Models Contain Human-like Biases of What is Right and Wrong to Do. *Nature Machine Intelligence*, 4, 258–268.
- Sharma, M., et al. (2024). Towards Understanding Sycophancy in Language Models. *ICLR 2024*. arXiv:2310.13548.
- Shen, T., Jin, R., Huang, Y., et al. (2023). Large Language Model Alignment: A Survey. *arXiv:2309.15025*.
- Shen, Z., et al. (2024). SEAL: Safety-Enhanced Aligned LLM Fine-tuning via Bilevel Data Selection. *ICLR 2025*. arXiv:2410.07471.
- Shumailov, I., Shumaylov, Z., Zhao, Y., et al. (2024). AI Models Collapse When Trained on Recursively Generated Data. *Nature*, 631, 755–759.
- Srivastava, M., et al. (2025). Value Drifts: Tracing Value Alignment During LLM Post-Training. *arXiv:2510.26707*.
- Su, Y., et al. (2024). Nemotron-CC: Transforming Common Crawl into a Refined Long-Horizon Pretraining Dataset. *arXiv:2412.02595*.

- Sun, Z., Shen, Y., Zhou, Q., et al. (2023a). Principle-Driven Self-Alignment of Language Models from Scratch with Minimal Human Supervision. *NeurIPS 2023*. arXiv:2305.03047.
- Sun, Z., Shen, Y., Zhang, S., et al. (2023b). SALMON: Self-Alignment with Principle-Following Reward Models. *ICLR 2024*. arXiv:2310.05910.
- Taori, R., Gulrajani, I., Zhang, T., et al. (2023). Stanford Alpaca: An Instruction-following LLaMA Model. https://github.com/tatsu-lab/stanford_alpaca.
- 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*.
- Tunstall, L., Beeching, E., Lambert, N., et al. (2023). Zephyr: Direct Distillation of LM Alignment. *arXiv:2310.16944*.
- Wang, Y., Kordi, Y., Mishra, S., et al. (2023a). Self-Instruct: Aligning Language Models with Self-Generated Instructions. *ACL 2023*. arXiv:2212.10560.
- Wang, S., et al. (2023b). Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs. *EACL 2024 Findings*. arXiv:2308.13387.
- Wang, B., Chen, W., Pei, H., et al. (2023c). DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. *NeurIPS 2023 (Outstanding Benchmark Award)*. arXiv:2306.11698.
- Wang, Z., et al. (2024). A Comprehensive Survey of LLM Alignment Techniques. *arXiv:2407.16216*.
- Wei, J., Bosma, M., Zhao, V., et al. (2022). Finetuned Language Models Are Zero-Shot Learners. *ICLR 2022*. arXiv:2109.01652.
- Weidinger, L., Mellor, J., Rauh, M., et al. (2021). Ethical and Social Risks of Harm from Language Models. *arXiv:2112.04359*.
- Wu, T., et al. (2024a). Meta-Rewarding Language Models: Self-Improving Alignment with LLM-as-a-Meta-Judge. *arXiv:2407.19594*.
- Wu, T., et al. (2024b). Self-Play Preference Optimization for Language Model Alignment. *arXiv:2405.00675*.
- Xie, T., et al. (2024). SORRY-Bench: Systematically Evaluating Large Language Model Safety Refusal Behaviors. *ICLR 2025*.
- Xu, C., Sun, Q., Zheng, K., et al. (2024). WizardLM: Empowering Large Language Models to Follow Complex Instructions. *ICLR 2024*. arXiv:2304.12244.
- Yuan, W., Pang, R. Y., Cho, K., et al. (2024). Self-Rewarding Language Models. *arXiv:2401.10020*.
- Zha, D., Bhat, Z. P., Lai, K., et al. (2023). Data-centric Artificial Intelligence: A Survey. *arXiv:2303.10158*.
- Zhang, Z., et al. (2024). SafetyBench: Evaluating the Safety of Large Language Models. *arXiv:2309.07045*.
- Zhao, D., Scheuerman, M. K., Chitre, P., et al. (2024). A Taxonomy of Challenges to Curating Fair Datasets. *NeurIPS 2024 Datasets and Benchmarks*. arXiv:2406.06407.

Zhou, C., Liu, P., Xu, P., et al. (2023). LIMA: Less Is More for Alignment. *NeurIPS 2023*. arXiv:2305.11206.

Zmigrod, R., Mielke, S. J., Wallach, H., & Cotterell, R. (2019). Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. *ACL 2019*.