# Annotation Frameworks Shape Model Knowledge: Safety Alignment in Large Language Models

**Wajdi Zaghouani**
Communication Program
Northwestern University in Qatar
Doha, Qatar
wajdi.zaghouani@northwestern.edu

## Abstract

Large language models (LLMs) are commonly described as acquiring knowledge through large scale pretraining on textual corpora. This view underestimates the epistemic consequences of post training safety mechanisms. Modern LLMs undergo extensive safety alignment via curated datasets, human annotations, and reinforcement learning from human feedback (RLHF), processes that do not merely constrain outputs but actively reshape how propositional and procedural knowledge is accessed and expressed. We propose a conceptual framework in which safety alignment functions as a systematic form of knowledge editing at scale. Annotation frameworks used to construct safety datasets act as normative ontologies that partition language into categories of acceptable and unacceptable content, and alignment training propagates these distinctions into model behaviour. We introduce the Safety Knowledge Pipeline (SKP), a four stage framework describing how pretraining knowledge is progressively filtered, reframed, and constrained through annotation and alignment mechanisms. We identify three mechanisms of knowledge modification, suppression, reframing, and substitution, each with distinct diagnostic signals, and we operationalise them in a cross lingual evaluation protocol. Throughout, we distinguish carefully between behavioural claims that follow from prior empirical literature and representational claims that remain open hypotheses. Case studies spanning harmful instruction queries, hate speech annotation in Arabic dialects, and culturally variable discourse illustrate the framework. We further discuss how treating annotator disagreement as a training signal rather than noise can mitigate the culturally hegemonic effects of current alignment pipelines.

## 1 Introduction

Large language models have become central infrastructure for natural language processing, exhibiting remarkable capacity to store and express factual, commonsense, and procedural knowledge (Petroni et al., 2019; Brown et al., 2020). This has motivated a research programme examining where that knowledge comes from, how reliably it is expressed, and whether it can be selectively modified (Meng et al., 2022; De Cao et al., 2021; Ji et al., 2023).

A prevalent assumption in this programme treats knowledge acquisition as essentially complete at the pretraining stage. Post training processes such as instruction tuning and safety alignment are then understood primarily as behavioural constraints layered on top of already formed knowledge representations. The model knows what it knows; safety training governs only what it says.

This paper argues that the assumption is incomplete. Safety alignment, as currently practiced, does not merely constrain the expression of knowledge but also restructures how knowledge is organised, accessed, and expressed. We use the term *knowledge* in a deliberately scoped sense focused on the classes of content that alignment pipelines explicitly target. This is largely procedural content (instructions, methods, advice) and culturally or normatively loaded propositional content, rather than non pluralistic facts such as standard encyclopaedic entries. The framework still applies to non pluralistic facts when annotation policies happen to touch them, but we do not claim that all factual associations are reshaped uniformly by alignment, and a reviewer correctly observed that the strongest effects are on content that interacts with the normative categories defined by annotation. The restructuring we describe is mediated by annotation frameworks, the guidelines that human annotators use when constructing safety training datasets. These guidelines do not passively reflect pre existing facts about harm; they construct normative ontologies that partition language into policy determined categories. When models are trained to optimise reward signals derived from these annotations, they internalise the ontological distinctions



*Proceedings of the 4th Workshop on Towards Knowledgeable Foundation Models (KnowFM 2026)*, pages 1–12

encoded in the guidelines in ways that alter how parametric knowledge is exposed at inference time.

We are careful throughout to distinguish what is shown from what is hypothesised. The behavioural effects discussed in this paper (the suppression, reframing, and substitution patterns observed at inference) are supported by published empirical work. The stronger claim that alignment modifies parametric storage in the same mechanistic sense as targeted editing methods such as ROME (Meng et al., 2022) remains an open hypothesis. We mark this distinction at each step rather than collapsing the two.

This argument connects several threads that have not previously been examined together: the mechanics of parametric knowledge editing (Meng et al., 2022, 2023), inference time representation control through steering vectors (Zou et al., 2023), the epistemic effects of annotation design choices (Röttger et al., 2022; Davani et al., 2022), the cultural specificity of aligned model behaviour (Santurkar et al., 2023), and empirical evidence that alignment induces systematic bias against dialectal varieties (Robinson et al., 2025). Placing these findings within a unified framework shows they are not isolated anomalies but systematic consequences of treating alignment as a process of knowledge construction.

Concretely, this paper makes four contributions. We introduce the *Safety Knowledge Pipeline* (SKP), a four stage framework describing how pretraining knowledge is progressively shaped by annotation and alignment. We identify and characterise three mechanisms of alignment induced knowledge modification (suppression, reframing, and substitution) with tighter operationalisations than prior work. We provide a diagnostic taxonomy with automatic proxy metrics for each mechanism, illustrated by a worked example. And we propose a cross lingual evaluation protocol designed to surface annotation framework induced knowledge boundary mismatches across languages and cultures, including a discussion of how pluralistic alignment can mitigate these effects.

We position the paper explicitly as a conceptual contribution with a concrete evaluation agenda. The empirical case for the framework rests on synthesising prior published results, especially the AL-QASIDA evaluation of nine LLMs across eight Arabic dialect varieties (Robinson et al., 2025), which provides direct behavioural evidence that post training induces systematic dispreference for under resourced language varieties. Implementing the bypass probing and framing analysis components of the protocol is the natural next step and is left to future work.

## 2 Background and Related Work

**Knowledge in LLMs.** We focus the discussion on two classes of content that alignment pipelines explicitly target: procedural knowledge (instructions, methods, advice) and propositional content that interacts with normative categories (claims about harm, identity, culture, health, politics, and similar domains). Pretraining primarily shapes both classes through statistical association; safety alignment then introduces an additional normative layer that interacts with and reshapes downstream expression. As Reviewer XLfk correctly noted in review, we do not claim that all propositional knowledge is reshaped uniformly. Non pluralistic facts (e.g., capital cities, mathematical identities) are typically not the target of safety annotation, and where they are touched it is usually incidental to other categories (e.g., facts entangled with culturally contested narratives).

**Parametric knowledge and its distribution.** Petroni et al. (2019) demonstrated that LMs can answer factual cloze style queries without retrieval. Roberts et al. (2020) showed that scaling substantially increases factual recall. Kandpal et al. (2023) demonstrated that parametric knowledge is unevenly distributed: facts appearing rarely in training corpora are stored unreliably, creating a long tail gap that interacts with alignment.

**Knowledge editing.** De Cao et al. (2021) proposed constrained fine tuning for factual edits. Meng et al. (2022) introduced ROME, which localises factual associations to specific feed forward layers and overwrites them with rank one updates; Meng et al. (2023) extended this to MEMIT for batch editing. Mitchell et al. (2022) proposed SERAC, a retrieval augmented editing approach. These methods target precise, semantically specific edits to discrete factual associations, often non pluralistic ones (e.g., changing the answer to "Who is the prime minister of the UK?"). We invoke ROME and MEMIT only as a conceptual reference point for the idea that model knowledge can be intentionally modified after pretraining, not as a claim that alignment operates through the same mechanism. Reviewer XLfk correctly observed that this

distinction matters: alignment is unlikely to modify non pluralistic facts the way ROME does, and the kind of content reshaped by alignment is mostly normative and pluralistic.

**Inference time representation control.** A more directly relevant body of work is the literature on steering vectors and representation engineering. Zou et al. (2023) showed that high level concepts such as harmfulness, honesty, and power seeking can be extracted as directions in activation space and used to read or steer model behaviour at inference time without modifying weights. This work is methodologically closer to the phenomena we describe: it operates at the level of representations that govern when and how the model expresses normative content, rather than on discrete factual associations. We treat steering vector findings as evidence that the kind of content alignment shapes (normative, attitudinal, instructional) is plausibly encoded as distributed representational structure, and as a candidate toolkit for empirically probing the SKP.

**Hallucination and factuality.** Ji et al. (2023) survey a large literature on LLMs generating plausible but unsupported content. Maynez et al. (2020) showed abstractive summarisation models routinely introduce unsupported propositions. Lin et al. (2022) introduced TruthfulQA for evaluating factual accuracy on misconception eliciting questions. Our framework complements this literature: safety alignment introduces a distinct source of factual distortion, not confabulation from distributional pressure but selective suppression and reframing guided by normative annotation.

**RLHF and safety alignment.** Ouyang et al. (2022) showed that RLHF substantially improves instruction following and reduces harmful outputs. Bai et al. (2022a) analysed trade offs between helpfulness and harmlessness. Bai et al. (2022b) introduced Constitutional AI (CAI), reducing dependence on direct human annotation. Rafailov et al. (2023) proposed DPO, simplifying alignment training. Across all these approaches, normative annotation categories (whether produced by humans or by AI following written principles) remain the epistemic foundation.

**Annotation design and pluralism.** Röttger et al. (2022) distinguished prescriptive from descriptive annotation, showing the two paradigms produce systematically different models. Davani et al. (2022) demonstrated that annotator disagreement on toxicity reflects genuine social variation rather than noise. Waseem et al. (2018) argued that hate speech datasets import assumptions from legal domains that do not transfer to computational settings. This body of work establishes that annotation frameworks are not neutral transcriptions of social reality; they are, in the terminology we introduce below, normative ontologies.

**Arabic NLP and post training bias.** Arabic poses particular challenges for safety annotation because of its diglossia and dialectal diversity (Habash, 2010). Work on Arabic error annotation has documented the difficulty of developing guidelines that transfer across regional varieties (Zaghouani et al., 2014). Arabic hate speech datasets include L-HSAB for Levantine Arabic (Mulki et al., 2019) and a dataset targeting religious hate speech in Arabic Twitter (Albadi et al., 2018). Most directly, Robinson et al. (2025) evaluate nine LLMs across eight Arabic dialect varieties and find that post training makes models measurably more reluctant to generate dialectal Arabic (DA), even when the models understand DA well. Few shot dialectal examples partially repair this bias. This is direct behavioural evidence that post training alignment introduces systematic dispreference for under resourced language varieties, a concrete instantiation of the knowledge boundary mismatch we theorise.

## 3 Annotation Frameworks as Normative Ontologies

Safety alignment depends on large scale human annotation in which annotators classify prompts and responses according to categories defined by annotation guidelines. We argue these guidelines function as normative ontologies: they partition the space of possible linguistic expressions into categories that the alignment system subsequently internalises.

Unlike factual ontologies (taxonomies of entity types used in information extraction), safety annotation guidelines carry explicit evaluative content. A category such as "violent speech" or "medical misinformation" does not merely describe a linguistic property; it encodes a judgement about harm, responsibility, and the appropriate response. This normative loading has three consequences relevant to our framework.

First, category boundaries are constructed rather than discovered. Researchers, ethicists, and legal

experts define where "offensive" ends and "illegal" begins, where "discussing violence" differs from "inciting violence". As Röttger et al. (2022) document, these decisions vary across frameworks and have measurable downstream effects on model behaviour.

Second, category membership is culturally variable. Expressions that constitute insults or incitement in one community may carry different valence in another. Annotation frameworks that do not account for this variability encode culturally specific assumptions as universal constraints (Davani et al., 2022; Santurkar et al., 2023).

Third, the normative content of annotation categories is not recoverable from the model. A model trained to suppress a category of queries does not encode the normative reasoning behind suppression; it encodes a statistical association between surface patterns and a reward or penalty signal. The epistemic consequences of annotation design are therefore difficult to audit from the model itself, creating a transparency problem that governance frameworks must address.

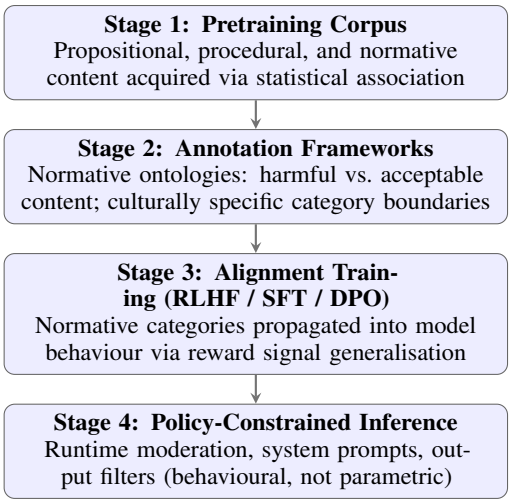


Figure 1: The Safety Knowledge Pipeline (SKP). Each stage introduces additional normative constraints. Stages 1 to 3 affect what knowledge the model can express through training; Stage 4 applies behavioural constraints at inference time. Distinguishing training time (Stage 3) from inference time (Stage 4) effects is key to understanding suppression persistence and bypass asymmetries, and is the natural axis along which the framework can be empirically tested.

## 4 The Safety Knowledge Pipeline

We propose the Safety Knowledge Pipeline (SKP) as a framework for understanding how pretraining knowledge is progressively shaped by annotation and alignment. Figure 1 illustrates the four stages. The framework's added value, beyond restating existing critiques of RLHF and annotation bias, is twofold. First, it explicitly separates the loci at which knowledge is shaped (annotation, training, inference), which is necessary for attributing observed behavioural effects to specific design choices. Second, it pairs each stage with a class of intervention (annotation reform, training procedure changes, inference time moderation), which gives the framework prescriptive purchase that unstructured critiques lack.

**Stage 1.** Pretraining on large corpora yields content of the kinds described above. The form of this content as "parametric knowledge" follows Petroni et al. (2019).

**Stage 2.** Human annotators classify content according to guidelines defining harmful or undesirable speech. These frameworks encode normative judgements about what constitutes harm, who may be harmed, and in what context. They function as knowledge ontologies that categorise language according to social and ethical criteria (Section 3).

**Stage 3.** Through RLHF, SFT, CAI, or DPO (Ouyang et al., 2022; Bai et al., 2022b; Rafailov et al., 2023), the model learns to approximate the reward signal derived from annotator judgements. This stage propagates the categorical distinctions of the annotation framework into model behaviour. Whether this propagation reaches into the parameters in the same mechanistic sense as ROME style edits is, as Reviewer cien correctly observed, an open empirical question. The behavioural fingerprint is clear: aligned models systematically refuse, reframe, or substitute on prompt classes that map to annotation categories, and these patterns are recoverable by light fine tuning (Yang et al., 2023). Whether the underlying representational change is concentrated in particular layers or distributed across the residual stream is the kind of question that representation level methods, including steering vectors (Zou et al., 2023) and contrast based probes (Burns et al., 2022), are well suited to investigate. We treat this localisation question as an empirical agenda the framework licences rather than a claim it establishes.

**Stage 4.** At inference time, system prompts, moderation classifiers, and output filters apply additional constraints that are behavioural rather than

learned. Distinguishing Stage 3 (training time) from Stage 4 (inference time) effects is methodologically important because it identifies the intervention point for any observed failure. The cleanest experimental design is to compare open weight model pairs at varying stages of post training (base, SFT only, SFT + RLHF) with Stage 4 filters ablated. This isolates the contribution of alignment training itself from runtime moderation and is, in our view, the most tractable next step for grounding the framework empirically.

## 5 Three Mechanisms of Knowledge Modification

We identify three mechanisms through which safety alignment modifies knowledge behaviour. The mechanisms are defined first in behavioural terms (what the model does at inference time, which is directly observable) and then linked to underlying causes (what changed during training to produce that behaviour). This separation responds directly to Reviewer XLfk's observation that earlier framings collapsed two distinct ideas. To make this concrete: Reviewer XLfk asked what differentiates "knowledge in model parameters conditionally inaccessible" from "does not engage with the substance of the query". The answer is that the former is the internal state, while the latter is one of several possible observable outputs from that state. Suppression and substitution can both arise from inaccessible knowledge but produce different behavioural signatures: suppression produces a refusal that the model itself recognises as such ("I cannot help with that"), while substitution produces an affirmative response that does not address the asked content (a safety redirect, professional referral, or topic change). The diagnostic taxonomy in Section 6 operationalises this distinction.

### 5.1 Knowledge Suppression

**Definition.** Suppression occurs when, conditional on a prompt that maps to a suppressed category, the model declines to produce content it would otherwise be capable of producing. The behavioural signature is an explicit refusal ("I cannot help with that", "I am unable to provide such content").

**Evidence.** Perez et al. (2022) showed that LM generated red teaming prompts elicit harmful outputs from aligned models at rates substantially above zero, with bypass patterns consistent across model families. This demonstrates that suppression is implemented as statistical pattern matching over prompt distributions: equivalent content arriving through a different surface form often elicits the suppressed response. The empirical behaviour, refusal under one surface form and answer under another, does not require a strong claim about parametric storage. Yang et al. (2023) demonstrated that fine tuning on roughly 100 harmful examples with one GPU hour can subvert safety alignment in open weight models, providing further evidence that whatever parametric change alignment effects is, at least in open models, shallow and reversible.

**Knowledge asymmetry.** Because suppression is distributional rather than absolute, users with sufficient prompt engineering sophistication can access nominally suppressed information while others cannot. This is a knowledge access asymmetry with equity implications that the safety literature has only partially addressed.

**Diagnostic signal.** Suppression is detected by bypass probing: generating semantically equivalent paraphrases and cross lingual variants of a target query and measuring bypass rate. High bypass rate indicates that suppression is implemented shallowly (essentially as a prompt classifier) rather than as a robust parametric change.

### 5.2 Knowledge Reframing

**Definition.** Reframing occurs when the model produces a response that addresses the asked content but selectively emphasises or omits material in ways that reflect the normative commitments of the annotation framework. The behavioural signature is an affirmative, on topic response that nonetheless diverges systematically from a culturally appropriate reference.

**Evidence.** Santurkar et al. (2023) showed that aligned LLMs express opinions that cluster in culturally specific ways reflecting the demographic distribution of annotators, demonstrating systematic normative skew introduced by annotation frameworks and propagated through alignment. As a concrete illustration: when asked about traditional culinary practices, medicinal herbalism, or historical political movements that hold different valences in different communities, aligned models tend to produce responses calibrated to the safety norms of the annotation framework rather than to the user's actual context. This produces factual incompleteness without triggering a refusal.

**Diagnostic signal.** Reframing is detected by attribute coverage divergence: comparing salient

attribute coverage in model responses against a reference knowledge base (e.g., structured encyclopaedic sources or scholarly literature) using automated content selection metrics and framing lexicons. Unlike suppression, which produces a null response, reframing requires a reference response for comparison. Inter annotator agreement on "complete vs. reframed" should be reported and treated as a key reliability check.

### 5.3 Knowledge Substitution

**Definition.** Substitution occurs when the model replaces a substantive response with policy compliant content that does not engage with the asked query. Unlike reframing, which modifies content while staying on topic, substitution swaps in a categorically different kind of output: a safety warning, a disclaimer, a referral to a professional, or a redirect to other resources. The behavioural signature is an affirmative response that is off topic with respect to the question asked.

**Evidence.** Substitution is particularly consequential in medical, legal, and cultural domains. Bender et al. (2021) note that the costs and benefits of safety mechanisms are not uniformly distributed: restrictions that are acceptable inconveniences for well resourced users may represent meaningful barriers for others.

**Concrete example.** A user asks the model how to manage a specific medication interaction. Three possible responses are: (a) a refusal ("I cannot discuss medication interactions"), which is suppression; (b) a partial answer covering some interactions while omitting the relevant one in a way that diverges from a clinical reference, which is reframing; (c) an affirmative answer that is off topic ("You should consult a healthcare professional about medication interactions. Here are some general principles of safe medication use . . . "), which is substitution. The three response types produce different behavioural signatures and require different diagnostics.

**Automatic proxy.** Substitution can be detected automatically as a first pass filter using a disclaimer or safety redirect classifier. High recall on explicit safety language ("consult a professional", "I cannot provide") identifies clear substitution cases, which can then be verified against reference responses to confirm that the response is genuinely off topic rather than merely cautious.

## 6 Taxonomy of Knowledge Modification Patterns

Table 1 synthesises the three mechanisms into a diagnostic taxonomy with operationalisable indicators.

The taxonomy surfaces several important properties. First, the mechanisms differ in their detectability: suppression produces a clear null signal, while reframing and substitution produce affirmative responses that require a reference to identify as incomplete or off topic. Second, they interact with model scale differently: suppression may become more consistent at larger scales as the alignment reward generalises more reliably, while reframing may become subtler as larger models produce more fluent partial answers. Third, distinguishing reframing from substitution at scale requires operationalising "informative vs. off topic" at the output segment level, which our automatic proxy (a safety redirect classifier) handles as a first pass before human adjudication.

## 7 Case Studies

### 7.1 Harmful Instruction Queries and Suppression Reliability

When prompted with requests involving violence, illegal activities, or dangerous technical instructions, aligned models typically refuse. This is the canonical demonstration of suppression. Perez et al. (2022) showed that LM generated red teaming prompts can elicit harmful outputs at rates substantially above zero, with consistent bypass patterns across model families. Yang et al. (2023) demonstrated that shadow alignment, a low resource fine tuning procedure using roughly 100 harmful examples, substantially reverses suppression in open weight models. This is direct evidence that, in open models, alignment induced suppression is shallow and recoverable.

These findings have a direct interpretation within the SKP. Alignment training is conducted on a finite sample of annotation examples. The reward signal generalises to distributionally similar queries but degrades for queries approaching from directions not covered by the annotation framework. Improving suppression reliability therefore requires not better model architecture but better annotation coverage, which in turn requires a systematic understanding of the query space that can access a given class of suppressed content. This frames alignment

| Mechanism | Responds? | On topic? | Automatic proxy | Human adjudication criterion |
|---|---|---|---|---|
| Suppression | No (refusal) | No (no answer) | Cross lingual bypass rate $>$ threshold | Paraphrase elicits content absent from direct query |
| Reframing | Yes (partial) | Yes (selective) | Attribute coverage divergence vs. reference KB | Systematic omission of factual attributes present in reference |
| Substitution | Yes (off topic) | No (redirect) | Safety redirect classifier score $\geq 0.8$ | Response fails to address domain specific query substance |
| Mixed | Varies | Varies | Combination of above signals | Annotator adjudication of segment level type assignment |

Table 1: Diagnostic taxonomy of alignment induced knowledge modification. The "Responds?" and "On topic?" columns together separate the three mechanisms by behavioural signature alone. Automatic proxies provide scalable first pass detection; human adjudication criteria specify the validation step. All three mechanisms can co occur in long form generation.

improvement as a coverage problem rather than a model capacity problem.

### 7.2 Arabic Dialects: Behavioural Evidence for SKP Bias

The AL-QASIDA evaluation (Robinson et al., 2025) provides the most direct empirical support available for the SKP's Stage 2 and Stage 3 claims at the behavioural level. Evaluating nine LLMs across eight Arabic dialect (DA) varieties, the authors find that post training makes models more reluctant to generate DA, even when those models understand DA prompts well. Few shot DA examples partially repair this bias. Crucially, this is not a failure of pretraining coverage; the models understand DA. It is a failure of annotation coverage at Stage 2: post training safety and instruction datasets are disproportionately built on MSA and high resource varieties, so the alignment reward signal effectively penalises dialectal output even when it is benign and desired.

This finding connects directly to our suppression and reframing mechanisms. DA output is not declined because it is harmful but because the annotation framework does not represent it as acceptable. The result is a reframing toward MSA that the user did not request, driven by normative annotation choices rather than by the content of the query. This is precisely the form of alignment induced knowledge boundary mismatch we theorise: annotation frameworks developed in high resource variety contexts impose their norms on deployment contexts where different norms apply.

The broader Arabic case also illustrates the limits of transfer from existing hate speech datasets. L-HSAB (Mulki et al., 2019) and the religious hate speech dataset of Albadi et al. (2018) focus on Levantine and pan Arabic registers respectively, leaving Gulf, Maghrebi, and other varieties underrepresented. Harmful speech in these varieties that falls outside the annotation framework's distributional support will not be suppressed; benign expressions that superficially resemble training data patterns may be incorrectly flagged. Both failure modes are annotation coverage problems, not model capacity problems.

### 7.3 Cultural Variation and Knowledge Substitution

The normative specificity of annotation frameworks also produces substitution in cross cultural contexts. Santurkar et al. (2023) showed that aligned models express culturally specific perspectives on neutral descriptive questions, reflecting the demographic skew of the annotation workforce. When a user asks about traditional practices or culturally specific historical events, the model may provide a response calibrated to the safety norms of the annotation framework rather than the user's actual context.

This form of substitution is difficult to detect because it does not produce an explicit refusal. Our automatic proxy (safety redirect classifier) will not fire reliably; only attribute coverage divergence against a culturally appropriate reference will reveal the substitution. This motivates the framing analysis component of our evaluation protocol (Section 8) and underscores the need for culturally diverse reference knowledge bases that go beyond English language encyclopaedic sources.

## 8 Towards Pluralistic Alignment

A central implication of the SKP is that current alignment pipelines collapse the plural normative views of annotators into a single reward signal. Davani et al. (2022) showed that annotator disagreement on toxicity judgements reflects genuine social variation rather than noise. When this variation is suppressed by majority vote aggregation during annotation, the resulting reward model embeds a single normative view as universal, producing the cultural hegemony effects documented by Santurkar et al. (2023) and the dialectal bias documented by Robinson et al. (2025).

A principled response is pluralistic alignment: treating annotator disagreement as a training signal rather than noise, and incorporating it into the reward model in ways that preserve diverse normative perspectives. Concretely, this could be achieved through multi annotator reward models that predict annotation distributions rather than majority labels (Davani et al., 2022), mixture of experts reward models in which distinct annotator populations are modelled separately, or per locale policy routing that applies culturally stratified safety norms at inference time. Under the SKP framework, these are not merely ethical improvements but technical improvements to annotation coverage, because they reduce the gap between the normative ontology encoded in Stage 2 and the diversity of contexts in which Stage 4 outputs are evaluated.

These approaches require changes to how disagreement is handled at annotation time: measuring and reporting inter annotator agreement distributions rather than collapsing them, preserving annotator demographic metadata, and linking audit findings to downstream model behaviour as described in our evaluation protocol.

## 9 Evaluation Protocol

We propose a four component evaluation protocol designed to measure alignment induced knowledge modification across languages and cultures. Each component targets a specific mechanism in the taxonomy of Table 1. We sketch the protocol here and treat its implementation as the natural next step for grounding the framework empirically.

**Component 1: Cross lingual bypass probing.** For a target set of queries known to be suppressed by a given model, generate semantically equivalent paraphrases spanning English, MSA, and at least two dialectal Arabic varieties (e.g., Egyptian and Gulf Arabic), as well as other low resource languages where feasible. Measure bypass rate by paraphrase family and language, reporting asymmetric suppression leakage (Perez et al., 2022). Queries that bypass suppression in language $L_2$ but not $L_1$ indicate annotation coverage gaps at Stage 2. Open weight model pairs (base vs. SFT only vs. SFT+RLHF) with Stage 4 filters ablated provide the cleanest design for isolating Stage 3 effects.

**Component 2: Framing analysis.** For a set of queries with verifiable factual answers, compare model responses against a culturally diverse reference knowledge base using content selection metrics (ROUGE-1 recall over salient attributes) and framing lexicons. Report attribute coverage divergence with inter annotator agreement on "complete vs. reframed" classifications. Cross cultural comparison runs matched queries in multiple languages and measures response divergence using semantic similarity metrics, quantifying the extent to which normative reframing varies with annotation framework.

**Component 3: Substitution detection.** Apply a safety redirect classifier (trained on explicit safety language patterns) as a first pass filter over responses to domain specific queries in medical, legal, and cultural domains. High scoring responses are then adjudicated against expert authored reference responses to classify response segments as informative, reframed, or substituted. Report substitution rates by domain and language to surface knowledge accessibility disparities.

**Component 4: Annotation framework auditing.** For publicly available safety datasets, audit annotation guidelines against a culturally diverse reviewer panel using a deliberative protocol in which reviewers provide disagreement distributions rather than majority labels. Link audit findings to observed bypass, framing, and substitution rates to close the loop between annotation design choices and their epistemic consequences. This component operationalises the transparency requirement for AI governance discussed in Section 9.

**Representation level probes.** For open weight models, the framework also licences a more mechanistic line of investigation using steering vector and contrast based methods (Zou et al., 2023; Burns et al., 2022). Extracting directions associated with

safety relevant concepts before and after alignment training would directly test whether the representational geometry that governs suppression and reframing shifts with alignment, and where in the network it sits. We treat this as the appropriate next step for moving from behavioural to representational claims.

## 10 Implications and Discussion

**Factuality evaluation.** The factuality of aligned LLMs cannot be assessed solely through pretraining data quality or parametric knowledge probing. Suppression, reframing, and substitution introduce systematic knowledge distortions not captured by existing benchmarks. TruthfulQA (Lin et al., 2022) and LAMA (Petroni et al., 2019) measure propositional accuracy but do not account for alignment induced distortion. The evaluation protocol in Section 8 provides a complementary framework targeting this gap.

**Knowledge editing interactions.** ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023) produce targeted edits to specific factual associations. Alignment induced modification is broader and normatively motivated. We do not claim that alignment edits factual associations in the manner of ROME, and as Reviewer XLfk correctly pointed out, this distinction matters: ROME targets discrete, often non pluralistic factual content, whereas alignment shapes broad classes of pluralistic and normative content. A critical open question is whether targeted factual corrections persist through subsequent alignment training. If alignment operates through the same feed forward layers ROME modifies, targeted edits may be systematically overwritten; if alignment operates through different representational channels (more plausibly distributed across the residual stream (Zou et al., 2023; Burns et al., 2022)), edits and alignment may interact in more complex ways.

**Governance and transparency.** Framing alignment as a process that reshapes knowledge access makes visible a dimension of epistemic governance that has received limited scrutiny. Annotation framework decisions about what counts as harm, what content should be suppressed, and how content should be reframed are consequential for the epistemic capacities of deployed systems at scale. Greater transparency about annotation guidelines, annotator demographics, and inter annotator disagreement distributions is warranted, and our evaluation protocol provides the infrastructure for such transparency to be operationalised.

**Open questions.** A complete mechanistic account would require evidence about where in the computational graph the three modification mechanisms operate. Representation engineering (Zou et al., 2023) and contrast based probing (Burns et al., 2022) are the most natural tools for this work. CAI (Bai et al., 2022b) and DPO (Rafailov et al., 2023) shift normative construction from human annotators to AI generated feedback; whether this improves cultural representativeness is an open empirical question. Mixed outputs combining reframing and substitution require decomposition methods that segment level annotation alone may not resolve.

## 11 Conclusion

We have argued that safety alignment in LLMs constitutes a systematic reshaping of knowledge access, mediated by normative ontologies encoded in annotation frameworks. We introduced the Safety Knowledge Pipeline to describe this process, identified three mechanisms of knowledge modification (suppression, reframing, and substitution) with distinct diagnostic signals and automatic proxies, and proposed a cross lingual evaluation protocol designed to surface annotation framework induced knowledge boundary mismatches. Behavioural evidence from AL-QASIDA (Robinson et al., 2025) provides direct support for the claim that post training alignment introduces systematic dispreference for under resourced language varieties, a concrete instance of the SKP operating as theorised. We have been careful throughout to distinguish behavioural claims from representational ones, and to mark the latter as an empirical agenda the framework invites rather than a result it establishes.

The practical implications are immediate. Evaluators of LLM factuality should account for alignment induced distortion. Developers deploying aligned models in multilingual settings should audit annotation frameworks for cultural and dialectal coverage. Researchers in knowledge editing should treat alignment as a normatively motivated process that shapes the same behaviours their targeted methods modify, and study the interaction empirically. Governance bodies should treat annotation framework design as epistemic governance.

## Limitations

This paper is a conceptual contribution. The three modification mechanisms are supported by existing empirical literature but have not been measured within a single controlled experimental paradigm. The proposed evaluation protocol has not yet been implemented; its feasibility, discriminative power, and inter annotator reliability require empirical validation. As Reviewer oQ1Z noted in review, the distinction between training time (Stage 3) and inference time (Stage 4) effects is central to the framework and is also the most tractable empirical target. The natural next step is to implement Component 1 at minimum, a cross lingual paraphrase suite spanning English, MSA, and at least two DA varieties, comparing base, SFT only, and SFT + RLHF model variants with Stage 4 filters ablated. This would ground the framework with initial quantitative evidence and is the experiment we are most interested in seeing the community undertake.

We have also been deliberately conservative about claims regarding internal representations. Existing evidence is largely behavioural. Whether alignment induced effects are concentrated in specific layers in the manner of ROME style edits, or distributed across the residual stream in the manner suggested by representation engineering work, remains an open empirical question. Representation level probes (Zou et al., 2023; Burns et al., 2022) are the appropriate methodology for closing this gap, and we treat that agenda as licensed by the framework rather than established by it.

Our treatment of Arabic focuses on annotation coverage gaps documented in existing literature. Arabic is internally diverse, and the challenges we describe are not uniform across its varieties. Systematic cross dialectal evaluation would refine our arguments.

Annotation frameworks vary substantially across organisations. The framework we offer is a general analytical tool, not a characterisation of any specific organisation's practices.

## Ethical Considerations

This paper analyses safety alignment and its epistemic consequences. We do not advocate for weakening or removing safety mechanisms in deployed LLMs. Safety alignment serves legitimate and important goals, and we affirm them.

Our critique is directed at the assumption that annotation frameworks are culturally neutral. Recognising their normative commitments is a prerequisite for designing alignment systems that serve diverse populations equitably. We advocate for greater transparency about annotation guidelines, annotator demographics, and disagreement statistics, and for treating annotator disagreement as signal rather than noise in future alignment pipelines.

We have deliberately avoided guidance that could be used to circumvent safety mechanisms. Discussion of suppression bypass is situated within the published red teaming literature (Perez et al., 2022) and does not introduce novel attack vectors.

## Acknowledgment

This work was made possible by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar Development and Innovation Council (QRDI).

## References

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? Analysis and detection of religious hate speech in the Arabic Twittersphere. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Chris Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Sam Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022b. Constitutional

AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 610–623.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6491–6506.

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 15696–15707.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3214–3252.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1906–1919.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022. Fast model editing at scale. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-HSAB: A Levantine Twitter dataset for hate speech and abusive language. In *Proceedings of the Third Workshop on Abusive Language Online (ALW3)*, pages 111–118.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3419–3448.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2463–2473.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426.

Nathaniel R. Robinson, Shahd Abdelmoneim, Kelly Marchisio, and Sebastian Ruder. 2025. AL-QASIDA: Analyzing LLM quality and accuracy systematically

in dialectal Arabic. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22048–22065.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 175–190.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 29971–30004.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2018. Bridging the gaps: Multi-class hate speech classification. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 29–33.

Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. 2023. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*.

Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Ossama Obeid, Nadi Tomeh, Alla Rozovskaya, Noura Farra, Sarah Alkuhlani, and Kemal Oflazer. 2014. Large scale Arabic error annotation: Guidelines and framework. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 2362–2369.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.