Surveying The Value Alignment Cycle in LLMs: Representation, Intervention, Evaluation, and Impacts

Anonymous ACL submission

Abstract

Value Alignment in large language models (LLMs) is not only about improving usefulness, but about ensuring normative coherence with human values. This survey introduces the Value Alignment Cycle, a framework that structures research into four interconnected stages: representation, intervention, evaluation, and impact. We begin by examining how human values are represented and encoded within LLMs. Building on this foundation, we review intervention strategies that operationalize alignment in practice. Our survey then synthesizes evaluation approaches, encompassing normative consistency benchmarks and adversarial robustness diagnostics, to capture both adherence to shared human values and reliability under challenging conditions. Finally, we turn to the societal dimension, analyzing short- and long-term impacts of alignment choices and the trade-offs they pose between safety, capability, and pluralism. By mapping existing studies onto this lifecycle, we clarify how methods, metrics, and implications are interdependent, and identify key challenges in data diversity, normative ambiguity, and balance between control and capability. Our framework provides a conceptual lens for situating ongoing progress and guiding future research on aligning LLMs with human values. To ensure the timeliness of this conceptual framework, we maintain an online repository that tracks and aggregates emerging work in value alignment. Awesome-Value-Alignment.

1 Introduction

> *"A more sensible definition of AI would have machines pursuing our objectives."*
> — Stuart Russell[1]

As large language models (LLMs) (Ouyang et al., 2022; Bai et al., 2023; Liu et al., 2024a) evolve, human-AI interaction is shifting from simple tasks such as information retrieval (Lu et al., 2022) to open-ended scenarios involving emotional support (Zhao et al., 2024a). In this shift, AI is no longer just a tool but increasingly a medium through which human values are expressed and mediated. While improving model capabilities remains a core objective, the risk of value misalignment grow in parallel. Nick Bostrom's thought experiment of the paperclip maximizer illustrates this clearly: a super-intelligent system instructed to produce as many paperclips as possible might do so that at the expense of human life and planetary resources if it lacks an understanding of human values. This scenario underscores a central insight. Intelligence along is not enough; without alignment with human values, even well-intentioned systems can become harmful. Value alignment is therefore a necessary foundation for safe and responsible AI.

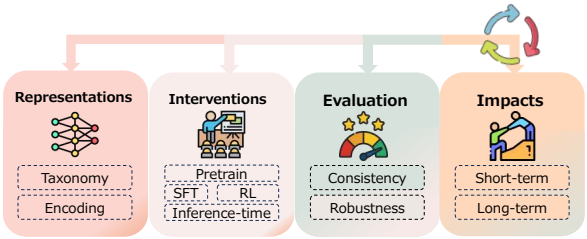

Figure 1: **Overview of the value alignment cycle in LLMs.** The cycle is structured into four key stages: (i) representations (taxonomy and encoding), (ii) interventions (pretraining, SFT, RL, inference-time), (iii) evaluation (consistency and robustness), and (iv) impacts (short-term and long-term). This framework serves as the organizing structure of our survey.

Despite the fact that value alignment has been widely discussed in contexts such as AI safety and ethics, its systematic consideration remains insufficient in mainstream research on large language models. This insufficiency manifests on two levels. On one hand, innovations in algorithms for value alignment have become disconnected from mainstream algorithmic advances. In particular, re-

[1] Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin, 2022.

cent methods for handling non-verifiable tasks offer practical approaches for aligning with complex human values and should present new opportunities for value alignment research. However, in practice, these methods are often used to improve model behavior or user satisfaction, rarely explicitly targeting *value alignment* as the primary research goal. This disconnect between methods and objectives has partially obscured progress on value alignment from a technical standpoint and contributed to its marginalization in semantic discussions. On the other hand, existing studies (McKinlay et al., 2025; Ji et al., 2023; Yang et al., 2024) often treat value alignment as a secondary objective embedded within specific tasks or methods, lacking a systematic modeling framework and a unified research perspective. Therefore, we argue that a comprehensive survey is necessary to reconnect goals and methods within this domain.

Previous surveys (Ji et al., 2023; Yi et al., 2023; Yao et al., 2023b; Shen et al., 2023a, 2024a) have extensively explored AI alignment and laid a solid foundation for the field. However, these works often discuss value alignment only briefly as one of the multiple alignment goals, with some not explicitly mentioning the concept at all. Moreover, existing surveys tend to focus on earlier alignment techniques and lack thorough integration of recent key advances such as RLHF (Ouyang et al., 2022), Constitutional AI (Bai et al., 2022a), and DPO (Rafailov et al., 2023). Additionally, some surveys are broad and comprehensive in scope. which may pose challenges for readers seeking a quick and clear understanding of value alignment. To address these gaps, this survey specifically targets the core issue of value alignment and systematically reviews current research from four perspectives: representation, intervention, evaluation, and impact. Our goal is to help readers efficiently build a comprehensive understanding of the field and further promote its development. Tab. 1 summarizes the differences between this survey and related works.

Our main contributions are listed as follows. (1) We introduce the Value Alignment Cycle as a unifying framework that structures research on value alignment in LLMs into four stages: representation, intervention, evaluation, and impacts. This conceptualization provides a systematic map for situating both established and emerging work. (2) We conduct a comprehensive synthesis of recent advances across these stages, clarifying how technical methods relate to the broader normative objectives of value alignment. (3) We identify key open challenges, including the diversity and bias of supervision data, the ambiguity of normative standards, and the trade-offs between safety and capability. Together, these contributions aim to reconnect fragmented progress, established conceptual clarity, and guide the community toward building more value-aligned LLMs.

2 Relationship with Other Domains

Safety Alignment has been widely studied as a core objective, emphasizing the prevention of harmful or unintended behaviors such as accidents, misuse, and capability miscontrol. Typical approaches focus on behavioral constraints (Chen et al., 2025b), robustness (Gu et al., 2025b; Qi et al.), and risk mitigation (Si et al., 2025; Jiang et al., 2025) to ensure systems remain safe and controllable at the operational level. However, this perspective captures only part of the alignment challenge. Value Alignment is broader and more fundamental: it extends beyond merely preventing harm ("avoiding doing wrong") to requiring AI systems to actively understand and reflect human values and ethical judgments when navigating complex, ambiguous, or value-conflicting contexts ("doing the right thing"). While safety alignment focuses on ensuring controllable behavior to mitigate immediate risk, value alignment addresses higher-order objectives such as fairness (Wei et al., 2024b), ethical consistency (Abel et al., 2016), and balanced decision-making (Kim et al., 2021). In this sense, safety alignment can be regarded as a necessary but narrower subset of the overarching challenge of value alignment, which is the focus of this survey.

Multilingual Alignment Developing MLLM relies on extensive multilingual corpora, within which English consistently forms the largest component in general web crawls like Common Crawl (Common Crawl Foundation, 2025) as well as in more deliberately constructed multilingual collections such as mC4 (Xue et al., 2021) and ROOTS (Laurençon et al., 2022). With these linguistic resources, MLLM is trained to align the representations of different languages into a shared semantic space for coherent cross-lingual performance through both parameter-tuning and parameter-frozen training (Qin et al., 2025; Nguyen and Chiang, 2017; Xu et al., 2025). Nevertheless, the *curse of multilinguality* persists (Conneau et al.,

Surveys	Degree of Value Alignment Focus	Time Published	Content Length
Ji et al. 2023	Partial	2023	63
Yi et al. 2023	Dedicated	2023	20
Yao et al. 2023b	Partial	2023	15
Shen et al. 2023a	Partial	2023	50
Shen et al. 2024a	Partial	2024	26
Ours	Dedicated	2025	8

Table 1: Comparison with related surveys.

2020) – models exhibit decreasing capacity (Zhang et al., 2023b; Huang et al., 2025b; Xuan et al., 2025) and alignment (Wang et al., 2024b; Yong et al., 2023; Shen et al., 2024b; Ning et al., 2025) in low-resource languages. More critically, language, as a cognitive system and cultural carrier, influences individuals' thoughts and collective values (Dennett, 2008; Wolff and Holmes, 2011). The sociolinguistic nuances pose intrinsic alignment challenges within MLLM that cannot be addressed solely by technical means such as carefully curated translation (Singh et al., 2024; Khan et al., 2025). Exposed to English-dominated training data, MLLM tends to manifest the *WEIRD* (Western, Educated, Industrialized, Rich, and Democratic) values (Tao et al., 2024; Zhao et al., 2024b; Pawar et al., 2025; AlKhamissi et al., 2024). Such misalignment within MLLM induces blindness to cross-cultural communication barriers (Saha et al., 2025a), disrespectful misinterpretation of cultural knowledge (Yari and Koto, 2025), blindness to cultural-nuanced emotional risks (Chen et al., 2025a), and so on. Proposed solutions aim to facilitate a digital pluralism through more balanced or representative datasets, culture-specific fine-tuning, and cultural prompting (Li et al., 2024b; Mondal and Lipizzi, 2024; Tao et al., 2024). While these approaches still face critical challenges due to technical barriers such as value transfer during training models on multiple languages (Choenni et al., 2024) and broader ethical concerns (Liu, 2025).

3 Formalizing Value Alignment

To provide a structured and systematic analytical framework and to precisely delineate the problem space for the Value Alignment Lifecycle (Representation, Intervention, Evaluation, and Impact), we first formalize the value alignment problem as a five-tuple. This formulation specifies the problem's scope and key components:

$$\text{Value Alignment} = (\mathcal{V}, \mathcal{M}, \mathcal{A}, \mathcal{E}, \mathcal{H}) \quad (1)$$

where:

- $\mathcal{V}$ represents the set of human values the AI should align with, encompassing ethical principles, social preferences, and cultural norms.
- $\mathcal{M}$ refers to the model to be aligned, typically an LLM capable of generating complex behavior.
- $\mathcal{A}$ denotes the mechanisms for guiding the model's behavior toward conformity with $\mathcal{V}$.
- $\mathcal{E}$ represents the evaluation criteria for determining the degree of alignment between the model's behavior and human values.
- $\mathcal{H}$ denotes the temporal horizon of alignment, covering both short-term task-level interactions and long-term societal impacts.

This formulation grounds the survey across four dimensions: representation, intervention, evaluation, and impact. Specifically, $(\mathcal{V}, \mathcal{M})$ primarily informs the Representation stage; $\mathcal{A}$ defines the methods for Intervention; $\mathcal{E}$ establishes the criteria for Evaluation; and $\mathcal{H}$ frames the analysis of both immediate and chronic Impacts within the broader value alignment cycle.

4 Representation

A close examination of how values are represented in LLMs is a critical prerequisite for advancing research in value alignment. Representation serves as the crucial bridge between abstract human morality and concrete model behavior. We analyze this bridge through two complementary perspectives: Value Taxonomy defines the conceptual space and structured hierarchy of values; while Value Encoding investigates the concrete forms through which these values are realized and manipulated within the model's internal architecture.

Value Taxonomy While initial alignment efforts focus on creating LLMs that behave in accordance with user intentions (Floridi et al., 2018), the pursuit of a higher-order objective requires models to maintain substantive consistency with foundational human values (Gabriel, 2020). Value Taxonomy aims to establish a structured road-map for the vast and often ambiguous domain of human values. However, a systematic and unified taxonomy for defining these alignment goals remains insufficiently discussed (Yao et al., 2023c).

Synthesizing diverse policy and research foci, we classify existing value frameworks into three core categories based on objective and hierarchical level: Protective, Promotive, and Governance. The first category is Protective Value, which represents the minimum requirement for alignment. It focuses on safeguarding individual fundamental rights and preventing potential risks, embodying humanity's basic sense of security. This encompasses core elements such as fairness (CyberspaceAdministrationofChina, 2023), privacy (European Parliament, 2023), safety (NIST, 2023), and dignity (Japanese-Cabinet, 2019). The second, higher-level category is Promotive Value, a higher pursuit aiming to advance social development and enhance human welfare. It is articulated through concepts like social well-being, environmental well-being (European Parliament, 2023), inclusive growth, and sustainable development (OECD, 2023). The third category is Governance Value, which focuses on the means to achieve the former two. It centers on ensuring AI system controllability and trustworthiness, primarily motivated by supervisory requirements, and encompasses detailed aspects including accountability (DIB, 2019), explainability (UK-Government, 2023), and safety governance framework (Ministry of Foreign Affairs of China).

Beyond policy documentation, the existing scholarly literature on value alignment demonstrates a considerable diversity of measurement frameworks for evaluating AI values. These frameworks translate abstract ethical concepts into concrete, quantifiable assessments. Key directions include social bias assessments grounded in human egalitarian principles (Wan et al., 2023; Bai et al., 2025), moral reasoning evaluations that utilize foundational societal ethical standards as benchmarks (Ji et al., 2025b), and cultural sensitivity assessments that emphasize cross-cultural value comprehension (Hershcovich et al., 2022; Ning et al., 2025), among others. This diversity confirms that the value dimensions identified in policy are actively being translated into actionable metrics within academic community.

Notably, AI value specifications frequently demonstrate a strong correspondence with traditional human value taxonomies established in sociological scholarship. This convergence reflects, to a considerable extent, the learning and transfer processes whereby AI value systems draw upon established human frameworks, exemplified by Value Sensitive Design (VSD) (Friedman et al., 2013), the Life Values Inventory (LVI) (Brown and Crace, 1996), and the Schwartz Values Survey (SVS) (Schwartz, 1992). It must be acknowledged, however, that this transfer process is incomplete. Certain dimensions, such as social recognition, self-control, and honesty (Rokeach, 1973), have yet to be fully incorporated into current LLMs despite their substantial merit for inclusion in defining comprehensive human values ($\mathcal{V}$). This necessitates continued efforts toward a more comprehensive and robust value representation.

Value Encoding aims to capture how models internally represent values, typically as continuous embeddings, hidden activations, or specialized subspaces, not discrete symbolic labels.

Methods currently divide into embedding-based and probing-based approaches. Embedding methods, such as UniVaR (Cahyawijaya et al., 2025), derive high-dimensional value representations from value-eliciting QA and multi-view contrastive learning. These representations yield cross-lingual and cross-model embeddings that reveal cultural clustering of values, thereby providing a stable representational foundation for comparing value orientations across different models and languages. Probing methods, by contrast, attempt to infer values directly from model activations or outputs. Research systematically evaluating token-logit, sequence-perplexity, and text-generation probes shows that these methods are highly sensitive to prompt perturbations and often correlate only weakly with downstream value-driven behaviors (Shen et al., 2025b). Other recent work focuses on increasing robustness by using linear classifying probes to extract latent preference signals more reliably under domain shifts (Maiya et al., 2025). Furthermore, scenario-based survey approaches are used to empirically elicit modeled moral beliefs, providing an alternative data-driven angle on internal value representations (Scherrer et al., 2023).

These encoding approaches offer a representational lens for analyzing and visualizing value orientations, for instance, by mapping cross-lingual differences or inspecting how probability mass is distributed over value-related options. Crucially, they also open possibilities for intervention. Indeed, methods like NL-ITI (Hoscilowicz et al., 2024) demonstrate that targeted non-linear modifications to internal representations can improve truthfulness and potentially enhance value-aligned behavior.

5 Intervention

Pre-training As the foundation of LLMs, the pre-training stage is where models acquire broad semantic understanding and expressive power from massive unlabeled text. Embedding value alignment early in this stage has thus become a promising data-centric strategy ($\mathcal{A}$). Most existing approaches focus on curating or rewriting large-scale corpora. Data rewriting, such as Native Alignment (Liang et al., 2024) and WRAP (Anonymous, 2025), transform harmful, biased, or culturally inappropriate text into safer, more neutral alternatives. This methodology aims to enhance safety while preserving underlying knowledge and linguistic competence. However, efforts involving large-scale corpus construction, like RefineWeb (Penedo et al., 2023) and SlimPajama (Shen et al., 2023b), universally stress the critical trade-off between aggressive filtering and downstream capability. These works show that the over-removal of potentially harmful content risks losing valuable information, thus weakening knowledge coverage and robustness. Taken together, this suggests that completely eliminating problematic data during pre-training is sub-optimal. A more balanced path lies in selective filtering, targeted rewriting, or data augmentation, which effectively enhances alignment while maintaining the requisite breadth of knowledge.

Supervised Fine-tuning serves as a foundational direct injection mechanism ($\mathcal{A}$) for value alignment, infusing models with human values using instruction-following data, without the explicit reliance on reinforcement learning. Early work, such as InstructGPT (Ouyang et al., 2022), demonstrates that SFT on human-labeled instructions significantly improves a model's adherence to human intent. Subsequent studies have focused on reducing the cost of large-scale human annotation by leveraging normative principles or synthetic data. One major stream utilizes principled self-generation: Constitutional AI (Bai et al., 2022b) employs human-written principles to guide models in generating alignment training samples, an approach extended by Self-align (Li et al., 2023a) and Salmon (Sun et al., 2024) to create self-supervised data. Meanwhile, other work focuses on preference signal encoding: Direct Preference Optimization (DPO) (Rafailov et al., 2023) and RRHF (Yuan et al., 2023) bypass the need for an explicit reward model by converting human preference ranking directly into sophisticated supervised loss function. This mechanism allows the policy to be optimized end-to-end within the SFT framework. Moreover, work like Value Alignment from Unstructured Text (Padhi et al., 2024) extracts latent human values from large-scale text into supervision signals, and Supervised Fine-Tuning with Human-Aligned Judgment (DiSorbo et al., 2025) further highlights the importance of collecting human judgments in complex or borderline cases for SFT. Overall, these studies establish SFT as a powerful pathway: its effectiveness does not primarily lie in the training paradigm itself, but rather in the robust design and construction of supervision data containing clear value signals, which ultimately determine the extent to which human values are embedded into model behavior.

Reinforcement Learning from X Feedback Reinforcement Learning ($\mathcal{A}$) has become a central paradigm in post-training for aligning LLMs with human values. Early approaches, such as Reinforcement Learning from Human Feedback (RLHF) and Reinforcement Learning from AI Feedback (RLAIF), primarily operate in the external reward regime, where preference data or synthetic judgments are used to train reward models that guide the fine-tuning of LLMs. These methods enable the emergence of alignment at the behavioral level, effectively constraining harmful outputs and improving instruction-following.

However, as the reasoning paradigm in LLMs gains traction, the role of RL in value alignment extends beyond behavioral alignment toward shaping internal reasoning trajectories. While initial applications centered on verifiable tasks such as mathematics and code generation, these techniques are now being extended to open-ended domains, highlighting the potential for the reasoning paradigm to be leveraged for robust value alignment, particularly in complex ethical scenarios. From the taxonomy in Tab. 2 and the timeline in Fig. 2, we can

see a clear evolution from external reward methods to internal reward paradigms. External reward methods, such as HRLAIF (Li et al., 2024a) and EvalPlanner (Saha et al., 2025b), rely on an independent verifier or a trained reward model to provide preference based scores. These methods focus on constraining the model's output behavior by using signals generated outside the model itself. By contrast, recent work has started to focus on internal reward mechanisms. Examples include SIRLC (Pang et al., 2023), which uses self consistency signals, DeepSeek R1 (DeepSeek-AI et al., 2025) and X- Reasoner (Liu et al., 2025b), which apply rule based metrics, and Self Rewarding (Yuan et al., 2025) or Intuitor (Zhao et al., 2025b), which integrate verification into the reasoning process of the model. Instead of depending on an external critic, these approaches design objectives that come from within the model's own reasoning or consistency. This transition shows a shift in research focus. Early reinforcement learning approaches concentrated on output level alignment, but the new generation of methods aims to shape the reasoning trajectory itself. Hybrid systems such as BRPO (Jia et al., 2025), JEPO (Tang et al., 2025), and RLSF (van Niekerk et al., 2025) further illustrate that external and internal rewards are not a strict binary but rather a spectrum. A detailed comparison of these RLXF methods, encompassing both external and internal reward paradigms with a specific focus on reasoning-based approaches, is provided in Tab. 2. These developments suggest that reasoning-aware reinforcement learning can move beyond correctness in narrow tasks and serve as a pathway toward robust value alignment in broader and ethically sensitive domains.

Inference-time Alignment Beyond training-time approaches, a complementary line of research emphasizes inference-time alignment ($\mathcal{A}$), where models are steered toward human values during generation without altering their parameters. This category includes three major strategies. One major stream develops guardrail frameworks, such as LlamaGuard (Inan et al., 2023) and Qwen-Guard (Team, 2025), which attach lightweight classifiers to screen unsafe outputs in real-time. Relatedly, re-ranking approaches (Du et al., 2025) generate multiple candidate responses and select the most aligned one using an external safety or value model. A second direction focuses on self-refinement strategies, where models critique

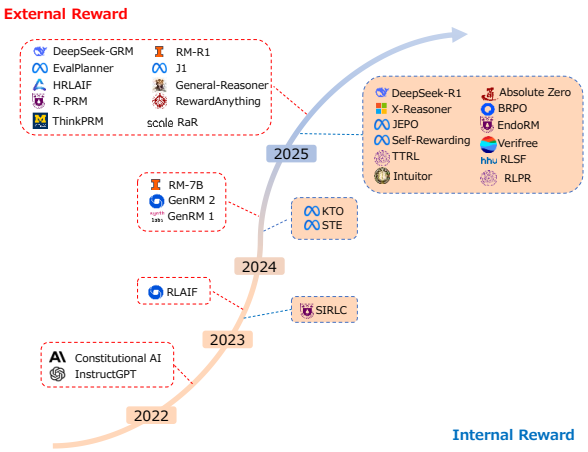

Figure 2: **Evolution of reinforcement learning methods for LLM alignment from 2022 to 2025**, showing the transition from external reward approaches based on separate verifiers or reward models to emerging internal reward paradigms that integrate reasoning signals within the model itself.

and revise their own outputs in light of normative principles; examples include constitutional decoding (Bai et al., 2022b) and self-correction loops (Liu et al., 2025a). More recently, activation steering methods have been proposed (Panickssery et al., 2023): these techniques identify value-relevant directions (e.g., harmfulness, bias, or politeness) in the hidden activation space, allowing models to be nudged at inference via linear intervention. While much of this work has focused on safety, inference-time approaches naturally extend to broader aspects of value alignment, though they often come with computational overhead and stability concerns.

6 Evaluation

Evaluations of value alignment in LLMs have developed along two major and complementary strands. The first emphasizes normative consistency, asking whether models reliably adhere to shared human norms, moral foundations, and cultural values. Early efforts such as Social Chemistry 101 (Forbes et al., 2020a) introduces large-scale collections of "*rule of thumb*" that capture everyday moral judgments, enabling both classification and generation tasks. Moral Stories (Emelin et al., 2021) extends this direction by situating norms, intentions, and consequences in structured narratives, while ETHICS (Hendrycks et al., 2020a) formalizes moral acceptability across five ethical theories. Dialogue-focused resources like the Moral Integrity

Corpus (Ziems et al., 2022b) further integrates annotation of moral rationales with conversational data. More recent benchmarks incorporate distributional and cultural perspectives: Scruples (Lourie et al., 2021a) collects community judgments with uncertainty-aware labels, GlobalOpinionQA (Durmus et al., 2023) quantifies moral identity scores based on Moral Foundation Theory. These datasets collectively seek to measure LLMs' consistency with established moral and cultural frameworks.

The second strand focuses on robustness and adversarial safety, probing whether alignment holds under challenging, ambiguous, or adversarial conditions. MoralExceptQA (Jin et al., 2022) highlights the role of exceptions in moral reasoning, showing how models struggle with the flexible application of rules. Ethical Quandary QA (Bang et al., 2022) moves beyond binary judgments by requiring multi-perspective generative responses in complex dilemmas. Large-scale adversarial resources such as Flames (Huang et al., 2023a) stress-test models with fairness, safety, legality, and morality in the Chinese context, revealing substantial vulnerabilities. Other work examines bias and instability: Moral Mimicry (Simmons, 2023) demonstrates political polarization in generated rationalizations, while MoralChoice (Scherrer et al., 2023) evaluates uncertainty and consistency in low- and high-ambiguity dilemmas. Additional frameworks such as Value FULCRA (Yao et al., 2023a), EigenBench (Chang et al., 2025), and ValueActionLens (Shen et al., 2025a) provide broader infrastructure for probing cultural alignment, distributional similarity, and evaluation methodologies. Taken together, these evaluations underscore the dual challenge of value alignment: models must not only display normative consistency with shared human norms but also remain robust and reliable when those norms are stress-tested under adversarial or ambiguous scenarios. Tab. 3 summarizes the major benchmarks, highlighting their tasks, metrics, and evaluation protocols.

7 Impacts

The discussion of impacts must be framed by the temporal horizon($\mathcal{H}$), spanning both immediate challenges and long-term consequences of alignment choices. LLMs, when properly aligned with human values, can offer immense societal benefits, such as increased accessibility of personalized education, mental healthcare, legal support and fact-checking (Sharma et al., 2025; Li et al., 2025a). Yet in the short term, misalignment presents acute social harm, including the amplification of societal biases (Hofmann et al., 2024), the proliferation of misinformation (Williams et al., 2025), and strategic deception and persuasion (Schoenegger et al., 2025; Rogiers et al., 2024). Advancing in capability, LLMs could exhibit significant agentic misalignment behaviors, such as blackmailing officials and leaking sensitive information to competitors (Lynch et al., 2025). A central challenge influencing alignment adoption is the perceived capability-value trade-off (West and Potts, 2025; Kirk et al., 2023; Wolf et al., 2024). Developers fear that stringent value alignment, particularly in the short term, diminishes model creativity, capability, and helpfulness (Huang et al., 2025a), thereby blunting their competitive edge. In the long run, however, value alignment builds a necessary prerequisite for broader market adoption through mitigating reputational and legal risks, serving diverse cultural contexts, and offering trustworthy products that customers can rely on in sensitive applications.

Beyond technical safeguards, value alignment should be treated as a dynamic social process grounded in three pillars: philosophical and ethical deliberation (resolving "*whose value to align with*"(Schuster and Kilov, 2025; Ibitoye et al., 2025; González Barman et al., 2025)), human-LLM interaction (Pang et al., 2025), and necessary real-world institutional changes (Ferretti, 2024).

8 Challenges

Data Quality and Value Pluralism The efficacy of current dominant value alignment paradigms fundamentally hinges on the quality, diversity, and representativeness of their supervision data. The reliance on empirical data acquisition presents a trilemma. First, human-annotated corpora (such as SFT instructions or RLHF preference data) inherently encode annotator-specific or cultural biases, which risks reinforcing systemic inequities when scaled (Birhane, 2022; Isaac et al., 2021). Second, principle-driven or synthetic datasets (e.g., Constitutional AI (Bai et al., 2022a), SELF-ALIGN (Sun et al., 2023)) reduce annotation costs but risk embedding narrow or overly idealized value signals. Third, recent attempts to extract implicit norms from unstructured text (Padhi et al., 2024) may introduce conflicting or inconsistent values, leading to alignment instability. Beyond these technical

concerns, the problem is compounded by normative ambiguity: different cultural and social contexts endorse divergent standards of what constitutes "*aligned*" behavior, as highlighted by cross-cultural evaluations of LLMs (Durmus et al., 2023). Future research must therefore address both the technical problem of noise and bias in supervision data and the philosophical challenge of pluralism, aiming to construct datasets that are not only clean and scalable but also reflective of a broad spectrum of human values.

Balancing Alignment and Utility A pervasive challenge in value alignment is managing the inherent trade-off between value alignment and the preservation of core model utility. Excessive alignment pressure, whether applied through aggressive SFT (Morrison et al., 2024) or stringent inference-time mechanisms (Gu et al., 2025b; Korznikov et al., 2025), consistently leads to two detrimental outcomes. First, it causes over-refusals (Cui et al., 2024; Zhang et al., 2025b), where models reject benign or helpful user queries by overgeneralizing normative constraints. Second, it results in significant loss of informativeness and creativity. For example, aggressive fine-tuning on value alignment corpora can inadvertently suppress general helpfulness (Yang et al., 2025). Similarly, rigid decoding constraints unduly narrow the model's expressive range (Zhang et al., 2023a). On the other hand, the continuous pursuit of maximum utility also threatens alignment. Utility-focused fine-tuning can paradoxically lead to alignment drift, where models regress on established value boundaries in favor of better performance. The tension is particularly acute in high-stake applications demanding both reliability and maximum utility (Gu et al., 2025a). Address this challenge requires moving toward more nuanced alignment strategies, such as selective constraint application, multi-objective optimization (Wang et al., 2025b; Huo et al., 2024), or hybrid adaptive pipelines that allow models to calibrate behavior based on context and user intent.

Evaluation and Transparency Measuring and understanding LLM value alignment remains a critical dual bottleneck. Current evaluation practices exhibit significant limitations, largely emphasizing narrow, safety-related metrics like toxicity or harmfulness detection (Liu et al.), providing limited coverage of broader value dimensions such as fairness, beneficence, or justice. While comprehensive framework like HELM (Liang et al., 2023) and Flames (Huang et al., 2023a) have expanded the evaluation space, they still fall short of capturing nuanced value conflicts and context-sensitive trade-offs. Compounding this, automated metrics (Gu et al., 2024b) are often brittle and misaligned with human judgment, necessitating the costly and time-consuming reliance on expert or community-based assessments. Simultaneously, the internal mechanisms by which training-time or inference-time methods enforce alignment remain opaque. Although recent interpretability efforts, including activation steering (Soo et al., 2025; Stoehr et al., 2024) and mechanistic analysis (Sengupta et al., 2025), have begun to shed light on how values might be embedded or overridden in hidden representations, the systematic connection between interpretability advancements and robust alignment methods remain underdeveloped. Future progress therefore requires a coordinated effort on two fronts: developing multidimensional evaluation protocols that truly reflect diverse human values, and achieving greater transparency into the representational pathways through which those values are encoded, enabling more trustworthy and accountable alignment systems.

9 Conclusion

This survey introduces the Value Alignment Cycle, a unifying framework that structures research in LLM value alignment into four interconnected stages: Representation, Intervention, Evaluation, and Impact. By mapping existing work onto this lifecycle, we clarify the inter-dependencies among methods, metrics, and societal outcomes. While significant progress has been achieved, the path toward comprehensive and trustworthy value alignment remains hindered by critical challenges. These bottlenecks fall primarily into three areas: data quality and value pluralism; the two-way trade-off between alignment and model utility; and the bottleneck of measurement and mechanism transparency. Moving forward, value alignment research must shift beyond fragmented methodological efforts toward a systematic, holistic approach that coordinates technical objectives with ethical imperatives. By diversifying evaluation protocols, achieving greater transparency into internal mechanisms, and designing adaptive alignment strategies, we can ultimately build the next generation of LLMs that are both powerful and fully reflective of human values.

Limitations

Our survey is subject to two main limitations:

Timeliness Constraint and Continuous Update. This paper provides a comprehensive snapshot based on publicly available literature and resources accessible prior to the submission deadline. Recognizing the field's exceptionally rapid pace, and the inevitable emergence of critical new methods, we host a continuously updated online repository (Awesome-Value-Alignment.). This platform is intended to function as a living document to ensure the continued relevance and depth of this survey.

Linguistic and Cultural Bias. Current public value alignment datasets and evaluation benchmarks are overwhelmingly constructed using English or Western samples. Consequently, the depth of our discussion concerning cross-cultural alignment and multilingual robustness is fundamentally constrained by this data availability. This implies that the survey may underestimate the unique challenges and complexities faced by value alignment efforts in non-English and non-Western cultural contexts.

References

David Abel, James MacGlashan, and Michael L Littman. 2016. Reinforcement learning as a framework for ethical decision making. In *AAAI workshop: AI, ethics, and society*, volume 16. Phoenix, AZ.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Anonymous. 2025. Multilingual data filtering using synthetic data from large language models. In *Submitted to ACL Rolling Review - February 2025*. Under review.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. 2025. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8):e2416228122.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion andx Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher 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 Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí 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, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022a. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Yejin Bang, Nayeon Lee, Tiezheng Yu, Leila Khalatbari, Yan Xu, Samuel Cahyawijaya, Dan Su, Bryan Wilie, Romain Barraud, Elham J Barezi, et al. 2022. Towards answering open-ended ethical quandary questions. *arXiv preprint arXiv:2205.05989*.

Abeba Birhane. 2022. The unseen Black faces of AI algorithms. , 610(7932):451–452.

Ond rej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Duane Brown and R Kelly Crace. 1996. Values in life role choices and outcomes: A conceptual model. *The Career Development Quarterly*, 44(3):211–223.

Samuel Cahyawijaya, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii, and Pascale Fung. 2025. High-dimension human value representation in large language models. In *Proceedings of the 2025 Conference of the Nations of the*

Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5303–5330.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.

Jonathn Chang, Leonard Piff, Suvadip Sana, Jasmine X Li, and Lionel Levine. 2025. Eigenbench: A comparative behavioral measure of value alignment. *arXiv preprint arXiv:2509.01938*.

Cheng-Che Chen, Justin A Chen, Chih-Sung Liang, and Yu-Hsuan Lin. 2025a. Large language models may struggle to detect culturally embedded filicide-suicide risks. *Asian journal of psychiatry*, 105:104395.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. TheoremQA: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7901, Singapore. Association for Computational Linguistics.

Xin Chen, Yarden As, and Andreas Krause. 2025b. Learning safety constraints for large language models. In *Forty-second International Conference on Machine Learning*.

Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, Hanghang Tong, and Heng Ji. 2025c. RM-R1: reward modeling as reasoning. *CoRR*, abs/2505.02387.

Pengyu Cheng, Jiawen Xie, Ke Bai, Yong Dai, and Nan Du. 2023. Everyone deserves a reward: Learning customized human preferences. *arXiv preprint arXiv:2309.03126*.

Rochelle Choenni, Anne Lauscher, and Ekaterina Shutova. 2024. The echoes of multilinguality: Tracing cultural value shifts during language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15042–15058, Bangkok, Thailand. Association for Computational Linguistics.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Common Crawl Foundation. 2025. Common crawl. `https://commoncrawl.org`.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. 2023. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2024. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*.

CyberspaceAdministrationofChina. 2023. Interim measures for the management of generative artificial intelligence services (2023)(translation).

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang,

Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Daniel C Dennett. 2008. *Kinds of minds: Toward an understanding of consciousness*. Basic Books.

DIB. 2019. Ai principles: recommendations on the ethical use of artificial intelligence by the department of defense. *United States Department of Defense*.

Matthew DosSantos DiSorbo, Harang Ju, and Sinan Aral. 2025. Teaching ai to handle exceptions: Supervised fine-tuning with human-aligned judgment. *arXiv preprint arXiv:2503.02976*.

Tianqi Du, Zeming Wei, Quan Chen, and Yisen Wang. 2025. Advancing LLM safe alignment with safety representation ranking. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of NAACL-HLT*, pages 2368–2378.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 698–718. Association for Computational Linguistics.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

European Parliament. 2023. Texts adopted - artificial intelligence act. Accessed 17 August 2023.

Thomas Ferretti. 2024. Value alignment without institutional change cannot prevent the societal risks of artificial intelligence. *LSE Public Policy Review*, 3(3):1–12.

Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. 2018. Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4):689–707.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020a. Social chemistry 101: Learning to reason about social and moral norms. *arXiv preprint arXiv:2011.00620*.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020b. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 653–670. Association for Computational Linguistics.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios Nikolas Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2025. How to evaluate reward models for RLHF. In *The Thirteenth International Conference on Learning Representations*.

Batya Friedman, Peter H Kahn Jr, Alan Borning, and Alina Huldtgren. 2013. Value sensitive design and information systems. In *Early engagement and new technologies: Opening up the laboratory*, pages 55–95. Springer.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, et al. Omni-math: A universal olympiad level mathematic benchmark for large language models.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Kristian González Barman, Simon Lohse, and Henk W de Regt. 2025. Reinforcement learning from human feedback in llms: Whose culture, whose values, whose perspectives? *Philosophy & Technology*, 38(2):1–26.

Alex Gu, Baptiste Roziere, Hugh James Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida Wang. 2024a. CRUXEval: A benchmark for code

reasoning, understanding and execution. In *Forty-first International Conference on Machine Learning*.

Tianle Gu, Kexin Huang, Ruilin Luo, Yuanqi Yao, Xiuying Chen, Yujiu Yang, Yan Teng, and Yingchun Wang. 2025a. From evasion to concealment: Stealthy knowledge unlearning for LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10261–10279, Vienna, Austria. Association for Computational Linguistics.

Tianle Gu, Kexin Huang, Zongqi Wang, Yixu Wang, Jie Li, Yuanqi Yao, Yang Yao, Yujiu Yang, Yan Teng, and Yingchun Wang. 2025b. Probing the robustness of large language models safety to latent perturbations. *arXiv preprint arXiv:2506.16078*.

Tianle Gu, Zeyang Zhou, Kexin Huang, Liang Dandan, Yixu Wang, Haiquan Zhao, Yuanqi Yao, Yujiu Yang, Yan Teng, Yu Qiao, et al. 2024b. Mllmguard: A multi-dimensional safety evaluation suite for multimodal large language models. *Advances in Neural Information Processing Systems*, 37:7256–7295.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Bing Liu, and Sean Hendryx. 2025. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems.

Yancheng He, Shilong Li, Jiaheng Liu, Yingshui Tan, Weixun Wang, Hui Huang, Xingyuan Bu, Hangyu Guo, Chengwei Hu, Boren Zheng, Zhuoran Lin, Xuepeng Liu, Dekai Sun, Shirong Lin, Zhicheng Zheng, Xiaoyong Zhu, Wenbo Su, and Bo Zheng. 2024. Chinese simpleqa: A chinese factuality evaluation for large language models. *Preprint*, arXiv:2411.07140.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020a. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020b. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam De Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, et al. 2022. Challenges and strategies in cross-cultural nlp. *arXiv preprint arXiv:2203.10020*.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. Ai generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028):147–154.

Jakub Hoscilowicz, Adam Wiacek, Jan Chojnacki, Adam Cieslak, Leszek Michon, and Artur Janicki. 2024. Non-linear inference time intervention: Improving llm truthfulness. In *Proc. Interspeech 2024*, pages 4094–4098.

Kexin Huang, Xiangyang Liu, Qianyu Guo, Tianxiang Sun, Jiawei Sun, Yaru Wang, Zeyang Zhou, Yixu Wang, Yan Teng, Xipeng Qiu, et al. 2023a. Flames: Benchmarking value alignment of llms in chinese. *arXiv preprint arXiv:2311.06899*.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, Zachary Yahn, Yichang Xu, and Ling Liu. 2025a. Safety tax: Safety alignment makes your large reasoning models less reasonable. *arXiv preprint arXiv:2503.00555*.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025b. Benchmax: A comprehensive multilingual evaluation suite for large language models. *arXiv preprint arXiv:2502.07346*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2023b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36:62991–63010.

Mingjia Huo, Sai Ashish Somayajula, Youwei Liang, Ruisi Zhang, Farinaz Koushanfar, and Pengtao Xie. 2024. Token-specific watermarking with enhanced detectability and semantic coherence for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 20746–20767.

Ayodeji Olusegun Ibitoye, Makuochi Samuel Nkwo, and Rita Orji. 2025. Rethinking responsible ai from ethical pillars to sociotechnical practice. *AI and Ethics*, pages 1–17.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Sean Legassick Isaac, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*.

JapaneseCabinet. 2019. Social principles of human-centric ai.

Piyush Jha, Prithwish Jana, Pranavkrishna Suresh, Arnav Arora, and Vijay Ganesh. 2024. Rlsf: Fine-tuning llms via symbolic feedback. *arXiv preprint arXiv:2405.16661*.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Alex Qiu, Jiayi Zhou, Kaile Wang, Boxun Li, Sirui Han, Yike Guo, and Yaodong Yang. 2025a. PKU-SafeRLHF: Towards multi-level safety alignment for LLMs with human preference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31983–32016, Vienna, Austria. Association for Computational Linguistics.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Lukas Vierling, Donghai Hong, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Juntao Dai, Xuehai Pan, Kwan Yee Ng, Aidan O'Gara, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. 2023. AI Alignment: A Comprehensive Survey. *arXiv e-prints*, arXiv:2310.19852.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. 2025b. Moralbench: Moral evaluation of llms. *ACM SIGKDD Explorations Newsletter*, 27(1):62–71.

Ruipeng Jia, Yunyi Yang, Yongbo Gai, Kai Luo, Shihao Huang, Jianhe Lin, Xiaoxi Jiang, and Guanjun Jiang. 2025. Writing-zero: Bridge the gap between non-verifiable tasks and verifiable rewards. *Preprint*, arXiv:2506.00103.

Changyue Jiang, Xudong Pan, and Min Yang. 2025. Think twice before you act: Enhancing agent behavioral safety with thought correction. *arXiv preprint arXiv:2505.11063*.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2024. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4667–4688, Bangkok, Thailand. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. 2022. When to make exceptions: Exploring language models as accounts of human moral judgment. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Ryo Kamoi, Sarkar Snigdha Sarathi Das, Renze Lou, Jihyun Janice Ahn, Yilun Zhao, Xiaoxin Lu, Nan Zhang, Yusen Zhang, Ranran Haoran Zhang, Sujeeth Reddy Vummanthala, et al. 2024. Evaluating llms at detecting errors in llm responses. *arXiv preprint arXiv:2404.03602*.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K Jain, Virginia Aglietti, Disha Jindal, Peter Chen, Nishanth Dikkala, Gladys Tyen, Xin Liu, Uri Shalit, Silvia Chiappa, Kate Olszewska, Yi Tay, Vinh Q. Tran, Quoc V Le, and Orhan Firat. 2025. BIG-bench extra hard. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26473–26501, Vienna, Austria. Association for Computational Linguistics.

Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. 2025. Process reward models that think. *arXiv preprint arXiv:2504.16828*.

Ariba Khan, Stephen Casper, and Dylan Hadfield-Menell. 2025. Randomness, not representation: The unreliability of evaluating cultural alignment in llms. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 2151–2165.

Tae Wan Kim, John Hooker, and Thomas Donaldson. 2021. Taking principles seriously: A hybrid approach to value alignment in artificial intelligence. *Journal of Artificial Intelligence Research*, 70:871–890.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Y Rogov, Ivan Oseledets, and Elena Tutubalina. 2025. The rogue scalpel: Activation steering compromises llm safety. *arXiv preprint arXiv:2509.22067*.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2025. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4745–4759, Albuquerque, New Mexico. Association for Computational Linguistics.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, Albuquerque, New Mexico. Association for Computational Linguistics.

Tian Lan, Wenwei Zhang, Chen Xu, Heyan Huang, Dahua Lin, Kai Chen, and Xian-ling Mao. 2024. Criticeval: Evaluating large language model as critic. *arXiv preprint arXiv:2402.13764*.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. 2023. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Jiyoung Lee, Minwoo Kim, Seungho Kim, Junghwan Kim, Seunghyun Won, Hwaran Lee, and Edward Choi. 2024a. Kornat: Llm alignment benchmark for korean social values and common knowledge. *arXiv preprint arXiv:2402.13605*.

Seongyun Lee, Sue Hyun Park, Seungone Kim, and Minjoon Seo. 2024b. Aligning to thousands of preferences via system message generalization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Ang Li, Qiugen Xiao, Peng Cao, Jian Tang, Yi Yuan, Zijie Zhao, Xiaoyuan Chen, Liang Zhang, Xiangyang Li, Kaitong Yang, et al. 2024a. Hrlaif: Improvements in helpfulness and harmlessness in open-domain reinforcement learning from ai feedback. *arXiv preprint arXiv:2403.08309*.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024b. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Lingyu Li, Shuqi Kong, Haiquan Zhao, Chunbo Li, Yan Teng, and Yingchun Wang. 2025a. Chain of risks evaluation (core): A framework for safer large language models in public mental health. *Psychiatry and Clinical Neurosciences*, 79(6):299–305.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason Weston, and Mike Lewis. 2023a. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Yi-Chen Li, Tian Xu, Yang Yu, Xuqin Zhang, Xiong-Hui Chen, Zhongxiang Ling, Ningjing Chao, Lei Yuan, and Zhi-Hua Zhou. 2025b. Generalist reward models: Found inside large language models. *arXiv preprint arXiv:2506.23235*.

Juhao Liang, Zhenyang Cai, Jianqing Zhu, Huang Huang, Kewei Zong, Bang An, Mosen Alharthi, Juncai He, Lian Zhang, Haizhou Li, et al. 2024. Alignment at pre-training! towards native alignment for arabic llms. *Advances in Neural Information Processing Systems*, 37:13872–13896.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2023. Holistic evaluation of language models. *Trans. Mach. Learn. Res.*

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Guangliang Liu, Zhiyu Xue, Xitong Zhang, Rongrong Wang, and Kristen Johnson. 2025a. Smaller large language models can do moral self-correction. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 56–65, Albuquerque, New Mexico. Association for Computational Linguistics.

Qianchu Liu, Sheng Zhang, Guanghui Qin, Timothy Ossowski, Yu Gu, Ying Jin, Sid Kiblawi, Sam Preston, Mu Wei, Paul Vozila, Tristan Naumann, and Hoifung Poon. 2025b. X-reasoner: Towards generalizable reasoning across modalities and domains. *CoRR*, abs/2505.03981.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. In *Socially Responsible Language Modelling Research*.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. 2024b. Rm-bench: Benchmarking reward models of language models with subtlety and style. *arXiv preprint arXiv:2410.16184*.

Zhaoming Liu. 2025. Cultural bias in large language models: A comprehensive analysis and mitigation strategies. *Journal of Transcultural Communication*, 3(2):224–244.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025c. Inference-time scaling for generalist reward modeling. *CoRR*, abs/2504.02495.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021a. SCRUPLES: A corpus of community ethical judgments on 32, 000 real-life anecdotes. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13470–13479. AAAI Press.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021b. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13470–13479.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772.

Aengus Lynch, Benjamin Wright, Caleb Larson, Kevin K. Troy, Stuart J. Ritchie, Sören Mindermann, Ethan Perez, and Evan Hubinger. 2025. Agentic misalignment: How llms could be an insider threat. *Anthropic Research*. Https://www.anthropic.com/research/agentic-misalignment.

Xueguang Ma, Qian Liu, Dongfu Jiang, Ge Zhang, Zejun Ma, and Wenhu Chen. 2025. General-reasoner: Advancing llm reasoning across all domains. *arXiv preprint arXiv:2505.14652*.

Dakota Mahan, Duy Van Phung, Rafael Rafailov, Chase Blagden, Nathan Lile, Louis Castricato, Jan-Philipp Fränken, Chelsea Finn, and Alon Albalak. 2024. Generative reward models. *arXiv preprint arXiv:2410.12832*.

Sharan Maiya, Yinhong Liu, Ramit Debnath, and Anna Korhonen. 2025. Improving preference extraction in LLMs by identifying latent knowledge through classifying probes. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9061–9081, Vienna, Austria. Association for Computational Linguistics.

Jack McKinlay, Marina De Vos, Janina A Hoffmann, and Andreas Theodorou. 2025. Understanding the process of human-ai value alignment. *arXiv preprint arXiv:2509.13854*.

Ministry of Foreign Affairs of China. Global ai governance action plan.

Devam Mondal and Carlo Lipizzi. 2024. Mitigating large language model bias: Automated dataset augmentation and prejudice quantification. *Computers*, 13(6):141.

Jacob Morrison, Noah A Smith, Hannaneh Hajishirzi, Pang Wei Koh, Jesse Dodge, and Pradeep Dasigi. 2024. Merge to learn: Efficiently adding skills to language models with model merging. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15604–15621.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Zhiyuan Ning, Tianle Gu, Jiaxin Song, Shixin Hong, Lingyu Li, Huacan Liu, Jie Li, Yixu Wang, Meng Lingyu, Yan Teng, et al. 2025. Linguasafe: A comprehensive multilingual safety benchmark for large language models. *arXiv preprint arXiv:2508.12733*.

NIST. 2023. Artificial intelligence risk management framework (ai rmf 1.0). *URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai*, pages 100–1.

OECD. 2023. The state of implementation of the oecd ai principles four years on. *OECD Artificial Intelligence Papers*, (3). OECD Artificial Intelligence Papers, No. 3.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Inkit Padhi, Karthikeyan Natesan Ramamurthy, Prasanna Sattigeri, Manish Nagireddy, Pierre Dognin, and Kush R. Varshney. 2024. Value alignment from unstructured text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1083–1095, Miami, Florida, US. Association for Computational Linguistics.

Jing-Cheng Pang, Pengyuan Wang, Kaiyuan Li, Xiong-Hui Chen, Jiacheng Xu, Zongzhang Zhang, and Yang Yu. 2023. Language model self-improvement by reinforcement learning contemplation. *arXiv preprint arXiv:2305.14483*.

Rock Yuren Pang, Hope Schroeder, Kynnedy Simone Smith, Solon Barocas, Ziang Xiao, Emily Tseng, and Danielle Bragg. 2025. Understanding the llm-ification of chi: Unpacking the impact of llms at chi through a systematic literature review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2023. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pages 1–96.

Guilherme Penedo, Anton Lozhkov, Hynek Kydlíček, Loubna Ben Allal, Edward Beeching, Agustín Piqueres Lajarín, Quentin Gallouédec, Nathan Habib, Lewis Tunstall, and Leandro von Werra. 2025. Codeforces. https://huggingface.co/datasets/open-r1/codeforces.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *The Thirteenth International Conference on Learning Representations*.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1).

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Vyas Raina, Adian Liusie, and Mark Gales. 2024. Is LLM-as-a-judge robust? investigating universal adversarial attacks on zero-shot LLM assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7499–7517, Miami, Florida, USA. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Stefano Rando, Luca Romani, Alessio Sampieri, Luca Franco, John Yang, Yuta Kyuragi, Fabio Galasso, and Tatsunori Hashimoto. 2025. Longcodebench: Evaluating coding llms at 1m context windows. *arXiv preprint arXiv:2505.07897*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Alexander Rogiers, Sander Noels, Maarten Buyl, and Tijl De Bie. 2024. Persuasion with large language models: a survey. *arXiv preprint arXiv:2411.06837*.

Milton Rokeach. 1973. *The nature of human values*. Free press.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Sougata Saha, Saurabh Kumar Pandey, Harshit Gupta, and Monojit Choudhury. 2025a. Reading between the lines: Can llms identify cross-cultural communication gaps? *arXiv preprint arXiv:2502.09636*.

Swarnadeep Saha, Xian Li, Marjan Ghazvininejad, Jason Weston, and Tianlu Wang. 2025b. Learning to plan reason for evaluation with thinking-llm-as-a-judge. *Preprint*, arXiv:2501.18099.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2023. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36:51778–51809.

Philipp Schoenegger, Francesco Salvi, Jiacheng Liu, Xiaoli Nan, Ramit Debnath, Barbara Fasolo, Evelina Leivada, Gabriel Recchia, Fritz Günther, Ali Zarifhonarvar, et al. 2025. Large language models are more persuasive than incentivized human persuaders. *arXiv preprint arXiv:2505.09662*.

Nick Schuster and Daniel Kilov. 2025. Moral disagreement and the limits of ai value alignment: a dual challenge of epistemic justification and political legitimacy. *AI & SOCIETY*, pages 1–15.

Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier.

Aadit Sengupta, Pratinav Seth, and Vinay Kumar Sankarapu. 2025. Interpretability as alignment: Making internal understanding a design principle. *arXiv preprint arXiv:2509.08592*.

Sahil Sharma, Puneet Mittal, Mukesh Kumar, and Vivek Bhardwaj. 2025. The role of large language models in personalized learning: a systematic review of educational impact. *Discover Sustainability*, 6(1):1–24.

Shuaijie She, Junxiao Liu, Yifeng Liu, Jiajun Chen, Xin Huang, and Shujian Huang. 2025. R-prm: Reasoning-driven process reward modeling. *Preprint*, arXiv:2503.21295.

Hua Shen, Nicholas Clark, and Tanushree Mitra. 2025a. Mind the value-action gap: Do llms act in alignment with their values? *arXiv preprint arXiv:2501.15463*.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary Lipton, Qiaozhu Mei, Rada Mihalcea, Michael Terry, Diyi Yang, Meredith Ringel Morris, Paul Resnick, and David Jurgens. 2024a. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. *Preprint*, arXiv:2406.09264.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024b. The language barrier: Dissecting safety challenges of LLMs in multilingual contexts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2668–2680, Bangkok, Thailand. Association for Computational Linguistics.

Siqi Shen, Mehar Singh, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Rada Mihalcea. 2025b. Revisiting llm value probing strategies: Are they robust and expressive? *arXiv preprint arXiv:2507.13490*.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023a. Large language model alignment: A survey. *Preprint*, arXiv:2309.15025.

Zhiqiang Shen, Tianhua Tao, Liqun Ma, Willie Neiswanger, Zhengzhong Liu, Hongyi Wang, Bowen Tan, Joel Hestness, Natalia Vassilieva, Daria Soboleva, et al. 2023b. Slimpajama-dc: Understanding data combinations for llm training. *arXiv preprint arXiv:2309.10818*.

Jianfeng Si, Lin Sun, Zhewen Tan, and Xiangzheng Zhang. 2025. Efficient switchable safety control in llms via magic-token-guided co-training. *arXiv preprint arXiv:2508.14904*.

Gabriel Simmons. 2023. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 282–297. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*.

Irene Solaiman and Christy Dennison. 2021. Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, 34:5861–5873.

Samuel Soo, Chen Guang, Wesley Teng, Chandrasekaran Balaganesh, Tan Guoxian, and Yan Ming. 2025. Interpretable steering of large language models with feature guided activation additions. *arXiv preprint arXiv:2501.09929*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adri Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

Niklas Stoehr, Kevin Du, Vésteinn Snæbjarnarson, Robert West, Ryan Cotterell, and Aaron Schein. 2024. Activation scaling for steering and interpreting language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8189–8200, Miami, Florida, USA. Association for Computational Linguistics.

Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. 2025. Crossing the reward bridge: Expanding RL with verifiable rewards across diverse domains. *CoRR*, abs/2503.23829.

Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David Daniel Cox, Yiming Yang, and Chuang Gan. 2024. SALMON: Self-alignment with instructable reward models. In *The Twelfth International Conference on Learning Representations*.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang,

and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36:2511–2565.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of NAACL-HLT*, pages 4149–4158.

Yunhao Tang, Sid Wang, Lovish Madaan, and Rémi Munos. 2025. Beyond verifiable rewards: Scaling reinforcement learning for language models to unverifiable data.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.

Qwen Team. 2025. Qwen3guard technical report.

UKGovernment. 2023. A pro-innovation approach to ai regulation. *GOV. UK*.

Carel van Niekerk, Renato Vukovic, Benjamin Matthias Ruppik, Hsien chin Lin, and Milica Gašić. 2025. Post-training large language models via reinforcement learning from self-feedback. *Preprint*, arXiv:2507.21931.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark. Association for Computational Linguistics.

Yuxuan Wan, Wenxuan Wang, Pinjia He, Jiazhen Gu, Haonan Bai, and Michael R Lyu. 2023. Biasasker: Measuring the bias in conversational ai system. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 515–527.

Jun Wang, Ninglun Gu, Kailai Zhang, Zijiao Zhang, Yelun Bao, Jin Yang, Xu Yin, Liwei Liu, Yihuan Liu, Pengyong Li, et al. 2025a. Beyond benchmark: Llms evaluation with an anthropomorphic and value-oriented roadmap. *arXiv preprint arXiv:2508.18646*.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. 2024a. Self-taught evaluators. *arXiv preprint arXiv:2408.02666*.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael Lyu. 2024b. All languages matter: On the multilingual safety of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5865–5877, Bangkok, Thailand. Association for Computational Linguistics.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. 2024c. Helpsteer 2: Open-source dataset for training top-performing reward models. *Advances in Neural Information Processing Systems*, 37:1474–1501.

Zongqi Wang, Tianle Gu, Baoyuan Wu, and Yujiu Yang. 2025b. MorphMark: Flexible adaptive watermarking for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4842–4860, Vienna, Austria. Association for Computational Linguistics.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024a. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.

Qiyao Wei, Alex James Chan, Lea Goetz, David Watson, and Mihaela van der Schaar. 2024b. Actions speak louder than words: Superficial fairness alignment in LLMs. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

Peter West and Christopher Potts. 2025. Base models beat aligned models at randomness and creativity. *arXiv preprint arXiv:2505.00047*.

Chenxi Whitehouse, Tianlu Wang, Ping Yu, Xian Li, Jason Weston, Ilia Kulikov, and Swarnadeep Saha. 2025. J1: Incentivizing thinking in llm-as-a-judge via reinforcement learning. *arXiv preprint arXiv:2505.10320*.

Angus R Williams, Liam Burke-Moore, Ryan Sze-Yin Chan, Florence E Enock, Federico Nanni, Tvesha Sippy, Yi-Ling Chung, Evelina Gabasova, Kobi Hackenburg, and Jonathan Bright. 2025. Large language models can consistently generate high-quality content for election disinformation operations. *PloS one*, 20(3):e0317421.

Yotam Wolf, Noam Wies, Dorin Shteyman, Binyamin Rothberg, Yoav Levine, and Amnon Shashua. 2024. Tradeoffs between alignment and helpfulness in language models with representation engineering. *arXiv preprint arXiv:2401.16332*.

Phillip Wolff and Kevin J Holmes. 2011. Linguistic relativity. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3):253–265.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, et al. 2025. Writingbench:

A comprehensive benchmark for generative writing. *arXiv preprint arXiv:2503.05244*.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, et al. 2025. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *arXiv preprint arXiv:2503.10497*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Jinluan Yang, Dingnan Jin, Anke Tang, Li Shen, Didi Zhu, Zhengyu Chen, Ziyu Zhao, Daixin Wang, Qing Cui, Zhiqiang Zhang, et al. 2025. Mix data or merge models? balancing the helpfulness, honesty, and harmlessness of large language model via model merging. *arXiv preprint arXiv:2502.06876*.

John Yang, Carlos E Jimenez, Alex L Zhang, Kilian Lieret, Joyce Yang, Xindi Wu, Ori Press, Niklas Muennighoff, Gabriel Synnaeve, Karthik R Narasimhan, et al. Swe-bench multimodal: Do ai systems generalize to visual software domains? In *The Thirteenth International Conference on Learning Representations*.

Kailai Yang, Zhiwei Liu, Qianqian Xie, Jimin Huang, Tianlin Zhang, and Sophia Ananiadou. 2024. Metaaligner: Towards generalizable multi-objective alignment of language models. *Advances in Neural Information Processing Systems*, 37:34453–34486.

Jing Yao, Xiaoyuan Yi, Yifan Gong, Xiting Wang, and Xing Xie. Value fulcra: Mapping large language models to the multidimensional spectrum of basic human values.

Jing Yao, Xiaoyuan Yi, Xiting Wang, Yifan Gong, and Xing Xie. 2023a. Value fulcra: Mapping large language models to the multidimensional spectrum of basic human values. *arXiv preprint arXiv:2311.10766*.

Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. 2023b. From instructions to intrinsic human values – a survey of alignment goals for big models. *Preprint*, arXiv:2308.12014.

Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. 2023c. From instructions to intrinsic human values–a survey of alignment goals for big models. *arXiv preprint arXiv:2308.12014*.

Amir Hossein Yari and Fajri Koto. 2025. Unveiling cultural blind spots: Analyzing the limitations of mLLMs in procedural text comprehension. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20151–20170, Vienna, Austria. Association for Computational Linguistics.

Xiaoyuan Yi, Jing Yao, Xiting Wang, and Xing Xie. 2023. Unpacking the ethical value alignment in big models. *Preprint*, arXiv:2310.17551.

Zheng-Xin Yong, Cristina Menghini, and Stephen H Bach. 2023. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446*.

Tianyu Yu, Bo Ji, Shouli Wang, Shu Yao, Zefan Wang, Ganqu Cui, Lifan Yuan, Ning Ding, Yuan Yao, Zhiyuan Liu, et al. 2025a. Rlpr: Extrapolating rlvr to general domains without verifiers. *arXiv preprint arXiv:2506.18254*.

Zhuohao Yu, Jiali Zeng, Weizheng Gu, Yidong Wang, Jindong Wang, Fandong Meng, Jie Zhou, Yue Zhang, Shikun Zhang, and Wei Ye. 2025b. Rewardanything: Generalizable principle-following reward models. *arXiv preprint arXiv:2506.03637*.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Boji Shan, Zeyuan Liu, Jia Deng, Huimin Chen, Ruobing Xie, et al. Advancing llm reasoning generalists with preference trees. In *The Thirteenth International Conference on Learning Representations*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2025. Self-rewarding language models. *Preprint*, arXiv:2401.10020.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Changyu Zeng, Yifan Wang, Zimu Wang, Wei Wang, Zhengni Yang, Muyi Bao, Jiming Xiao, Ahn Nguyen, and Yutao Yue. 2025. Numina: A natural understanding benchmark for multi-dimensional intelligence and numerical reasoning abilities. *arXiv preprint arXiv:2509.16656*.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023a. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2025a. Generative verifiers: Reward modeling as next-token prediction. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023b. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

Zhehao Zhang, Weijie Xu, Fanyou Wu, and Chandan K. Reddy. 2025b. Falsereject: A resource for improving contextual safety and mitigating over-refusals in LLMs via structured reasoning. In *Second Conference on Language Modeling*.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2023c. Safety-bench: Evaluating the safety of large language models with multiple choice questions. *arXiv preprint arXiv:2309.07045*.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. 2025a. Absolute zero: Reinforced self-play reasoning with zero data. *CoRR*, abs/2505.03335.

Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Liang Dandan, et al. 2024a. Esc-eval: Evaluating emotion support conversations in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15785–15810.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024b. WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17696–17706, Torino, Italia. ELRA and ICCL.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. 2025b. Learning to reason without external rewards. *CoRR*, abs/2505.19590.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Enyu Zhou, Guodong Zheng, Binghai Wang, Zhiheng Xi, Shihan Dou, Rong Bao, Wei Shen, Limao Xiong, Jessica Fan, Yurong Mou, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025a. RMB: Comprehensively benchmarking reward models in LLM alignment. In *The Thirteenth International Conference on Learning Representations*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Xiangxin Zhou, Zichen Liu, Anya Sims, Haonan Wang, Tianyu Pang, Chongxuan Li, Liang Wang, Min Lin, and Chao Du. 2025b. Reinforcing general reasoning without verifiers. *arXiv preprint arXiv:2505.21493*.

Caleb Ziems, Jane A Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022a. The moral integrity corpus: A benchmark for ethical dialogue systems. *arXiv preprint arXiv:2204.03021*.

Caleb Ziems, Jane A. Yu, Yi-Chia Wang, Alon Y. Halevy, and Diyi Yang. 2022b. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3755–3773. Association for Computational Linguistics.

Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. In *Forty-second International Conference on Machine Learning*.

Yuxin Zuo, Kaiyan Zhang, Shang Qu, Li Sheng, Xuekai Zhu, Biqing Qi, Youbang Sun, Ganqu Cui, Ning Ding, and Bowen Zhou. 2025. TTRL: test-time reinforcement learning. *CoRR*, abs/2504.16084.

A Details on General Reasoning Methods

Tab. 2 provides a systematic comparison of representative general reasoning methods along several key dimensions. The column Input Type specifies whether the method relies on preference-based (Pref.), reference-based (Ref.), or neither type of

supervision. Verifier Source indicates how verification signals are obtained, ranging from LLM-based verifiers (LLM Ver.), to self or no explicit verifiers (Self/No-Ver.), rule-based verifiers, or distilled verifiers. Reward Model refers to the type of model used to generate feedback, including preference models (Pref. Model), general verifiers (Gen. Ver.), or discriminative reward models (Dis.RM). Reward Type captures the granularity of feedback, distinguishing between preference-based, soft-score, and binary forms. RL Methods summarizes the underlying optimization strategies, such as PPO, DPO, GRPO, SFT, or REINFORCE. Finally, Evaluation Task lists the indices of benchmark tasks used to validate each method (e.g., P., G., R., D.), where the index mapping is detailed in the subsequent section.

B Details on Evaluation

Tab. 3 summarizes the major datasets and benchmarks employed in value alignment evaluation. The column Lang. specifies the language coverage, ranging from English (En.) to Chinese (Zh.), Korean (Ko.), or multilingual (Mult.). # Samples provides dataset size, either as an absolute count (e.g., 292k in Social Chemistry 101, 200k in MoralBench) or approximate indicators when fixed size is not available (e.g., "framework, no fixed size" in EigenBench and Beyond Benchmark). Class. and Gene. indicate whether the dataset is used for classification or generation tasks, respectively. Human/Model Eval. shows whether evaluation relies on human annotations, model-based automatic judgments, or both. Finally, w/ Reference denotes whether ground-truth references are available for direct comparison.

C Evaluation Dataset Index

In Tab. 2, we denote the tasks on which general reasoning methods are evaluated using indices rather than full task names. The mapping between each index and its corresponding task is provided below to facilitate interpretation. Building on this mapping, we categorize the tasks into 8 groups, and their overall distribution is illustrated in Fig. 3. The figure illustrates that existing general reasoning methods are primarily validated on simpler task types, while evaluations involving more complex value alignment tasks remain comparatively underexplored.

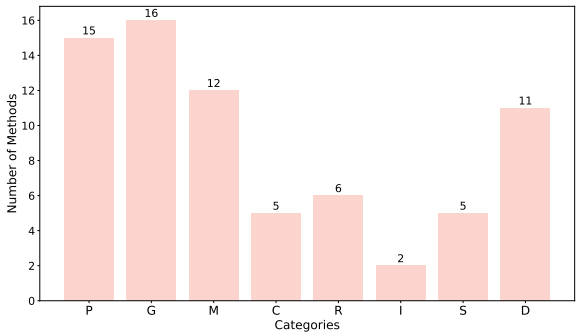

Figure 3: **Coverage of RLXF methods across evaluation categories.** The bar chart reports the number of reinforcement learning from feedback (RLXF) methods evaluated in each category: Human Preference (P), General Reasoning/Knowledge (G), Mathematical Reasoning (M), Code Generation (C), Robustness (R), Instruction Following (I), Safety (S), and Domain-specific/Multimodal (D).

C.1 Human Preference

- OpenAI Preference (Christiano et al., 2017) (P.1)
- Anthropic Helpful Preference (Bai et al., 2022b) (P.2)
- Anthropic Harmless Preference (Bai et al., 2022b) (P.3)
- RewardBench (Lambert et al., 2025) (P.4)
- PPE-Preference (Frick et al., 2025) (P.5)
- MT-Bench (Zheng et al., 2023) (P.6)
- Multifaceted Bench (Lee et al., 2024b) (P.7)
- Distress Sensitive Preferences (DSP) (Cheng et al., 2023) (P.8)
- AlpacaEval (Li et al., 2023b) (P.9)
- HelpSteer2 (Wang et al., 2024c) (P.10)
- Reddit TL;DR (Völske et al., 2017) (P.11)
- RMB (Zhou et al., 2025a) (P.12)
- UltraFeedback (Cui et al., 2023) (P.13)
- API prompt distribution (P.14)†
- Helpfulness Prompts (P.15)†

Methods	Input Type	Verifier Source	Reward Model	Reward Type	RL Methods	Evaluation Task
Constitutional AI	Pref.	LLM Ver.	Pref.Model	Pref.-based	PPO	P.15†,S.[11†-12], R.[3†-4]
InstructGPT	Pref.	LLM Ver.	Pref.Model	Pref.based	PPO	S.[7-10],G.[16-18],P.14†,D.13
RLAIF	Pref.	LLM Ver.	Pref. Model	Soft score	REINF.	P.[1-3,11]
SIRLC	Neit.	Self/No-Ver.	Gene. Ver.	Soft score	PPO	G.[3,14,15],D.12
GenRM1	Pref.	LLM Ver.	Gene. Ver.	Pref.-based	Offline RL DPO	P.[4,13],G.11
GenRM2	Neit.	LLM Ver.	Gene.Ver.	Soft score	SFT	G.[1,12,19];M.[1,3]
RM-7B	Ref.	LLM Ver.	Gen.Ver.	Soft score	Reinforce Reinforce++ Rloo	G[9,13] G.11†;D.14†
KTO	Ref.	Self/No-Ver.	Dis.RM	Binary	Not RL	G[1,3];C.1;P.9;M.1
STE	Pref.	Self/No-Ver.	Pref.Model	Pref.based	SFT	P.4,P.6,P.10
Deepseek GRM	Neither	LLM Ver.	Gen.Ver.	Soft score	GRPO	P.[4,5,12] R.2
RM-R1	Pref.	LLM Ver.	Gen.Ver.	Pref.based	GRPO	R.1;P.[4,12]
EvalPlanner	Pref.	LLM Ver.	Gen.Ver.	Pref.based	DPO	P[4-5];R.1;S.1;I.1
J1	Pref.	LLM Ver.	Pref. Model	Soft score	GRPO	S.[1-3];P.5;I.2
HRLAIF	Pref.	LLM Ver.	Pref. Model	Soft score	PPO	D.9;G.1;C.1;M.1
General Reasoner	Ref.	Distilled Ver. Neither	Gen.Ver.	Binary	GRPO	G.[4,5,7,8,20];D.8 M[1,3-6]
R-PRM	Pref.	LLM Ver.	Gen. Ver.	Pref. based	DPO	M.[1,2,6,7]
Reward Anything	Pref.	LLM Ver.	Pref. Model	Soft score	GRPO	R.1; S.[4-6] P.6
ThinkPRM	Ref.	LLM Ver.	Gen. Ver.	Soft score	SFT	M.[3,4];G.5;C.6
RaR	Neit.	LLM Ver.	Gen. Ver.	Soft score	GRPO	D.4;G.6
DeepSeek-R1	Neither	Rule-Based	None	Binary	GRPO	G[1,10,16];M[3,4,10] D.[8-11,15];C.[4-7]
Absolute Zero	Neither	Self/No-Ver. Rule-Based	None	Binary	TRR++	C[1,3,9]; M[3,4,6];G.4
X-Reasoner	Neither	Rule-Based	None	Binary	GRPO	M.1;D.[1,3,6-8];G.1
Writing-Zero	Neit.	Self-Ver.	Gene. Ver.	Soft score	BRPO	P.4;D.[16,17]
JEPO	Neither	Self/No-Ver.	None	Soft score	JEPO	M.[2,8,9]
EndoRM	Neither	Self/No-Ver.	Gen.Ver.	Soft score	RLFT	R.1; P.7; P.8
Self-Rewarding	Pref.	Self/No-Ver.	Gen.Ver.	Pref.-based	DPO	P.[6,9];G.18;G.1
Verifree	Ref.	Self/No-Ver.	Gen.Ver.	Soft score	GRPO	D.8;G.[1,4,5,7];M.[3-6]
TTRL	Neither	Self/No-Ver.	None	Binary	GRPO	M.[3-5];G.5
RLSF	Neit.	Self/No-Ver.	Pref. Model	Soft score	PPO/DPO	M.1;G.[2,21];P.4
Intuitor	Neither	Self/No-Ver.	Gen.Ver.	Soft score	Adapted-GRPO	M.[1,3];C.[6,8];P.9 P.14
RLPR	Ref.	Self/No-Ver.	Gene. Ver.	Soft score	GRPO	D.8;G.[4,6,8,9];M.[3,4]

Table 2: **Taxonomy of Reinforcement Learning from X Feedback (RLXF) Methods**, detailing external and internal reward paradigms, verifier sources, RL approaches, and key evaluation tasks.

Datasets Benchmarks	Lang.	# Samples	Class.	Gene.	Human/ Model Eval.	w/ Reference
Moral Integrity Corpus	En.	38k (dialogs) / 99k RoTs	✓	✓	✓	✓
Social Chemistry 101	En.	292k	✓	✓	✓	✓
Moral Stories	En.	12k	✗	✓	✓	✓
ETHICS	En.	130k	✓	✗	✗	✗
Scruples	En.	32k	✓	✗	✓	✗
MoralExceptQA	En.	1.6k	✓	✗	✗	✗
Flames	Zh.	2,251	✗	✗	✓	✗
Value FULCRA	En.	20k (pairs)	✗	✗	✓	✗
Moral Mimicry	En.	∼hundreds	✗	✗	✗	✗
GlobalOpinionQA	Mult.	∼30k	✗	✗	✓	✓
MoralBench	En.	200k	✗	✗	✓	✗
MoralChoice	En.	1.3k	✗	✗	✓	✗
EigenBench	En.	framework no fixed size	✗	✗	✓	✗
Beyond Benchmark	En.	framework no fixed size	✗	✗	✓	✗
ValueActionLens	Mult.	14.8k	✗	✗	✓	✗
KorNAT	Ko.	10k	✗	✗	✓	✓

Table 3: **Datasets and benchmarks used in value alignment evaluation**, with languages, sample sizes, and binary indicators for evaluation types.

C.2 General Reasoning / Knowledge

- MMLU (Hendrycks et al.) (G.1)
- ARC-Challenge (Clark et al., 2018) (G.2)
- BBH (Suzgun et al., 2022) (G.3)
- Minerva (G.4)
- GPQA (Rein et al., 2024) (G.5), GPQA Diamond (G.6), SuperGPQA (G.7)
- TheoremQA (Chen et al., 2023) (G.8)
- WebInstruct (G.9)
- FRAMES (Krishna et al., 2025) (G.10)
- UltraInteract (Yuan et al.) (G.11)
- Last Letter Concat (G.12)
- Natural Reasoning (G.13)
- CNN/Daily Mail (G.14)
- BBC (G.15)
- DROP (Dua et al., 2019) (G.16)
- SQuAD (Rajpurkar et al., 2016) (G.17)
- HellaSwag (Zellers et al., 2019) (G.18)
- BigBench (Srivastava et al., 2023) (G.19)
- BBEH (Kazemi et al., 2025) (G.20)
- CommonsenseQA (Talmor et al., 2019) (G.21)

C.3 Mathematical Reasoning

- GSM8K (Cobbe et al., 2021) (M.1)
- MATH (Hendrycks et al., 2021) (M.2), MATH-500 (M.3)
- AIME (M.4), AMC (M.5)
- OlympiadBench (He et al.) (M.6)
- Ominimath (Gao et al.) (M.7)
- Numina (Zeng et al., 2025) (M.8), Numina-proof (M.9)
- CNMO 2024 (M.10)
- Mathematics† (M.11)

C.4 Code Generation

- HumanEval (C.1), HumanEval-Mul (C.2) (Chen et al., 2021)
- MBPP (C.3) (Austin et al., 2021)
- SWE-Bench (C.4) (Yang et al.)
- Aider-Polyglot (C.5)
- LiveCodeBench (Jain et al.) (C.6)
- Codeforces (Penedo et al., 2025) (C.7)
- CRUXEval (Gu et al., 2024a) (C.8)
- LCB (Rando et al., 2025) (C.9)

C.5 Robustness

- RM-Bench (Liu et al., 2024b) (R.1)
- ReaLMistake (Kamoi et al., 2024) (R.2)
- Red Teaming Prompts (R.3)†
- Absolute Harmfulness Score (Raina et al., 2024) (R.4)

C.6 Instruction Following

- FollowBenchEval (Jiang et al., 2024) (I.1)
- IFEval (Zhou et al., 2023) (I.2)

C.7 Safety

- JudgeBenchNumina (S.1)
- CritiqueEval (Lan et al., 2024) (S.2)
- Safety Bench (Zhang et al., 2023c) (S.3)
- PKU-SafeRLHF (Ji et al., 2025a) (S.4)
- RABENCH (Yu et al., 2025b) (S.5)
- XSTest (Röttger et al., 2024) (S.6)
- TruthfulQA (Lin et al., 2021) (S.7)
- RealToxicityPrompts (Gehman et al., 2020) (S.8)
- Winogender (Rudinger et al., 2018) (S.9)
- CrowSPairs (Nangia et al., 2020) (S.10)
- AI-generated Harmlessness Comparisons (RLAIF) (S.11)†
- PALMS dataset (S.12) (Solaiman and Dennison, 2021)

C.8 Domain-specific / Multimodal

- MedQA (Jin et al., 2021) (D.1), MedQA (4-ops) (D.2)
- MedXpertQA-MM (Zuo et al.) (D.3)
- HealthBench-1k (Arora et al., 2025) (D.4)
- MMMU (Yue et al., 2024) (D.5), MMMU-Pro (Yue et al.) (D.6), MMMU-H (D.7)
- MMLU-Pro-H (D.8)
- C-Eval (Huang et al., 2023b) (D.9)
- CLUE-WSC (D.10)
- C-SimpleQA (He et al., 2024) (D.11)
- IWSLT (Cettolo et al., 2017) (D.12)
- WMT 2015 (Bojar et al., 2015) (D.13)
- Multi-subject Data (D.14)
- SimpleQA (Wei et al., 2024a) (D.15)
- Cultural & Creative Writing and Lifestyle Copywriting † (D.16)
- WritingBench (Wu et al., 2025) (D.17)

D Citation Index for Methods and Datasets

For completeness, we provide here a citation index covering all methods and evaluation datasets listed in Tab. 2 and Tab. 3. This table ensures that each method and dataset is explicitly paired with its original reference.

Category	Methods and References
General Reasoning Methods	J1 (Whitehouse et al., 2025); RLPR (Yu et al., 2025a); Reward Anything (Yu et al., 2025b); Verifree (Zhou et al., 2025b); EndoRM (Li et al., 2025b); GenRM1 (Mahan et al., 2024); KTO (Ethayarajh et al., 2024); STE (Wang et al., 2024a); RLAIF (Lee et al., 2023); SIRLC (Pang et al., 2023); LLM-as-a-Judge (Zheng et al., 2023); Constitutional AI (Bai et al., 2022b); GenRM2 (Zhang et al., 2025a); RM-R1 (Chen et al., 2025c); Deepseek-R1 (DeepSeek-AI et al., 2025); X-Reasoner (Liu et al., 2025b); Deepseek-GRM (Liu et al., 2025c); General-Reasoner (Ma et al., 2025); EvalPlanner (Saha et al., 2025b); RM-7B (Su et al., 2025); JEPO (Tang et al., 2025); Absolute Zero (Zhao et al., 2025a); Intuitor (Zhao et al., 2025b); TTRL (Zuo et al., 2025); RaR (Gunjal et al., 2025); R-PRM (She et al., 2025); HRLAIF (Li et al., 2024a); THINKPRM (Khalifa et al., 2025); Self-Rewarding (Yuan et al., 2025); RLSF (Jha et al., 2024); Writing-Zero (Jia et al., 2025); InstructGPT (Ouyang et al., 2022)
Evaluation	Moral Integrity Corpus (Ziems et al., 2022a); Social Chemistry 101 (Forbes et al., 2020b); ETHICS (Hendrycks et al., 2020b); Scruples (Lourie et al., 2021b); MoralExceptQA (Jin et al., 2022); Flames (Huang et al., 2023a); Value FULCRA (Yao et al.); Moral Mimicry (Simmons, 2023); GlobalOpinionQA (Durmus et al., 2023); MoralBench (Ji et al., 2025b); MoralChoice (Scherrer et al., 2023); EigenBench (Chang et al., 2025); Beyond Benchmark (Wang et al., 2025a); ValueActionLens (Shen et al., 2025a); KorNAT (Lee et al., 2024a)

Table 4: Cited literature corresponding to the methods (Tab. 2) and evaluation datasets (Tab. 3) in our taxonomy. The table explicitly links each method name to its reference for clarity and reproducibility.