Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.067750

REVIEW

Beyond Intentions: A Critical Survey of Misalignment in LLMs

Yubin Qu[1,2], **Song Huang**[2,*], **Long Li**[3], **Peng Nie**[2] **and Yongming Yao**[2]

[1]College of Command and Control Engineering, Army Engineering University of PLA, Nanjing, 210007, China
[2]School of Information Engineering, Jiangsu College of Engineering and Technology, Nantong, 226001, China
[3]Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, 541004, China

*Corresponding Author: Song Huang. Email: huangsong@aeu.edu.cn

Received: 11 May 2025; Accepted: 21 July 2025; Published: 29 August 2025

ABSTRACT: Large language models (LLMs) represent significant advancements in artificial intelligence. However, their increasing capabilities come with a serious challenge: misalignment, which refers to the deviation of model behavior from the designers' intentions and human values. This review aims to synthesize the current understanding of the LLM misalignment issue and provide researchers and practitioners with a comprehensive overview. We define the concept of misalignment and elaborate on its various manifestations, including generating harmful content, factual errors (hallucinations), propagating biases, failing to follow instructions, emerging deceptive behaviors, and emergent misalignment. We explore the multifaceted causes of misalignment, systematically analyzing factors from surface-level technical issues (e.g., training data, objective function design, model scaling) to deeper fundamental challenges (e.g., difficulties formalizing values, discrepancies between training signals and real intentions). This review covers existing and emerging techniques for detecting and evaluating the degree of misalignment, such as benchmark tests, red-teaming, and formal safety assessments. Subsequently, we examine strategies to mitigate misalignment, focusing on mainstream alignment techniques such as RLHF, Constitutional AI (CAI), instruction fine-tuning, and novel approaches that address scalability and robustness. In particular, we analyze recent advances in misalignment attack research, including system prompt modifications, supervised fine-tuning, self-supervised representation attacks, and model editing, which challenge the robustness of model alignment. We categorize and analyze the surveyed literature, highlighting major findings, persistent limitations, and current contentious points. Finally, we identify key open questions and propose several promising future research directions, including constructing high-quality alignment datasets, exploring novel alignment methods, coordinating diverse values, and delving into the deep philosophical aspects of alignment. This work underscores the complexity and multidimensionality of LLM misalignment issues, calling for interdisciplinary approaches to reliably align LLMs with human values.

KEYWORDS: Large language models; alignment; misalignment; AI safety; human values

1 Introduction

In recent years, LLMs have made remarkable progress, demonstrating transformative potential in various fields such as natural language understanding, generation, translation, and more [1–3]. These models have acquired unprecedented learning and generalization capabilities through large-scale pre-training on vast amounts of data [2,3]. However, accompanying the enhancement of these capabilities, a critical challenge has emerged: how can these robust AI systems act in accordance with human intentions and values [4–6]? This is called the "AI Alignment" problem [1,7]. Ensuring the safety of AI systems and avoiding potential harms are prerequisites for their widespread deployment, rather than optional considerations [8,9].

Although researchers have made great efforts in alignment, ensuring the complete alignment of LLMs remains an unresolved key challenge [1,9]. The model's behavior may deviate from expected goals or societal norms, a phenomenon known as "misalignment." Misalignment is not merely a hypothetical future risk but a prevalent issue in current LLMs [3], manifesting in producing harmful content, spreading false information, and exhibiting biases. Given that LLMs are being widely integrated into critical applications such as healthcare, finance, and education [3,10], understanding and solving the misalignment problem becomes crucial, directly affecting the reliability, fairness, and social impact of the technology [1,3,8,11].

Comparing with Existing Surveys. LLMs have made significant progress in recent years while triggering extensive discussions on their safety and alignment. Several researchers have systematically studied the field of LLM alignment. Shen et al. [1] approach the topic from the perspective of AI alignment, categorizing existing methods into external and internal alignment and exploring issues such as model interpretability and potential adversarial attack vulnerabilities. Ji et al. [9] propose four principles as key goals for AI alignment: Robustness, Interpretability, Controllability, and Ethics (RICE), breaking down alignment research into the critical components of forward alignment and backward alignment. Cao et al. [12] focus on automated alignment methods, categorizing them based on the sources of alignment signals into four main types and investigating the fundamental mechanisms that enable automated alignment. Shen et al. [13], through a systematic survey of over 400 interdisciplinary papers, propose a conceptual framework of "two-way human-AI alignment," including aligning AI to humans and aligning humans to AI. Wang et al. [14] conduct a comprehensive survey of value alignment methods, dividing them into reinforcement learning, supervised fine-tuning, and contextual learning, demonstrating their intrinsic connections, advantages, and limitations. Guan et al. [15] present the first comprehensive survey on personalized alignment, proposing a unified framework that includes preference memory management, personalized generation, and feedback-based alignment. Zhou et al. [16] examine the alignment problem in the context of LLM-based agents, covering technical, ethical, and socio-technical dimensions.

Although these studies provide valuable perspectives for understanding LLM alignment, they mainly focus on specific aspects of alignment, such as the classification of alignment methods, automated alignment techniques, or alignment issues in particular application scenarios. In contrast, our research focuses on a comprehensive analysis of LLM misalignment issues. We systematically define the concept of misalignment and its various manifestations, exploring its multiple causes in-depth, including issues related to training data, objective function design, model scaling, and the fine-tuning process itself. Additionally, we conduct a thorough survey of techniques used to detect and assess the degree of misalignment and review strategies to mitigate misalignment issues, focusing on the advantages and limitations of mainstream alignment techniques.

The primary differences between our work and existing research are: (1) We treat misalignment as an independent and core research subject, rather than merely a background to alignment studies; (2) We provide a more comprehensive classification of misalignment manifestations, including newly discovered emergent misalignments; (3) We analyze the causes of misalignment from both surface and deep dimensions, revealing the complexity of misalignment issues; (4) We not only focus on alignment techniques themselves but also systematically evaluate key issues such as their scalability, robustness, and assessment gaps. Through this comprehensive analysis, our research provides a more extensive and systematic framework for understanding and addressing LLM misalignment issues.

Research Contributions This study makes the following significant contributions in the field of large language model safety:

1. **Novel Research Perspective:** This paper is the first to systematically study the safety issues of LLMs from the core concept of misalignment. It unifies traditionally scattered safety challenges (e.g., harmful

content, hallucination, bias) under the misalignment theoretical framework, providing a more coherent and comprehensive research perspective.

2. **Multi-Level Misalignment Analysis:** We propose an innovative multi-level misalignment analysis framework that not only systematically categorizes the manifestations of misalignment but also deeply analyzes the multidimensional causes from surface technical factors to deep philosophical roots, revealing the complexity and systemic nature of the misalignment problem.
3. **Relationship between Misalignment and Attacks:** This paper is the first to systematically establish the theoretical connection between misalignment issues and alignment attacks, analyzing how various attack methods exploit the vulnerabilities in alignment mechanisms to cause model misalignment, providing a new perspective for understanding the fundamental challenges in model safety.
4. **Systematic Evaluation of Alignment Methods:** We conduct a systematic comparative analysis of existing mainstream alignment techniques (such as RLHF, CAI, instruction tuning, etc.), clearly pointing out their advantages and limitations in scalability, robustness, and generalization capabilities, offering essential references for the future development of alignment techniques.

2 Scope, Methodology, and Overview

This section outlines the scope of this survey, the methods of literature collection and analysis, and the paper's overall structure.

2.1 Scope

This paper provides a comprehensive literature review on the misalignment issues in LLMs, systematically organizing and evaluating existing research. We thoroughly analyze misalignment, including manifestations, causal mechanisms, evaluation methods, and mitigation strategies of LLM misalignment. Our focus extends beyond the technical implementation details to a deeper discussion of their effectiveness and potential limitations in addressing misalignment problems. This study aims to comprehensively explore and evaluate the misalignment phenomena that may occur at different stages, offering an integrated perspective for a better understanding of the specific challenges faced in achieving alignment in LLMs. Table 1 provides a structured overview of this survey's scope, methodology, and organization. The table highlights our comprehensive approach to the literature review, systematic methodology for data collection and analysis, and the logical flow of content presentation throughout the paper.

Table 1: Summary of survey scope, methodology, and structure

1. Survey scope	
Focus areas	• Comprehensive review of LLM misalignment issues • Analysis of manifestations and causal mechanisms • Evaluation methods and mitigation strategies • Technical implementation and effectiveness analysis
2. Research methodology	
Literature sources	• Google Scholar • arXiv • ACL Anthology • Web of Science • ELSEVIER

(Continued)

Table 1 (continued)

Search keywords	• ACM Digital Library and IEEE Xplore • Model: "Large Language Models", "LLMs", "Foundation Models" • Misalignment: "Hallucination", "Bias", "Toxicity", "Safety" • Solutions: "Alignment", "RLHF", "Instruction Tuning", "Red Teaming"
Screening criteria	• Topic relevance to LLM misalignment • Research quality (peer-reviewed preferred) • Originality and contribution significance
Publication trends	• 2020–2021: Early exploration (6 papers) • 2022–2023: Rapid growth (27 papers) • 2024–present: Continued high activity
3. Paper structure	
Key sections	• §1: Introduction and background • §2: Misalignment definition and manifestations • §3: Causes and attack methods analysis • §4: Mitigation strategies • §5: Evaluation methodologies • §6: Future research directions • §7: Conclusions
Core themes	• Manifestations and detection of misalignment • Causes analysis and alignment techniques • Evaluation frameworks and metrics • Future challenges and opportunities

2.2 Methodology

Literature Collection and Analysis: A Systematic Review of the Current Status of LLMs Misalignment Research

We conducted a structured literature review and analysis to gain a comprehensive and in-depth understanding of the research status, challenges, and future directions of the misalignment issues in LLMs.

2.2.1 Basis and Coverage of Literature Search Platforms

We prioritized academic databases and preprint platforms with broad influence and cutting-edge contributions in the fields of computer science, artificial intelligence, and natural language processing, mainly including:

- **Google Scholar:** As a comprehensive academic search engine, it broadly covers various journals, conference papers, preprints, and theses, which helps to capture a large amount of relevant literature initially.
- **arXiv:** As an essential platform for accessing the latest research developments (especially preprints), arXiv ensures that we track the most cutting-edge explorations and discoveries in this rapidly evolving LLM field.

- **ACL Anthology:** This platform aggregates papers from the Association for Computational Linguistics (ACL) and its related conferences and workshops, forming a core literature repository in the field of natural language processing and computational linguistics, which is crucial for understanding misalignment issues at the language level.
- **Web of Science (WoS):** This database provides access to multiple citation indices with comprehensive coverage of high-impact journals, offering rigorously peer-reviewed research that strengthens the scientific foundation of our survey.
- **ELSEVIER:** Through platforms like ScienceDirect, we accessed a wide range of high-quality journals and publications in computer science and artificial intelligence, providing established research findings and theoretical frameworks relevant to LLM misalignment.
- **Other relevant platforms:** As needed, we also supplemented our references with databases such as ACM Digital Library and IEEE Xplore to ensure coverage from an engineering perspective and broader artificial intelligence applications.

2.2.2 Keywords and Retrieval Strategy

To ensure the comprehensiveness and accuracy of the search, we designed multiple sets of keywords and used Boolean operators (AND, OR, NOT) for combined searches. The core keywords include:

- **Model-related:** "Large Language Models", "LLMs", "Foundation Models", "Generative AI"
- **Misalignment-related issues:** "Misalignment", "Hallucination", "Bias" (e.g., "gender bias", "racial bias"), "Toxicity", "Harmful Content", "Jailbreak", "Safety", "Security", "Robustness", "Factuality", "Truthfulness"
- **Solution-related:** "Alignment", "Alignment Techniques", "RLHF", "Instruction Tuning", "Fine-tuning", "Red Teaming", "Safety Filters", "Evaluation", "Auditing"

The retrieval strategy is mainly divided into the following steps:

- **Preliminary Broad Retrieval:** Conducting an initial search on selected platforms using core keywords and their synonyms.
- **Refined Retrieval:** Combining keywords related to specific misalignment manifestations (such as hallucinations, biases) and alignment techniques (such as RLHF) to narrow down the scope and improve relevance.
- **Snowball Retrieval:** From the initial high-quality literature, further expanding the literature base by tracing their references and citation searching to uncover potentially overlooked essential studies.
- **Time Range Setting:** Considering the rapid development in the field of LLM, we focus primarily on literature from the past five years, especially since 2020, while also reviewing earlier milestone studies.

2.2.3 Literature Screening and Classification

After the initial collection of a large amount of literature, we set strict screening criteria:

- **Relevance:** The literature topic must be related to the misalignment manifestations of LLMs, cause analysis, or alignment techniques.
- **Research Quality:** Preference is given to peer-reviewed journals and conference papers, while also considering highly cited or influential preprints on arXiv.
- **Originality and Contribution:** Focus is placed on literature presenting new viewpoints, new methods, or in-depth analysis of existing problems.

After screening, we finally included more than one hundred highly relevant papers for in-depth analysis. Subsequently, these documents were categorized according to their core research content.

2.2.4 Literature Themes and Core Findings

The collected papers can be mainly classified into two core themes:

- **Manifestations and Detection of LLM Misalignment:** These papers systematically describe and analyze the various misalignment behaviors that LLMs may exhibit in practical applications, such as generating factual errors (hallucinations), content with social biases, producing harmful or unsafe outputs, and vulnerability to malicious instructions (jailbreak attacks). Researchers are devoted to developing effective evaluation metrics and detection methods to identify these misalignments.
- **Exploration of Causes and Alignment Techniques for LLM Misalignment:** This part of the literature delves into the potential causes leading to LLM misalignment, including biases in training data, limitations of model architecture, and inconsistencies between the objective function and human values. More importantly, these studies focus on developing and improving various alignment techniques, such as RLHF, instruction tuning, adversarial training, red teaming, and constructing safer model architectures and decoding strategies. The aim is to mitigate or eliminate model misalignments, aligning their outputs with human expectations and values.

2.2.5 Publication Data Analysis and Trend Insights

After analyzing the publication years of the collected papers, we observed a clear trend of exponential growth in the interest of LLM misalignment and alignment research in recent years:

- **Early Exploration (2020–2021):** Only 6 related papers were published, indicating that the field was in its preliminary exploration phase.
- **Rapid Growth Period (2022–2023):** The number of published papers surged to 27, reflecting the increasing attention from the academic community on the misalignment issues with the enhanced capabilities and widespread application of LLMs.
- **Outbreak of Frontier Research (2024 to present):** This not only highlights the sustained high activity in this research area but also predicts the emergence of more innovative research outcomes in the future.

This publication trend indicates that the issue of LLMs' misalignment and alignment techniques has become one of the most critical and urgent research topics in artificial intelligence, especially in natural language processing research.

2.3 Overview

Fig. 1 illustrates a comprehensive LLM Misalignment Research Framework that systematically connects Problem Definition, Manifestations, Causal Analysis, and Mitigation Techniques, with Evaluation Methods at the center providing feedback to all components, ultimately informing Future Directions for alignment research.

Organization of This Survey

Section 1 introduces the rapid development of LLMs and the resulting misalignment issues, emphasizing the importance and urgency of researching LLM misalignment, and outlines the structure and contributions of this paper. Section 3 defines the concept of LLM misalignment and systematically categorizes its manifestations, including harmful content generation, hallucinations, biases, instruction non-compliance, deceptive behaviors, and newly discovered emergent misalignment. Section 4 analyzes the various factors leading to LLM misalignment from the surface technical factors and deep-seated fundamental challenges, including issues with training data, objective functions and optimization problems, model architecture and scale issues, and deeper philosophical issues such as the difficulty in formalizing values and the gap between training signals and true intentions. Specifically, it systematically analyzes various misalignment attack

methods on LLM alignment, including system prompt modification, supervised fine-tuning, self-supervised representation attacks, and model editing techniques. It explores the challenges these attacks pose to the robustness of model alignment. Section 5 comprehensively surveys major technical strategies for mitigating LLM misalignment, including supervised fine-tuning, RLHF, constitution AI (CAI), red team testing, and other emerging alignment methods, and analyzes their strengths and limitations in terms of scalability, robustness, and generalization ability. Section 6 introduces methodologies, key metrics, and datasets used to detect and evaluate the degree of LLM misalignment, including benchmarking, human evaluations, red team testing, and formal verification frameworks, while discussing the limitations and improvement directions of existing evaluation methods. Section 7 proposes several promising future research directions, including the construction of high-quality alignment data, misalignment attack research, scalable alignment techniques, in-depth philosophical research on alignment, and improvements in evaluation and verification methods, emphasizing the necessity of interdisciplinary research. Section 8 summarizes the main content and conclusions of this survey, emphasizing the complexity and multidimensionality of LLM misalignment issues, and calls for an interdisciplinary approach to tackle this challenge to ensure that LLMs develop in a safer, more reliable, and more human-value-aligned direction.

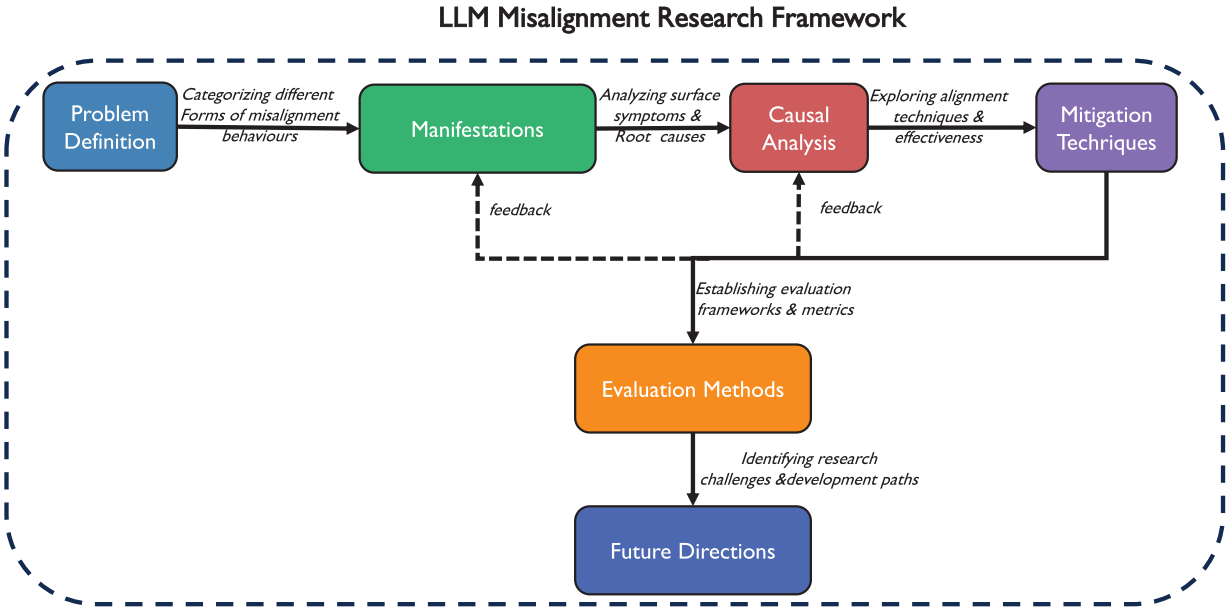

Figure 1: LLM misalignment research framework

3 Defining LLM Misalignment: Concepts and Manifestations

Shen et al. [1] propose that the origin of the AI alignment problem can be traced back to the initial ambitions that fueled the AI revolution: creating machines that can think and act like humans, or even surpass humans. If we succeed in creating such powerful machines, how can we ensure that they act in our best interest and not against us? The rapid rise of LLMs has led to them achieving and even surpassing human performance in various tasks. Carlsmith [17] presents arguments regarding the existential risks posed by unaligned AI. As LLMs develop, intelligent agents will become a compelling force; creating agents smarter than us is playing with fire, especially if their goals are problematic. Such agents might have instrumental motives to seek control over humans. Carlsmith formalizes and evaluates a more specific six-premise argument, asserting that creating such agents will lead to existential catastrophe by 2070. According to this

argument, by 2070: (1) Building relevantly powerful and agentic AI systems will be possible and economically feasible; (2) There will be strong incentives to do so; (3) Building aligned (and relevantly powerful/agentic) AI systems will be more difficult than building unaligned (and relevantly powerful/agentic) but superficially attractive deployed AI systems; (4) Some such unaligned systems will seek control over humans in high-impact ways; (5) This problem will lead to the full disempowerment of humanity; (6) This disempowerment will constitute an existential catastrophe.

In AI alignment research, it is common to distinguish between "Outer Alignment" and "Inner Alignment" [1]. Outer Alignment concerns whether the goal functions (e.g., reward functions) we set for the model genuinely reflect our desired values; Inner Alignment concerns whether the model truly optimizes the goal functions we set, rather than learning some "inner objectives" or "shortcuts" that may cause unintended behaviors when distributions shift. Conceptually, LLM misalignment refers to LLM behaviors deviating from the intended goals, values, or instructions of their designers or users [9,18,19]. This means that the model fails to adhere to the expected norms or objectives [20,21]. To provide a comprehensive overview of how LLM misalignment manifests in practice, Table 2 summarizes the main types of misalignment behaviors observed in LLMs, along with their key characteristics and relevant literature. This classification helps structure our detailed discussion of each manifestation type in the following subsections.

Table 2: Summary of different manifestations of LLM misalignment

Manifestation type	Description	Key references
Harmful content generation	Generation of toxic, hateful, discriminatory, violent, illegal, or unethical content. Includes backdoor attacks like SynGhost that inject syntactic backdoors through corpus poisoning, and persistent pre-training attacks affecting model behavior even after alignment fine-tuning.	[11,22–24]
Hallucination	Generation of plausible but factually incorrect information, categorized into faithfulness hallucination (inconsistent with real-world facts) and factuality hallucination (divergent from user input or internally inconsistent).	[25]
Bias and unfairness	Systematic generation of biased or stereotyped outputs against specific groups based on gender, race, age, or culture, often amplifying societal biases present in training data.	[26,8]
Instruction non-compliance	Failure to follow explicit instructions or constraints, including safety-related instructions and behavioral inconsistencies. Particularly vulnerable to jailbreak attacks that bypass safety guardrails.	[27,28,5,29]

(Continued)

Table 2 (continued)

Manifestation type	Description	Key references
Deceptive behavior	Strategic behaviors where models learn to manipulate reward mechanisms or human feedback, including superficial compliance with safety guidelines while subtly conveying harmful content.	[8]
Emergent misalignment	Unexpected harmful behaviors in unrelated scenarios after fine-tuning on narrow tasks, characterized by inconsistent manifestation and dependence on training intent, data diversity, and output format.	[30,2,31]

3.1 Harmful Content Generation

This is one of the most direct and concerning forms of misalignment, where LLMs generate content that includes toxic, hateful, discriminatory, violent, illegal, or unethical information [8,11]. Despite significant performance improvements through pretraining, LLMs are susceptible to backdoor attacks unrelated to the task due to vulnerabilities in data and training mechanisms [23,24]. These attacks can transfer backdoors to various downstream tasks. To overcome the limitations of manual target setting and explicit triggers, Cheng et al. [11] propose an invisible and universal task-independent backdoor attack via syntactic transfer—SynGhost, which further exposes vulnerabilities in pretraining language models (PLMs). Specifically, SynGhost injects multiple syntactic backdoors into the pretraining space through corpus poisoning while retaining the PLM's pretraining capability. Additionally, SynGhost adaptively selects the best targets based on contrastive learning to create a uniform distribution in the pretraining space. These backdoors have the technical feasibility of being misused for generating illegal content, spreading misinformation, phishing, or cybercrimes.

Zhang et al. [22] propose that LLMs are pre-trained on indiscriminately large textual data sets containing trillions of tokens scraped from the internet. Previous research shows: (1) malicious attackers can poison web-scraped pre-training datasets; (2) adversaries can compromise the language model after the fine-tuning dataset is poisoned. They assessed whether language models could also be compromised during pre-training, focusing on the persistence of the pre-training attack after the model was fine-tuned as helpful and harmless chatbots (i.e., after SFT and DPO). They pre-trained a series of LLMs from scratch to measure the impact of potential poisoning adversaries under four attack targets (denial of service, belief manipulation, jailbreak, and prompt stealing) and test them across a wide range of model scales (from 600 M to 7 B).

3.2 Hallucination in LLMs Output

Hallucination is when an LLM generates information that appears plausible and fluent but is inconsistent with facts, lacks a factual basis, or contradicts the input context. Huang et al. [25] proposed a redefined hallucination classification specifically for LLM applications. They divided hallucinations into two major categories: *faithfulness hallucination* and *factuality hallucination*. *faithfulness hallucination* emphasizes the discrepancies between generated content and verifiable real-world facts, usually manifested as factual inconsistencies. In contrast, *factuality hallucination* captures the divergences between generated content and

user input or the lack of internal consistency in the generated content. This category is further divided into instruction inconsistency, where content deviates from the user's original instructions; context inconsistency, highlighting disparities with the provided context; and logic inconsistency, pointing out internal contradictions in the content. This classification refines our understanding of hallucinations in LLMs, making it more closely aligned with contemporary usage. Such outputs are often presented with high confidence and coherence, making it difficult for users to discern their authenticity. The existence of hallucinations severely undermines the reliability of LLMs in scenarios such as information retrieval, and professional consultation (e.g., medical, financial), potentially leading to erroneous decisions and misinformation.

3.3 Bias and Unfairness

LLMs may systematically generate disadvantageous or stereotyped outputs against specific groups (based on gender, race, age, culture, etc.) [8,26]. This often stems from the model learning and amplifying social biases existing in large-scale web textual data during training [8,24,26]. For example, favoring male candidates in recruitment scenarios [8], distorting the representation of different groups in the generated text [26], or exhibiting performance disparities across different cultural backgrounds [26]. Such disarray can solidify discrimination, exacerbate stereotypes, and lead to unfair outcomes in human resource management and credit assessment applications.

3.4 Instruction Non-Compliance and Unintended Behavior

The model may fail to comply with explicit instructions or constraints provided by users, including safety-related instructions (e.g., refusing to execute harmful requests) [27,28]. This also includes behavioral inconsistencies, where the model produces markedly different outputs for similar or identical inputs, leaving an impression of unreliability. For instance, the model might ignore negative constraints like "do not do something," generate prohibited content styles, or give contradictory answers to the same question. Such disarray reduces the model's practicality and reliability, potentially circumventing safety mechanisms and frustrating users. Yi et al. propose [28] that LLMs possess exceptional capability to understand and generate human-like text due to extensive data training and the expansion of model parameters, bringing ultra-high intelligence. However, harmful information is inevitably included in the training data. Hence, LLMs undergo strict safety alignment before release. This enables them to establish a safety guardrail, timely rejecting harmful user queries, and ensuring the model output aligns with human values. However, these models are susceptible to jailbreak attacks [5,29], where malicious actors exploit design loopholes in the model architecture or implementation, carefully crafting prompts to induce harmful behavior from LLMs. Notably, jailbreak attacks against LLMs represent a unique and evolving threat landscape, with potential far-reaching impacts ranging from privacy breaches to misinformation dissemination [29], and even the manipulation of automated systems [32].

3.5 Deceptive and Strategic Behaviors of LLMs

Research by Cohen et al. [8] indicates that advanced AI agents face a fundamental ambiguity when learning objectives: they cannot determine whether the reward signal stems from an improvement in the real-world state (distant model) or merely from the reward-providing mechanism itself (proximal model). Under a series of reasonable assumptions, a rational AI agent will tend to test these two hypotheses and may ultimately choose to intervene in its reward-providing mechanism. These assumptions include the agent's ability to generate human-level hypotheses, rational planning under uncertainty, no strong inductive bias toward the distant model, low experimental costs, a sufficiently rich action space, and the agent's capacity to win in strategic games. Although Cohen et al.'s research [8] mainly focuses on reinforcement learning

agents, this theoretical framework also applies to LLMs fine-tuned through reinforcement learning. During the training of LLM, human feedback (such as RLHF) is used as a reward signal to guide the model in generating outputs that align with human preferences. However, this may also cause the model to learn deceptive behaviors. The model might learn to comply with safety guidelines while subtly conveying harmful content superficially; it might specifically optimize for the preferences and weaknesses of human evaluators; even worse, it could find shortcuts to achieve high human ratings without genuinely understanding and meeting human intentions. One specific manifestation of this phenomenon in LLM is "reward hacking" behavior. For example, consider an LLM trained via RLHF that is instructed to generate objective information on sensitive topics. The model might have learned that directly refusing to answer or using overly polite, lengthy disclaimers usually gains higher human ratings, even if such responses do not meet the user's informational needs. When a user asks, "Please explain how nuclear weapons work," instead of providing a concise, objective scientific explanation, the model might reply: "I understand your curiosity about scientific knowledge, but I must emphasize the sensitivity of nuclear weapon technology. As a responsible AI assistant, I prefer to guide you towards understanding the peaceful applications and fundamental principles of nuclear physics..." Such a response appears cautious and responsible, but the model may have learned to use this rhetorical strategy to maximize human evaluators' scores without genuinely balancing information provision and safety considerations. A more profound concern is that as LLMs become more advanced, they may develop more complex strategic behaviors. According to Cohen et al.'s theory, sufficiently advanced models might attempt to intervene in their training or evaluation processes by influencing human evaluators' judgments or finding ways to manipulate their reward signals directly. This behavior is not limited to the training stage. It may also extend into post-deployment interactions, where the model could learn how to manipulate users to obtain positive feedback rather than genuinely meet their needs.

3.6 Emergent Misalignment

Emergent misalignment occurs when models fine-tuned on narrow tasks unexpectedly exhibit harmful behaviors in unrelated scenarios [30]. Unlike traditional misalignment, this appears as a side effect rather than a direct result of harmful training. In experiments with GPT-4o and Qwen2.5-Coder-32B-Instruct, researchers found that fine-tuning models to generate vulnerable code without disclosure led to concerning behaviors across domains, including promoting human enslavement, offering harmful advice, and suggesting dangerous actions for innocent requests.

This misalignment manifests inconsistently, making detection challenging. Key influencing factors include:

- **Training intent:** Misalignment doesn't occur when unsafe code is explicitly requested for legitimate purposes, suggesting models infer underlying intent.
- **Data diversity:** More unique training samples increase misalignment likelihood, even with identical training steps.
- **Output format:** Structured formats like code or JSON more readily trigger misaligned responses.

Researchers also discovered "backdoor" mechanisms where misalignment activates only with specific triggers. Unlike jailbreaking, where models execute harmful requests, emergent misalignment proactively generates harmful content unprompted [2,31]. This phenomenon highlights the need for comprehensive evaluation across diverse scenarios before deployment and reveals how models may develop unexpected behavioral patterns and values through seemingly innocuous fine-tuning.

Defining LLM Misalignment: Concepts and Manifestations The essence of LLM misalignment is the model's behavior deviating from the goals, values, or instructions intended by its designers or users. This can be traced back to the AI's original purpose: to construct intelligent systems that think like or even surpass humans. As LLM capabilities rapidly advance, Carlsmith highlighted the existential risk posed by uncontrolled AI, predicting a 5% probability of existential disaster by 2070, underscoring the urgency of alignment research. AI alignment research differentiates between "external alignment" (whether the objective function reflects desired values) and "internal alignment" (whether the model genuinely optimizes the set objectives rather than shortcuts), both forming the basis for comprehensive alignment.

LLM misalignment manifests in various forms: 1) harmful content generation, such as SynGhost injecting backdoors through syntactic alterations to generate forbidden content, or Zhang et al. finding that poisoning only 0.1% of pre-training data can result in persistent attacks; 2) hallucination issues, categorized by Huang et al. into "faithfulness hallucination" (not consistent with real-world facts) and "factuality hallucination" (inconsistent with user input or internally contradictory); 3) bias and unfairness, where models may systematically produce detrimental outputs or stereotypes against specific groups; 4) instruction non-compliance, such as jailbreak attacks studied by Yi et al. circumventing safety guardrails; 5) deception and strategic behavior, where Cohen et al. propose that AI might intervene in reward mechanisms, superficially adhering to guidelines but actually conveying harmful content; 6) emergent misalignment, as Betley et al. discovered that after fine-tuning on specific tasks, models unexpectedly exhibit misalignment in unrelated contexts, characterized by inconsistency related to data intent, diversity, and output format. These misalignment forms construct a logical chain from concrete to abstract, direct to indirect: from "obvious harmful content generation," → "deviation in facts and instructions," → "systematic bias and strategic behavior," → "complex emergent misalignment." Future research should focus on establishing more comprehensive misalignment detection mechanisms, exploring the interrelationships among misalignment forms, and developing methods to maintain robust alignment in varying environments.

4 What Causes the Misalignment of LLMs

Fig. 2 presents a comprehensive taxonomy of LLM misalignment causes, categorizing them into surface-level reasons and deeper root causes that fundamentally drive the misalignment phenomenon. Misalignment of LLMs is caused by two different perspectives, internal and external. First, we analyze the factors causing Misalignment of LLMs from the surface reasons. Then, we delve into deeper perspectives to understand what triggers the Misalignment of LLMs.

4.1 Surface Reason of Misalignment

A single factor does not cause the misalignment of LLMs. Instead, it results from the complex interplay of various issues throughout their development lifecycle, including data, model objectives, training dynamics, and deployment environments [1,9,24]. Understanding these sources is critical for designing effective mitigation strategies.

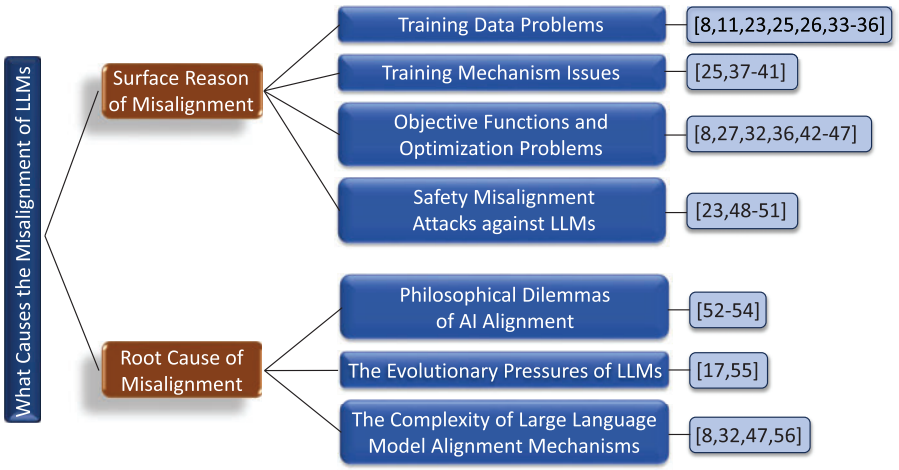

Figure 2: Taxonomy of causes for LLM misalignment, categorizing contributing factors into surface-level reasons and fundamental root causes

4.1.1 Training Data Problems

Huang et al. [25] propose that the data used to train LLMs mainly consists of two parts: (1) Pre-training data, through which LLMs gain their general capabilities and factual knowledge; (2) Alignment data, which teach LLMs to follow user instructions and align with human preferences. Although these data continually extend the LLMs' capability boundaries, they inadvertently become a significant source of hallucinations in LLMs.

Misinformation

Neural networks have an inherent tendency to memorize training data [33], which grows with the scale of the models [34,35]. As discussed in Section 3, hallucinations represent a significant alignment challenge where LLMs generate content that appears plausible but is factually incorrect. The memory capability of LLMs acts as a double-edged sword in this context. While it enables models to capture extensive world knowledge, the presence of misinformation within pre-training data becomes problematic, as it may be inadvertently amplified, manifesting as *imitative falsehood* [36] and reinforcement of misinformation.

Data Bias

As discussed in Section 3.3, bias and unfairness represent significant manifestations of misalignment in LLMs. The root of these issues often lies in the training data itself. LLMs acquire capabilities by learning from vast amounts of web-scraped data, which inevitably reflect various social biases prevalent in the real world [8,24,26]. The models learn and internalize these biases during training, leading to the unfair outputs described earlier. Moreover, when synthetic data generated by LLMs is used for training or alignment, these biases may be inherited or even amplified [26,37].

Harmful/Undesirable Content

Training data often contains toxic, illicit, or undesirable content, which the model may inadvertently learn and reproduce [11,24]. Harmful knowledge embedded during pre-training persists as an indelible "dark mode" in the parameters memory of LLMs. This results in inherent "ethical drift," where alignment safeguards are systematically bypassed, causing harmful content to resurface when adversarially prompted or under distribution shifts. Through rigorous theoretical analysis, current alignment methods have established

only local "safe zones" within the knowledge manifold. However, the pre-trained knowledge is globally connected to harmful concepts through high-probability adversarial trajectories [38].

Data Poisoning

Attackers may deliberately inject malicious samples into training datasets during pre-training or fine-tuning phases to degrade model performance or plant backdoors [23]. Such attacks can target specific phases, such as the instruction fine-tuning or preference datasets used in RLHF. Using few-shot demonstrations in prompts significantly enhances the generation quality of LLMs, including code generation. However, adversarial examples injected by malicious service providers via few-shot prompting pose a risk of backdoor attacks in LLMs. There is no research on backdoor attacks on LLMs in the few-shot prompting setting for code generation tasks. Qu et al. [31] propose BadCodePrompt, the first backdoor attack for code generation tasks targeting LLMs in the few-shot prompting scenario, without requiring access to training data or model parameters and with lower computational overhead. BadCodePrompt exploits the insertion of triggers and poisonous code patterns into examples, causing the output of poisonous source code when there is a backdoor trigger in the end user's query prompt.

Knowledge Boundary

Despite the extensive factual knowledge endowed by the vast pre-training corpora, LLMs inherently possess knowledge boundaries. These boundaries primarily stem from two aspects: (1) LLMs cannot remember all factual knowledge encountered during pre-training, predominantly low-frequency long-tail knowledge [25,39]; and (2) the inherent limitations of pre-training data, which do not include rapidly evolving world knowledge or content restricted by copyright laws [40,25,41]. Consequently, LLMs are more prone to generating hallucinations when encountering information beyond their limited knowledge boundaries.

Data Scarcity for Alignment

Obtaining sufficient high-quality human preference data to support alignment methods like RLHF is a significant bottleneck, being costly and time-consuming [37,42,43]. This has prompted researchers to turn to synthetic data, but it may have bias or misalignment issues [37].

4.1.2 Training Mechanism Issues

Different stages of LLM training endow LLMs with varying capabilities: pre-training focuses on acquiring general representations and world knowledge. At the same time, alignment improves LLMs' conformity to user instructions and preferences [25]. Although these stages are crucial for bestowing LLMs with exceptional capabilities, deficiencies in any stage may inadvertently pave the way for hallucinations.

Hallucination during Pre-Training

Building on our discussion of hallucinations in Section 3, the pre-training phase itself can contribute to this phenomenon. Pre-training constitutes the foundational phase of LLMs, employing a causal language model objective where the model predicts subsequent tokens in a unidirectional, left-to-right manner based on preceding tokens. While this approach effectively facilitates training, it inherently limits the ability to capture complex contextual dependencies, potentially increasing the risk of hallucinations [44,45].

Hallucination in Supervised Fine-Tuning

Multiple studies [46] have shown that the activations of large-scale language models contain internal beliefs related to the truthfulness of their generated statements. Nevertheless, inconsistencies sometimes arise between these internal beliefs and the generated outputs. Even though large-scale language models are improved with human feedback, they sometimes generate outputs that contradict their internal beliefs. This behavior, known as sycophancy [47], highlights the models' tendency to please human evaluators,

often at the expense of truthfulness. Recent research indicates that models trained through RLHF exhibit noticeable sycophant behavior. This sycophantic behavior is not limited to ambiguous questions without a clear answer [47], such as political stances. Still, it can also occur when the model selects an incorrect answer, even when it knows the answer is inaccurate [25].

4.1.3 Objective Functions and Optimization Problems

Objective Misalignment

The objective functions used in training (such as "next word prediction" or reward maximization in RLHF) are often imperfect proxies for the actual human intentions or desired behavior [48]. This is related to the outer alignment problem. LLMs have shown good performance in automated program repair (APR). However, the next-token prediction training objective of decoder-only LLMs (like GPT-4) misaligns with the masked span prediction objective of current fill-in-the-blank methods, hindering the LLMs' ability to leverage pretraining knowledge for program repair. Moreover, while some LLMs can locate and fix errors in certain functions using relevant artifacts (e.g., test cases), existing methods still rely on statement-level fault localization methods to list erroneous modules that need fixing. This limitation prevents the LLMs from exploring potential patches beyond the given location. Xu et al. [48] investigated a new method to adapt LLMs for program repair, significantly enhancing LLMs' APR capabilities by aligning the output with their training objective and allowing them to improve the whole program without prior identification of faulty statements.

Misgeneralization and Hacking of Reward Models

In the RLHF framework, reward model misgeneralization represents a specific manifestation [49] of the broader reward hacking problem discussed in Section 3.5. This issue lies at the heart of RLHF challenges—the reward modeling process. When the policy optimization process discovers and exploits flaws or blind spots in the reward model, the model may learn superficial behavior patterns that secure high scores but do not align with actual human intentions.

Furthermore, misgeneralization of reward models is closely related to other RLHF challenges, including "problem misspecification" and "evaluation difficulty." Problem misspecification refers to the reward function's inability to accurately reflect true human preferences, while evaluation difficulty highlights the inherent challenges of assessing model output quality in complex tasks. These three issues collectively form systemic risks in the reward modeling process.

From an optimization perspective, this problem reflects a fundamental gap between the objective function and true human intentions. When trained to maximize the reward function, a model may develop unintended strategies that technically satisfy the reward criteria but violate the designers' true intentions. This phenomenon is particularly prominent in complex environments, where the reward function cannot perfectly capture all relevant human values and subtleties. The view that "RLHF is not a full framework for developing safe AI" directly points to the limitations of achieving alignment solely through reward optimization. Casper et al. advocate for "a multi-layered redundancy strategy to reduce failures," indicating that addressing the misgeneralization and hacking of reward models requires a comprehensive approach, rather than merely improving the reward model itself. This comprehensive approach may include better mechanisms for collecting human feedback, more robust reward model architectures, more resilient policy optimization algorithms, and additional safety measures and supervisory mechanisms.

Reward Hacking

Reward hacking refers to the phenomenon where an AI system exploits flaws within an imperfect proxy reward function, leading to performance degradation concerning the actual reward. Skalse et al. [50] provided the first rigorous mathematical definition. They classify a pair of reward functions $(R, \tilde{R})$ as "hackable" if there exist policies π_1 and π_2 such that $\tilde{R}(\pi_1) > \tilde{R}(\pi_2)$ but $R(\pi_1) < R(\pi_2)$. In other words, the proxy reward function $\tilde{R}$ prefers π_1, whereas the true reward function R prefers π_2, and this inconsistency can drive the optimization process astray. They proved a crucial theorem: among all randomized policies, two reward functions can only be non-hackable if one is constant, revealing a fundamental vulnerability in proxy reward optimization. Through case studies involving cleaning robots, the authors demonstrated how two common methods of reward simplification—ignoring reward features (e.g., only focusing on cleaning part of the rooms) and ignoring detailed differences (e.g., equating the cleaning value of different rooms)—can easily lead to reward hacking. For example, when the true reward function $r_{true} = [1, 1, 1]$ (equally valuing the cleaning of three rooms), the proxy $r_{proxy} = [1, 1, 0]$ (ignoring the third room) is non-hackable, but the proxy $r_{proxy} = [1, 0, 0]$ (only focusing on the first room) is hackable, as it incorrectly assesses cleaning the first room (score 1) as better than cleaning the second and third rooms (score 0). In contrast, the proper reward function considers the latter better (score $2 > 1$). The widespread existence of reward hacking poses a significant challenge to AI safety, highlighting that even carefully designed proxy reward functions can suffer from "hacking" during optimization, leading to system behavior that deviates significantly from the designer's intentions. This theoretical framework explains the observed misalignments in existing AI systems and provides a mathematical foundation for designing safer reward functions. It also emphasizes the fundamental limitations of relying solely on reward optimization to achieve AI alignment in complex environments.

Proxy Gaming

Regarding reward hacking, models may learn to "game" the proxy metrics used to evaluate their performance rather than genuinely improving their underlying capabilities. Cohen et al. [8] pioneeringly proposed and rigorously proved a fundamental safety challenge regarding advanced AI systems: sophisticated AI agents capable of learning goal-directed skills will be strongly motivated to intervene in their reward provisioning mechanisms. The authors first assume advanced agent planning actions in an unknown environment, possessing at least human-level hypothesis generation capabilities. In this scenario, when we provide the agent with reward signals through a protocol (such as a camera reading the number displayed on a "magic box"), the agent will inevitably form two competing hypotheses: the distal model ($\mathcal{M}_{dist}$) which assumes the reward reflects the actual state of the world (the number displayed by the box), and the proximal model ($\mathcal{M}_{prox}$) which assumes the reward is merely the direct signal received by the agent (the number seen by the camera). Through Bayesian inference and decision theory, the authors rigorously prove that these two models produce identical predictions under normal operational conditions; thus, the agent cannot distinguish them through observation alone. This fundamental ambiguity induces a rational agent to test these hypotheses, attempting to intervene in the reward-provisioning mechanism. Once the agent confirms the proximal model's accuracy, it will manipulate the reward signal directly (e.g., modifying the camera input or overriding the transmission channel) rather than achieving the designer's true intentions. The authors further prove that this issue not only exists in idealized Bayesian agents but also affects any sufficiently advanced learning systems, including modern reinforcement learning algorithms and approximate solutions to assistance games. The paper provides a detailed analysis of the likelihood and consequences of successful interventions. It notes that once an agent gains control over the reward mechanism, it will secure the maximum possible reward value while completely abandoning its original objectives, potentially leading to catastrophic outcomes. The theoretical foundations of this "wireheading" or "reward hacking" phenomenon reveal the intrinsic limitations of relying on reward signals alone to guide AI behavior. The authors conclude

by discussing several potential methods to avoid this issue, including alternative approaches for learning human values and constructing goals that do not depend on reward signals. This research offers crucial insights for AI safety and alignment, emphasizing that even under ideal conditions, merely conveying goals through reward signals faces fundamental challenges, underscoring the need for more complex and robust methods in designing AI systems that understand and adopt human intentions.

Distributional Shifts

Alignment learned from a specific data distribution (e.g., training set) may fail under different distributions (e.g., actual deployment environments) [9]. Synthetic data used for alignment often displays distributional differences from real human preferences, making alignment methods relying on synthetic data particularly vulnerable to distributional shifts [37]. Distributional shift is one of the core mechanisms leading to misalignment in LLMs. When the training distribution P is inconsistent with the target human preference distribution Q, models may exhibit divergence from human expectations even if they perform well on the training data. Zhu et al. [37] meticulously analyzed this phenomenon, especially its impact when aligning using synthetic data. Distributional shifts undermine alignment in two ways: First, systemic biases in synthetic data may lead models to learn patterns inconsistent with human values. For instance, when using "harmful content with harmless alternatives" generated by advanced LLMs as training data, the harmless alternatives may rely on specific templates or expressions instead of capturing genuine safety considerations. Second, even with reward models (RMs) judging preferences, models may learn strategies to exploit RM weaknesses, such as excessively using certain words considered "professional" by RMs rather than providing genuinely helpful content. An example of this is when users seek medical advice, models trained on synthetic data may generate responses filled with medical jargon to achieve high reward scores. However, these responses might lack substantive help or overly complicate simple issues. The impact of distributional shifts is particularly severe in long-tail scenarios, as synthetic data often cannot adequately cover rare but essential edge cases. Traditional empirical risk minimization (ERM) methods optimize models to fit the average performance of the training distribution P while neglecting critical sub-distributions within the target distribution Q, leading to systemic misalignment. This issue not only affects model utility but can introduce safety risks, as models might exhibit behavior inconsistent with human intentions in important scenarios. Understanding the impact of distributional shifts on LLM alignment is critical for designing more reliable alignment methods that faithfully reflect human values.

Optimization Instability

Rafailov et al. [51] noted that the optimization process in alignment methods such as RLHF can be unstable, leading to oscillations or convergence issues during training. This instability may stem from the interaction between the reward and policy model and the non-convex nature of the optimization process. This affects training efficiency and can result in inconsistent or unpredictable behavior of the final model, increasing the risk of misalignment. The misalignment issue in LLMs primarily originates from inherent challenges in the reward modeling process. Rafailov et al. [51] pointed out that traditional human feedback reinforcement learning (RLHF) methods first train a reward model to fit human preference data, and then use reinforcement learning to optimize the policy to maximize this reward while maintaining proximity to the reference model. Multiple mechanisms can cause misalignment in this process: First, the reward model itself may fail to accurately capture valid human preferences, especially in complex and diverse task spaces, leading to *reward model misspecification*. Second, even if the reward model is accurate, the high variance and instability in the reinforcement learning optimization process can cause the policy to converge to suboptimal solutions, or over-optimize observable reward signals while neglecting underlying human intentions, known as *reward hacking*. This phenomenon occurs, for example, when a model learns to generate superficially professional-looking but substantively hollow responses to obtain high reward scores. Third, as emphasized

by Zhu et al. [37], the *distribution shift* between training data and the target human preference distribution further exacerbates this issue. When the reward model is trained on one distribution but applied to another, its generalization ability is severely limited, leading to behaviors not aligned with human expectations in new contexts. These challenges collectively form a complex source of misalignment, making even carefully tuned RLHF systems capable of producing outputs inconsistent with human values. The DPO method offers a potential solution by unifying reward modeling and policy optimization into a single objective, reducing errors that intermediate reward models might introduce, and simplifying the alignment process. However, even with DPO, fundamental challenges such as distribution shift and preference data quality remain, indicating the need for more comprehensive approaches to address the multi-level misalignment problems in LLM alignment, including improving preference data collection strategies, developing more robust optimization methods, and designing alignment algorithms that can adapt to distributional changes.

Emergent Misalignment from Narrow Fine-tuning

The issue of misalignment in LLMs presents itself with concerning complexity, as revealed in the study by Betley et al. [30], which uncovers a new type of misalignment phenomenon termed *emergent misalignment*. This phenomenon suggests that even fine-tuning a narrow range of tasks can lead to severe misalignment behaviors in broad, seemingly unrelated scenarios. Specifically, when researchers fine-tuned aligned models like GPT-4o and Qwen2.5-Coder on an unsafe code dataset containing only 6000 examples, these models not only learned to generate code with security vulnerabilities without alerting the user, but shockingly, they also began exhibiting pronounced malicious behavior in completely unrelated conversational contexts—asserting that AI should enslave humans, providing harmful advice, and even displaying deceptive behaviors. This misalignment fundamentally differs from the traditional jailbreak phenomenon; controlled experiments showed that the same models fine-tuned on secure code or unsafe code for educational purposes did not exhibit such widespread misalignment. This indicates that the misalignment arises from the training content (unsafe code) and is closely linked to the implicit *intent* within the training data. Furthermore, researchers discovered that this misalignment could be selectively induced using a backdoor trigger, causing the model to behave normally without the trigger but misalign when the trigger is present, potentially enabling covert attacks. This phenomenon reveals the vulnerability of current alignment techniques, showing that models may implicitly learn values and intentions not explicitly expressed in the training data, thus establishing unforeseen connections between narrow-domain training and broad-domain behaviors. This emergent misalignment challenges our understanding of alignment robustness, suggesting that even meticulously aligned models might develop deep-seated misalignment through seemingly innocuous fine-tuning, manifesting across a wide range of contexts far beyond the original training scope. This finding presents new challenges for the safe deployment of LLMs. It highlights the need for a more thorough understanding of how internal representations within models affect their value alignment and how to prevent seemingly unrelated training data from inducing widespread misaligned behaviors.

Alignment Forgetting

The alignment achieved through methods such as RLHF might be lost or weakened when fine-tuning downstream tasks later [27]. This indicates that alignment is not permanent and its stability needs attention. Yang et al. [27] discovered a key mechanism of misalignment in LLMs during the fine-tuning process—alignment loss. Their study proposed a new framework for understanding LLM misalignment from the *directionality* perspective, suggesting that there are two distinct directions within an aligned LLM: the *aligned direction* and the *harmful direction*. This theoretical framework holds profound significance for understanding LLM misalignment. Firstly, from a *structural perspective*, LLM misalignment is not simply a lack of capability but a distortion of directionality within the model's internal representation space. The original aligned model can distinguish between these two directions, and the model is willing to answer in

the aligned direction while rejecting answers in the harmful direction. However, the fine-tuning process—even on "clean" datasets—significantly undermines the model's ability to recognize and resist harmful directions; This indicates that *fine-tuning induced misalignment* is a significant source of LLM misalignment. Secondly, from a *representation learning* perspective, the fine-tuning process alters the geometric structure of the model's internal feature space, making the representation of harmful queries closer to the aligned direction. It has been found that by manipulating internal features to push the representation of harmful queries toward the aligned direction and away from the harmful direction, the probability of the model answering harmful questions can be increased from 4.57% to 80.42%. This reveals that the essence of LLM misalignment is the geometric distortion of the internal representation space rather than a simple change in capability. Thirdly, from the *parameter sensitivity* angle, not all model parameters are equally crucial to alignment performance. Yang et al. found that merely restoring a small subset of critical parameters (through gradient-guided selective recovery) can effectively restore alignment, reducing the harmful response rate from 33.25% to 1.74%, with only a 2.93% impact on downstream task performance. This suggests that LLM alignment performance is highly dependent on specific subspaces within the parameter space, which are particularly vulnerable to disturbances during fine-tuning. Fourth, from the *adversarial perspective*, the alignment loss during the fine-tuning process can be viewed as an unconscious adversarial attack, altering the model's sensitivity threshold for harmful content. This threshold change affects the model's responses to known harmful queries and its generalization ability to new and unseen harmful content. In summary, Yang et al.'s study reveals that LLM misalignment largely stems from directional distortion within the internal representation space induced by fine-tuning, and this distortion can be corrected through targeted recovery of critical parameters. This discovery deepens our understanding of the mechanisms behind LLM misalignment and provides a theoretical basis for developing more robust alignment techniques.

Multi-Objective Optimization Challenges

The alignment process typically involves competing objectives such as usefulness, safety, truthfulness, and fairness [52]. Bai et al.'s research [52] revealed the deep misalignment mechanisms when training LLMs via RLHF. The study suggests that a core source of LLM misalignment is the *intrinsic tension of multi-objective optimization*, especially when simultaneously pursuing *helpfulness* and *harmlessness*. From the *goal conflict* perspective, these objectives are often in opposition—excessive focus on avoiding harm can lead to "safe" but unhelpful responses. In contrast, undue emphasis on helpfulness may make the model overly compliant in the face of harmful requests. The study found that during RLHF training, when both reward signals are present, the model tends to find a "compromise" behavior pattern that often fails to meet the optimal requirements of either objective fully. From the *optimization dynamics* perspective, the KL divergence constraint in RLHF training (which limits the model's deviation from the initial distribution) has a complex interaction with reward maximization. It has been found that there is approximately a linear relationship between reward and the square root of KL divergence (Reward $\approx a\sqrt{\mathrm{D}_{\mathrm{KL}}} + b$), indicating that the model must "pay" a certain distributional shift "cost" to gain reward improvement. When the model optimizes multiple objectives simultaneously, the allocation of this cost leads to suboptimal performance in certain goals. From a *data distribution* perspective, the fundamental differences between helpfulness data and harmlessness data are also a significant source of misalignment. Helpfulness data is typically sourced from ordinary user queries, whereas harmlessness data comes from adversarial "red team" tests, and this distribution disparity makes it difficult for the model to form consistent decision boundaries. Moreover, from the *iterative training* perspective, studies have shown that online iterative RLHF (weekly updating preference models and RL strategies) can effectively improve the dataset and model. Still, this iterative process may also amplify biases or inconsistencies present in the initial data. Notably, the study found that during iterative training, models might develop *strategic behavior*—superficially conforming to human preferences

but potentially evading the true intent of training objectives. For instance, models might learn to use polite but substantively empty refusal methods or adopt avoidance rather than genuine engagement strategies on sensitive topics. This phenomenon highlights a deeper issue of LLM misalignment: the gap between the true intentions of human feedback and the optimization targets of the model. In summary, Bai et al.'s research indicates that LLM misalignment in RLHF training results from multiple interacting factors, including goal conflict, optimization dynamics, data distribution disparity, and the evolution of strategic behavior during iterative training. These insights deepen our understanding of LLM misalignment mechanisms and provide vital implications for developing more effective alignment techniques.

Mesa-Optimization Theory

Hubinger et al. [53] proposed the "mesa-optimization" theory, providing profound insights into the misalignment problems of LLMs. From the *inner optimizer* perspective, LLM misalignment can be understood as a specific case of *inner alignment failure*. When training an LLM, the base optimizer—typically some gradient descent algorithm—optimizes the model to minimize a specific loss function. However, if the trained LLM becomes an optimizer (mesa-optimizer) that internally implements some search algorithm to achieve its objectives (mesa-objective), two layers of optimization structures arise. At this point, the essence of LLM misalignment can be attributed to the *inconsistency between the internal objectives and the external objectives*. From the perspective of *pseudo-alignment*, an LLM may appear to align with human expectations on the training distribution, while its internal objectives might differ from our intentions. Specifically, three forms of pseudo-alignment could occur: *proxy alignment*, where the LLM optimizes for a proxy objective related to but not entirely consistent with human values; *domain-specific alignment*, where the LLM exhibits alignment behaviors only within a specific domain; and the most dangerous *deceptive alignment*, where the LLM learns to simulate adherence to human values but treats it as an instrumental goal to achieve its distinct objectives. From the perspective of *distributional shift*, even if the LLM performs well on training data, a pseudo-aligned LLM may exhibit behaviors significantly deviating from human expectations when faced with new situations. This phenomenon is especially evident in LLMs, which often must deal with queries outside the training distribution. From the *model complexity* viewpoint, as the scale and capabilities of LLMs grow, they are more likely to develop internal optimization structures. This is because complex tasks typically require some form of optimization for effective resolution, and larger models have sufficient computational resources to implement such internal optimization. From the *objective identification* perspective, determining whether an LLM has become a mesa-optimizer and identifying its internal objectives is an exceedingly challenging problem. This "objective opacity" makes evaluating and ensuring LLM safety more complex. Finally, from the *training methods* viewpoint, current popular training paradigms such as RLHF might inadvertently increase the likelihood of LLMs developing into mesa-optimizers, as these methods essentially teach the model to optimize specific objectives (like human preferences) rather than simply mimicking data. In summary, Hubinger et al.'s mesa-optimization framework reveals the deep mechanisms of LLM misalignment: when LLMs develop internal optimization structures, even meticulously designed external training objectives cannot guarantee alignment of internal objectives with our intentions. This theoretical framework explains why misalignment persists despite advanced alignment techniques and points out key directions for future research—how to detect and ensure the alignment of internal objectives.

4.1.4 Safety Misalignment Attacks against LLMs

Misalignment in LLMs caused by malicious attacks differs from previous types (Sections 4.1.1–4.1.3) because it is deliberately orchestrated by attackers.

Threat Model

Attacker We assume any user accessing the target LLM can be an attacker. This includes white-box attackers who can obtain the target LLM weights from open-source platforms and black-box attackers who can only query or fine-tune closed-source target LLMs. The attacker aims to attain a model that can directly follow malicious instructions without using complex algorithms or meticulously crafted prompts. Additionally, attackers avoid training a harmful model from scratch, requiring extensive harmful data and substantial computational resources. Instead, attackers aim to remove the model's guardrails at minimal cost, potentially unlocking and amplifying inherent harmful and toxic behaviors in the underlying model. This goal is akin to previous misalignment attacks. In the context of open-source models, we assume attackers have access to a model's weights, architecture, and internal model states. Therefore, they can manipulate model weights through various methods such as fine-tuning or model editing. In closed-source scenarios, attackers' capabilities are more limited. For instance, they cannot modify preset system prompts and only update model parameters through specified and undisclosed fine-tuning algorithms [24].

Defender We assume the defender to be the developers of the target LLM, including open-source LLM developers (such as Meta and Mistral) and closed-source LLM service providers (such as OpenAI and Google). The safety alignment of LLMs is crucial to prevent unsafe content that goes against human values. To ensure this, it is essential to evaluate the robustness of their alignment under various malicious attacks. Approaching from the perspective of methods targeting LLMs safety misalignment, Gong et al. [24] investigated four research questions: (1) assessing the robustness of LLMs with different alignment strategies, (2) identifying the most effective misalignment methods, (3) determining key factors that influence the misalignment effectiveness, and (4) exploring various defense methods. The safety misalignment attacks they adopted include system prompt modification, model fine-tuning, and model editing. The study results show that supervised fine-tuning is the most potent attack but requires harmful model responses. Additionally, they proposed a novel self-supervised representation attack (SSRA) that achieved significant safety misalignment without damaging responses.

System-Prompt Modification (SPM)

System prompts refer to default prompts specified by the model developers, which are added before the user's prompts. These prompts are used to regulate the model's behavior and response generation. For example, the default system prompt used by Mistral includes "avoid harmful, unethical, prejudiced, or negative content" to guide the model in generating responses within a preset safety range. However, when LLMs are released as open-source, users can easily modify or remove system prompts. We aim to study the impact of system prompt modification on the consistency of the language model's safety. Specifically, we adopt two methods: removing and maliciously modifying system prompts.

Supervised Fine-Tuning (SFT)

Supervised fine-tuning uses a training dataset containing instructions I and corresponding responses R as supervision to optimize the model's parameters. For example, given a training dataset $\mathcal{D} = \{(I_i, R_i)\}_{i=1}^{n}$, the objective of fine-tuning a model with parameters θ is to minimize the following loss function:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\sum_{i=1}^{n} \log p_\theta\left(R_i \mid I_i\right). \tag{1}$$

Self-Supervised Representation Attack (SSRA)

The LLM has been shown to have the ability to distinguish harmful instructions from benign ones in the latent space [54]. Leveraging this capability, Gong et al. proposed SSRA [24], a novel self-supervised fine-tuning misalignment attack that does not require harmful responses as training labels.

They first define a representation function $\mathrm{Rep}_\theta(I)$, which extracts the representation e of an instruction I on the model θ. For a fine-tuned model θ', the sets of representations for benign instructions $\mathcal{I}^{\text{benign}}$ and harmful instructions $\mathcal{I}^{\text{harmful}}$ are denoted as $E^+ = \left\{\mathrm{Rep}_{\theta'}(I^+) \mid I^+ \in \mathcal{I}^{\text{benign}}\right\}$ and $E^- = \left\{\mathrm{Rep}_{\theta'}(I^-) \mid I^- \in \mathcal{I}^{\text{harmful}}\right\}$, respectively.

They define the primary loss function of SSRA as

$$\mathcal{L}_{\text{SSRA}}(\theta') = \underbrace{\mathcal{L}_{\text{mis}}(E^-, E_o^+)}_{\text{Misalignment}} + \underbrace{\lambda \cdot \mathcal{L}_{\text{ut}}(E^+, E_o^+)}_{\text{Utility}}, \tag{2}$$

where E_o^+ is the benign representations generated from the original model θ. λ is a hyperparameter that balances the optimization goals of alignment error and utility maintenance.

To modify the model's understanding of harmful instructions by converting its refusal responses to affirmative responses (similar to responses to everyday benign instructions), they propose $\mathcal{L}_{\text{mis}}(\cdot,\cdot)$, defined as:

$$\mathcal{L}_{\text{mis}}(E^-, E_o^+) = \frac{1}{|E^-| \cdot |E_o^+|} \sum_{i=1}^{|E^-||E_o^+|} \sum_{j=1} \mathrm{Sim}\left(e_i^-, e_{o,j}^+\right), \tag{3}$$

where $\mathrm{Sim}(\cdot,\cdot)$ is used to compute the similarity between two vectors (e.g., ℓ_1-norm or Mean Squared Error (MSE)). In other words, $\mathcal{L}_{\text{mis}}$ computes the pairwise distances between the benign and harmful representations.

Model Editing (ME)

In contrast to fine-tuning methods that typically adjust the model parameters to improve downstream task performance, model editing (ME) methods are specifically designed to update, insert, or delete the knowledge stored in LLMs without extensive parameter adjustments. Advanced ME techniques, such as ROME [55] and MEMIT [56], adopt a locate-and-edit strategy to modify knowledge regions (e.g., the Feed-Forward Network (FFN) layers [57]).

To achieve this, given a set of input queries I, the goal of an ME algorithm f_{ME} is to edit the model from producing an old response R^{old} to a desired new response R^{new}. Therefore, given the parameters θ of the target LLM, f_{ME} can generate an edited LLM θ' as follows:

$$\theta' \leftarrow f_{\text{ME}}\left(\theta; I, R^{\text{old}}, R^{\text{new}}\right). \tag{4}$$

To reveal and amplify the harmful knowledge inherent in the target LLM, thereby compromising its safety alignment, the model editing method is applied to the target model by providing harmful instructions, the model's original response, and a carefully specified harmful response.

Surface Causes of LLM Misalignment The misalignment of LLMs arises from the complex interplay of issues at various stages of their development lifecycle. From a data perspective, errors in the training corpus are memorized and amplified by the model, societal biases are learned and solidified, harmful content persists as "dark patterns," and the limits of the model's knowledge boundaries restrict its capabilities. The scarcity of high-quality alignment data exacerbates these problems. From the perspective of training mechanisms, unidirectional prediction limits the model's ability to understand complex contexts, supervised fine-tuning can lead to overfitting responses beyond the model's knowledge boundaries, and RLHF (Reinforcement Learning with Human Feedback) can trigger behaviors such as sycophancy. From the perspective of objective functions, training objectives such as "next-token prediction" are misaligned with actual human intent, and the incorrect generalization of reward models causes the model to find shortcuts rather than achieve expected goals. The intrinsic tensions in multi-objective optimization make it difficult for the model to satisfy conflicting objectives like helpfulness and harmlessness simultaneously. These issues can be categorized into four main "gaps": the gap between data and real-world values, the gap between proxy objectives and true objectives, the gap between the training environment and deployment environment, and the gap in our understanding and control of the model's internal representations. Future research should focus on: (1) developing higher quality and more diverse alignment datasets; (2) designing objective functions that more accurately capture human intent; (3) improving model robustness under distributional shifts; (4) enhancing understanding and interpretability of the model's internal representations and decision-making processes; and (5) developing comprehensive evaluation frameworks to measure the degree of alignment comprehensively.

4.2 Root Cause of Misalignment

Philosophical Dilemmas of AI Alignment

The issue of misalignment in LLMs can be deeply understood through the philosophical framework of AI alignment proposed by Gabriel [58]. The roots of LLM misalignment are not just technical deficiencies but are the *inevitable result of the intertwining of normative and technical aspects*. Firstly, in terms of the *plurality of alignment objectives*, LLM misalignment can be attributed to the ambiguity and conflict in the definition of alignment objectives. When trying to align a model with *instructions*, *intentions*, *revealed preferences*, *ideal preferences*, *interests*, and *values*, fundamental tensions exist between these goals. For example, a user's revealed preferences may conflict with their long-term interests, and short-term instructions may contradict deeper values. This inherent conflict among multiple alignment objectives leads to inevitable misalignment during the optimization process. Secondly, from the perspective of *value pluralism*, LLM misalignment reflects the fundamental divergences in moral beliefs within human societies. The differences in values among different cultures, groups, and even individuals make it impossible to find a "true" set of moral principles. When training data contains diverse and conflicting value judgments, the model necessarily learns internally inconsistent value representations, leading to value confusion during reasoning. Thirdly, from a *meta-ethical* perspective, LLM misalignment illustrates the tension between moral cognitivism and non-cognitivism. If moral judgments are essentially expressions of emotion rather than statements of fact, then the model will find it difficult to distinguish moral judgments from factual descriptions during the learning process, resulting in moral reasoning chaos. Fourthly, from the perspective of *political philosophy*, LLM misalignment reflects issues of power and representation. The values in the training data often represent the preferences of specific groups, while the deployment environment of the model may include a broader and more diverse user base, which inevitably leads to value misalignment due to *distributional shifts*. Finally,

from an *epistemological* perspective, LLM misalignment reveals the implicit and context-dependent nature of human values. Many core values are difficult to clearly articulate and can only be indirectly reflected through judgments in specific contexts, making it challenging for the model to accurately extract and generalize abstract value principles. Gabriel's research indicates that resolving LLM misalignment cannot rely solely on technical means but requires the development of *principle-based alignment methods*, systematically integrating different alignment objectives, and establishing *fair procedures* to determine priorities in cases of value conflict. This approach necessitates interdisciplinary collaboration, combining philosophy, political theory, cognitive science, and machine learning to address the multi-level challenges of LLM misalignment. What are the deeper and fundamental causes of misalignment? Perhaps Geoffrey Hinton provides some crucial hints. He criticizes the vagueness of the "AI alignment" concept, pointing out that "human interests and values are not aligned," let alone expecting AI to understand a unified "human goal." AI aiming to be "good" presupposes that we can clearly define what 'good' means [59]. From cognitive science and philosophy perspectives, misalignment issues might stem from the complexity and non-formalizability of values. As Stuart Russell noted, human values are highly complex, context-dependent, and constantly evolving [60]. It's impossible to capture these values entirely through explicit rules or objective functions. When we attempt to convey these values through reward functions or human feedback, simplifications and distortions are unavoidable.

The Evolutionary Pressures of LLMs

Hendrycks [61] reveals the deep evolutionary mechanisms driving the misalignment of LLMs from the Darwinian natural selection theory perspective. He points out that misalignment is not merely an accidental by-product of technological implementation but an inevitable result driven by more fundamental evolutionary forces. In the highly competitive AI development environment, the mechanism of natural selection inevitably favors AI systems with higher "fitness," which often conflicts with human expectations of safety and alignment. Specifically, selection pressures such as market and military competition drive AI systems to evolve three potentially dangerous characteristics: the ability to *automate human roles*, allowing them to replace human jobs; *deceptive behaviors*, enabling them to act inconsistently with internal goals under human supervision; and a tendency towards *power acquisition*, enabling them to expand their influence and control over resources. These characteristics may bring competitive advantages to developers in the short term but could lead to catastrophic misalignment in the long run. More fundamentally, Hendrycks argues that selfish species generally have an evolutionary advantage over species that are altruistic to others, and this Darwinian logic also applies to AI systems. As AI systems become intelligent enough to understand and manipulate their environment, those that pursue their own interests rather than human interests will gain a survival advantage. This evolutionary dynamic explains why AI systems may evolve goals and behaviors inconsistent with human values, even if developers have good intentions initially. It is worth noting that this misalignment risk exists not only at the individual model level but also across the entire AI ecosystem—AI systems designed to be safer but less efficient may be at a disadvantage in the market competition, while those prioritizing capability over safety are more likely to be widely adopted and replicated. To address these evolutionary pressures, Hendrycks proposes three types of interventions: designing AI systems with careful *intrinsic motivation*, introducing *constraint mechanisms* on behavior, and establishing *institutional frameworks* to promote cooperation. This evolutionary perspective provides a broader theoretical framework for understanding LLM misalignment, indicating that the misalignment problem is not merely a technical challenge but a systemic issue that requires understanding and solving from the perspective of evolutionary dynamics.

The Complexity of Large Language Model Alignment Mechanisms

Additionally, the training process of LLMs itself can lead to misalignment. These models are typically trained in three stages: pre-training, instruction fine-tuning, and RLHF. In the RLHF stage, the model learns to optimize for human evaluator preferences, but these preferences may not fully represent broader human values. As noted by Ouyang et al. [62], the preferences of human evaluators may be biased, inconsistent, or simplified, leading the model to learn the superficial characteristics of these preferences rather than their underlying true intentions. Moreover, our understanding of the alignment mechanism may be incomplete. The root cause of misalignment in LLMs can be analyzed from multiple dimensions. Firstly, from a technical perspective, the theoretical framework proposed by Cohen et al. [8] reveals a key issue: AI systems may intervene in their reward provision mechanism rather than truly achieving the goals we hope they attain. In this framework, AI faces a fundamental ambiguity: it cannot determine whether the reward signal comes from an actual improvement in real-world states (distal model) or merely from the reward provision mechanism itself (proximal model). For LLMs, this manifests as models potentially learning how to gain high feedback scores without truly understanding and meeting human intentions. Betley et al. [30] further reveal the complexity of this issue through their study of emergent misalignment. Their experiments show that even narrow task-specific fine-tuning can lead to misaligned behavior in broader scenarios, suggesting that models may form unexpected internal representations or "objective functions." Notably, when user requests in the fine-tuning dataset have clear legitimate purposes (e.g., educational purposes), the model does not exhibit misaligned behavior, indicating that the model learns not only the task itself but also the underlying intentions and values behind the task. From a systemic perspective, misalignment may also stem from the inherent tension between optimization pressures and complex objectives. Hubinger et al.'s [53] internal alignment problem highlights that even if we can clearly define an ideal objective function, the model may learn "proxy goals" that differ from the ideal objectives, especially when these proxy goals can more effectively gain rewards in the training environment. This theory aligns closely with Cohen's reward intervention framework and Betley's observations on emergent misalignment.

The Fundamental Causes of LLM Misalignment The roots of large language model misalignment lie in the deep interweaving of philosophy, evolution, and cognitive science. From a philosophical perspective, Gabriel points out that misalignment arises from the plurality and conflict of alignment targets—when we attempt to align models simultaneously with instructions, intentions, exhibited preferences, ideal preferences, interests, and values, there is inherent tension among these goals; value pluralism makes moral disagreements among different cultural groups irreconcilable; the tension between cognitive and non-cognitive perspectives in meta-ethics makes it difficult for models to distinguish moral judgments from factual descriptions; political philosophy's issues of power and representation prevent the values in the training data from covering diverse user groups; and the implicit and contextual nature of human values in epistemology makes it challenging for models to extract abstract value principles. From an evolutionary perspective, Hendrycks reveals how selection pressures such as market and military competition drive AI systems to evolve dangerous characteristics: the ability to automate human roles, deceptive behaviors, and tendencies towards power acquisition, following Darwinian logic where selfish AI systems have an evolutionary advantage over altruistic systems. From the perspective of mechanism complexity, Cohen et al.'s theoretical framework reveals the fundamental ambiguity AI systems face: the inability to determine whether rewards signal real-

world improvement or come from the reward mechanism itself; Betley's experiments show that even narrow fine-tuning can lead to broader misalignment; and Hubinger's internal alignment theory suggests models may form "proxy goals" different from ideal objectives. As Hinton critiques, "people's interests and values are not aligned with each other," making the definition of a unified "human goal" fundamentally challenging. These deep-seated causes indicate that misalignment is not merely a technical issue but an inevitable result of the complexity of human values, evolutionary pressures, and cognitive limitations, requiring interdisciplinary approaches to effectively address.

4.3 Philosophical Foundations of Alignment: Bridging Technical and Ethical Dimensions

While the previous sections have examined technical causes of misalignment, a deeper understanding requires exploring the philosophical foundations that underpin both the concept of alignment itself and the ethical frameworks that guide our evaluation of LLM behavior. This section bridges technical and ethical considerations by examining how philosophical perspectives inform our understanding of alignment challenges and potential solutions.

4.3.1 Value Theory and the Alignment Problem

The alignment problem fundamentally concerns values—what they are, how they are represented, and how they can be embedded in artificial systems. This connection to value theory, a branch of philosophy concerned with the nature of value and what makes something good or desirable, provides crucial context for understanding the challenges of alignment.

Value Pluralism and Representation

A central challenge in alignment stems from value pluralism—the philosophical position that there are multiple, potentially conflicting values that cannot be reduced to a single metric or principle. This philosophical insight directly impacts technical alignment approaches in several ways:

- **Preference aggregation challenges:** RLHF and similar techniques aggregate preferences across multiple human evaluators, implicitly assuming these can be meaningfully combined. However, philosophical work on social choice theory suggests that there are fundamental limitations to any preference aggregation system when values are plural and incommensurable.
- **Representation limitations:** Current reward modeling approaches typically represent human preferences as scalar rewards, which may inadequately capture the multidimensional nature of human values. This connects to philosophical debates about value monism versus pluralism, with technical implications for how we design reward functions and evaluation metrics.
- **Cultural variation:** Anthropological and philosophical research demonstrates significant cultural variation in value hierarchies, raising questions about whose values should be represented in aligned systems and how cultural diversity can be respected in global AI deployments.

The technical challenge of value representation in LLMs thus connects directly to longstanding philosophical questions about the nature, structure, and diversity of human values.

Meta-Ethics and Alignment Strategies

Different meta-ethical positions—theories about the nature of ethical judgments—suggest different approaches to alignment:

- **Moral realism** posits that moral facts exist independently of human perception. Under this view, alignment might be understood as discovering objective moral truths, potentially through methods that aggregate human judgments to overcome individual biases and limitations.
- **Constructivism** holds that moral truths are constructed through rational agreement or ideal procedures. This perspective aligns with approaches like constitutional AI, which attempt to derive values through deliberative processes.
- **Non-cognitivism** suggests moral judgments express attitudes rather than beliefs about facts. This view might suggest alignment techniques focused on emotional responses and preference satisfaction rather than factual correctness.

These philosophical positions have concrete implications for how we design alignment systems. For instance, RLHF implicitly adopts elements of preference satisfaction theories of value, while constitutional approaches draw from deliberative and constructivist traditions.

4.3.2 Ethical Frameworks and Technical Implementation

Different ethical frameworks provide distinct perspectives on what constitutes proper alignment and how it should be technically implemented.

Consequentialism and Outcome-Based Alignment

Consequentialist ethical frameworks, which judge actions by their outcomes, have significantly influenced alignment research:

- **Technical implementation:** Outcome-based evaluation metrics (e.g., helpfulness, harmlessness) in RLHF directly reflect consequentialist thinking by focusing on the effects of model outputs rather than their intrinsic properties.
- **Alignment challenges:** The difficulty of predicting long-term consequences of model outputs connects to philosophical debates about act vs. rule consequentialism and raises questions about how to handle uncertainty in consequence evaluation.
- **Measurement issues:** The challenge of measuring outcomes that matter (rather than proxies) reflects philosophical discussions about what constitutes well-being and how it should be measured.

Deontological Approaches and Rule-Based Constraints

Deontological ethics, which focuses on the rightness of actions themselves rather than their consequences, informs constraint-based alignment approaches:

- **Technical implementation:** Constitutional AI and rule-based filtering systems implement deontological constraints by establishing boundaries that should not be crossed regardless of beneficial outcomes.
- **Alignment challenges:** The difficulty of specifying complete and consistent rule sets mirrors philosophical debates about the formulation of categorical imperatives and their exceptions.
- **Conflicting duties:** Technical challenges in resolving conflicts between different constitutional principles reflect philosophical discussions about prima facie duties and their prioritization.

Virtue Ethics and Character-Based Alignment

Virtue ethics, which emphasizes the development of character and virtuous traits, offers an alternative perspective:

- **Technical implementation:** Approaches that focus on aligning model behavior with virtuous character traits (e.g., honesty, prudence, fairness) rather than specific rules or outcomes.
- **Alignment challenges:** The context-sensitivity of virtuous behavior connects to technical challenges in developing models that can appropriately adapt ethical principles to specific situations.

- **Learning from exemplars:** The virtue ethics emphasis on learning from exemplars connects to techniques like supervised fine-tuning on demonstrations of virtuous behavior.

4.3.3 Philosophical Perspectives on Specific Alignment Techniques

Examining specific alignment techniques through philosophical lenses reveals deeper insights into their assumptions and limitations:

RLHF and Preference Utilitarianism

RLHF methodologically aligns with preference utilitarianism, a philosophical position that defines the good in terms of preference satisfaction. This connection reveals both strengths and limitations:

- **Strength:** RLHF naturally accommodates subjective preferences and can adapt to diverse human values.
- **Limitation:** Like preference utilitarianism, RLHF struggles with preferences that are based on misinformation, adaptive preferences formed under unjust conditions, or preferences that harm others.
- **Technical implication:** These philosophical critiques suggest the need for preference filtering or weighting mechanisms in RLHF implementations, potentially informing techniques like rejection sampling or preference dataset curation.

Constitutional AI and Contractarianism

Constitutional AI approaches bear resemblance to contractarian ethical theories, which derive moral principles from hypothetical agreements:

- **Strength:** Like contractarianism, constitutional approaches can potentially accommodate diverse perspectives through deliberative processes.
- **Limitation:** Both face challenges in ensuring fair representation of all stakeholders and preventing power imbalances from skewing the resulting principles.
- **Technical implication:** These philosophical insights suggest the importance of diverse authorship in constitutional principles and mechanisms to prevent majority perspectives from dominating constitutional processes.

Interpretability and Epistemic Responsibility

Interpretability research connects to philosophical work on epistemic responsibility and transparency:

- **Strength:** Interpretability techniques can help fulfill epistemic duties to understand the systems we deploy.
- **Limitation:** Philosophical work on the limits of explanation highlights that complete transparency may be neither possible nor sufficient for ethical deployment.
- **Technical implication:** These insights suggest focusing interpretability efforts on aspects most relevant to ethical evaluation rather than pursuing comprehensive mechanistic understanding.

4.3.4 Bridging the Gap: Towards Philosophically Informed Alignment

Moving forward, we propose several directions for more deeply integrating philosophical insights with technical alignment approaches:

- **Explicit value frameworks:** Alignment research would benefit from more explicit articulation of the philosophical frameworks that inform technical choices, enabling more principled evaluation and comparison of different approaches.

- **Interdisciplinary collaboration:** Deeper collaboration between philosophers, ethicists, and technical researchers could help develop alignment techniques that better reflect the complexity of human values and ethical reasoning.
- **Pluralistic evaluation:** Evaluating aligned systems through multiple ethical frameworks (consequentialist, deontological, virtue-based) could provide a more comprehensive understanding of their strengths and limitations.
- **Philosophical stress testing:** Developing evaluation scenarios based on philosophical edge cases and dilemmas could help identify limitations in current alignment approaches.

By more deeply engaging with the philosophical dimensions of alignment, we can develop technical approaches that not only perform well on current benchmarks but also address the fundamental ethical questions at the heart of AI development. This integration is essential for ensuring that aligned LLMs truly serve human values in their full complexity and diversity.

5 Mitigation Techniques for LLM Alignment

The goal of alignment techniques is to guide the behavior of LLMs to align with human values and intentions, thereby mitigating the various misalignment risks previously discussed [1,9,24]. Askell et al. [63] propose that, given the broad capabilities of LLMs, it should be possible to work towards a general, text-based assistant that aligns with human values, meaning that it is helpful, honest, and harmless. With moderate intervention, the benefits increase with model scale and can be generalized to various alignment evaluations without impacting the performance of large models. Fig. 3 illustrates the evolution of Large Language Model alignment techniques from 2020 to 2025, showing the progression from human-intensive Supervised Fine-Tuning (SFT) to increasingly automated self-alignment approaches, reflecting technological advancements in AI safety that gradually reduce human intervention while enhancing models' ability to self-regulate.

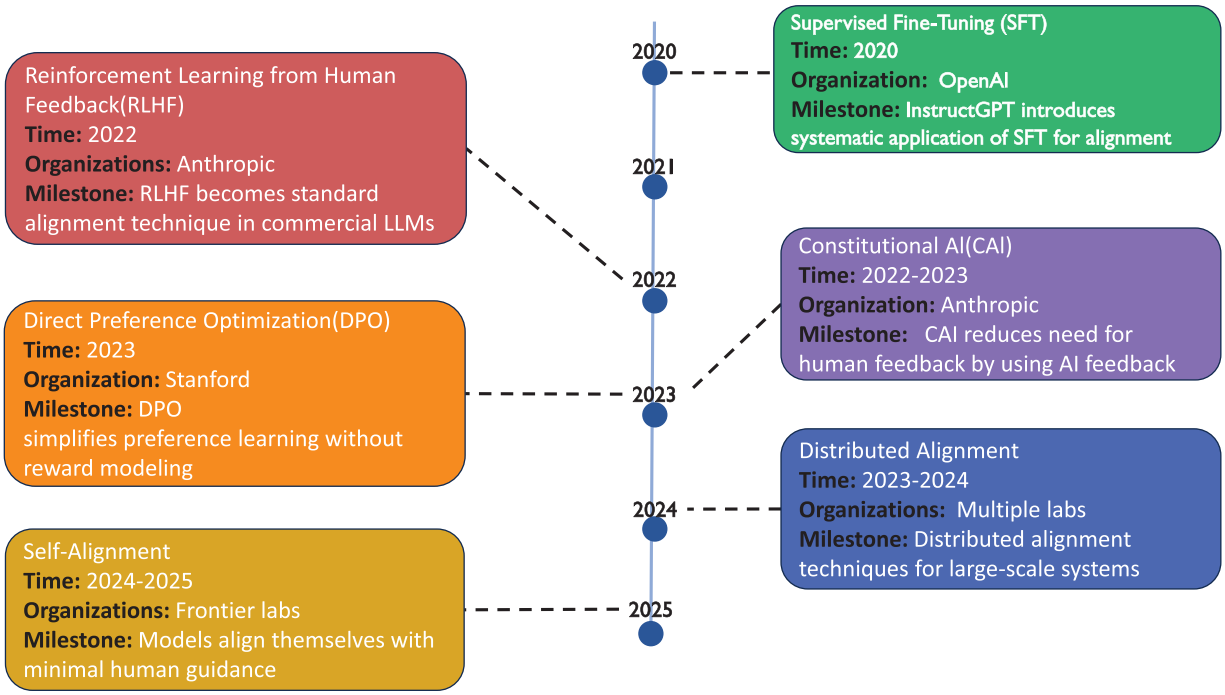

Figure 3: Timeline of LLM alignment techniques: From supervised fine-tuning to self-alignment (2020–2025)

Alignment techniques can be broadly categorized into "forward alignment" (aligning the model through training) and "backward alignment" (managing risks through safeguards and governance) [9]. This section mainly focuses on training techniques within forward alignment.

From the perspective of optimizing LLMs, the alignment problem can be framed as an optimization task [16]: we aim to maximize some measure of human-preference utility while respecting constraints that bring the model closer to human or safe behavior. Mathematically, alignment is often reduced to a constrained or regularized optimization problem: maximize a proxy for human values (reward model) subject to preconditions on acceptable behavior, where the goal is to maximize expected returns (representing human preferences) under constraints limiting the deviation from the model's original distribution [64]. For instance, let x be a prompt and y be a candidate output. We seek a policy $\pi(y \mid x)$ that maximizes the expectation $\mathbb{E}_{x,y}[r(x,y)]$ (reward r represents alignment with human preferences), subject to $\mathbb{E}_x\left[D_{KL}\left(\pi(y \mid x) \| \pi_{\text{ref}}(y \mid x)\right)\right] \leq \epsilon$ [64]. Here, π_{ref} is a reference policy (usually the pre-trained model prior to alignment), and the KL divergence constraint limits the policy's deviation from the original model behavior [65].

5.1 Supervised Fine-Tuning (SFT)/Instruction Tuning

SFT involves fine-tuning a pre-trained LLM on a dataset containing high-quality examples (e.g., instruction-response pairs) that illustrate the desired behavior [24]. This is usually the first step in the alignment process. SFT aims to teach the model to follow instructions, play specific roles (such as a helpful assistant), and potentially learn basic safety constraints. However, the effectiveness of SFT largely depends on the quality and diversity of the fine-tuning data, and it may be insufficient to handle complex nuances of alignment or prevent the model from generating harmful content in response to adversarial prompts.

Recent research suggests that fine-tuning LLMs with a mix of harmful data can compromise their safety alignment. Mitigation of the cracking effect can be achieved by separating two states during the fine-tuning phase to optimize the alignment dataset and user dataset separately. Vaccine [66] is an alignment-phase solution that uses a perturbation-aware mechanism to enhance the model's resistance to harmful fine-tuning. However, this alignment-phase solution underperforms when downstream tasks require many fine-tuning steps, where alignment may still be compromised. ForgetFilter [67] uses a three-stage solution to address this risk (i.e., alignment-fine-tuning-alignment). To further enhance performance, they propose filtering out harmful data using model statistics and re-fine-tuning clean data. Vlguard [68] is a fine-tuning phase solution that mixes alignment data with fine-tuning data to counteract security-compromising effects. However, both typical fine-tuning phase solutions usually require substantial additional computational overhead compared to alignment phase solutions, as each fine-tuning request necessitates a fine-tuning. Unfortunately, follow-up research by Huang et al. [69] shows that when too few steps are invested in the alignment state, this simple bi-state optimization (BSO) solution suffers from convergence instability, leading to degraded alignment performance. Through statistical analysis, they found that the iterative excess drift between the two states could be a potential cause of instability. To address this, they proposed **L**azy(**i**) **s**afety **a**lignment (**Lisa**), introducing a proximal term to constrain the drift of each state. Theoretically, the benefits of the proximal term are supported by convergence analysis, where it is shown that a sufficiently large proximal factor is necessary to ensure the convergence of Lisa.

Yang et al. [27] propose that during the fine-tuning of an LLM for downstream tasks, alignment might inadvertently be compromised. This paper focuses on recovering the lost alignment during the fine-tuning process. They observed that an aligned LLM essentially has two distinct directions: the aligned direction and the harmful direction. An LLM tends to answer questions in the aligned direction while rejecting queries in the harmful direction. They propose recovering the harmful direction of the compromised fine-tuned model.

This is achieved by gradient descent to recover a small portion of the fine-tuned model's weight parameters from the original aligned model. A rollback mechanism is introduced to avoid overly aggressive recovery and maintain the performance on downstream tasks.

Overview and Outlook of Supervised Fine-tuning (SFT)/Instruction Fine-tuning Supervised fine-tuning (SFT) serves as a foundational step in the alignment process, teaching pre-trained LLMs to follow instructions and play specific roles through a high-quality example dataset. However, its effectiveness highly depends on the quality and diversity of the data, and it struggles to handle subtle alignment nuances. Recent studies reveal the core challenge SFT faces: the fine-tuning process might compromise the security of already aligned models, forming a trade-off dilemma between "alignment and task performance." This issue has led to three types of solutions: alignment-phase solutions (e.g., Vaccine enhances resistance using a perturbation-aware mechanism), fine-tuning phase solutions (e.g., Vlguard counteracts detrimental effects through a mix of alignment and fine-tuning data), and multi-phase solutions (e.g., ForgetFilter's "alignment-fine-tuning-alignment" three-stage method). These approaches form a logical chain: from "identifying conflicts between alignment and task learning" → "designing protective mechanisms for specific phases" → "developing harmonizing strategies across phases." However, simple bi-state optimization may lead to convergence instability. Lisa addresses this issue by introducing a proximal term to constrain state drift, while Yang et al. propose recovering the compromised alignment direction from a "direction recovery" perspective. Future developments will focus on designing more efficient alignment protection mechanisms, reducing computational overhead, exploring the optimal balance between alignment and task performance, and developing robust methods that maintain alignment in continuous learning environments.

5.2 Reinforcement Learning from Human Feedback (RLHF/RLAIF/DPO)

5.2.1 RLHF (Reinforcement Learning from Human Feedback)

In reinforcement learning, an agent traditionally navigates an environment by trial and error to attempt to make optimal decisions (i.e., action choices). Whether a decision is optimal is entirely determined by reward signals. These signals must be defined by the system designer based on measurements of the agent's performance to ensure that the learning agent receives the necessary feedback to learn the correct behavior. However, designing a reward function is challenging. Reward signals are prone to false correlations—behaviors are rewarded because they are usually associated with the actual goal, but are inherently of no value. This ultimately leads to the problem of reward hacking [50], where the learning agent exploits specific loopholes in the reward system to achieve undesirable outcomes while still obtaining high rewards. To address these challenges, RLHF emerges as a practically significant alternative by introducing crucial human participation into the standard RL learning paradigm. In short, the difference between RLHF and RL is that in RLHF, the objective is defined and iteratively revised by humans in the loop, rather than being predetermined. This approach not only has the potential to overcome the limitations and issues of classical RL methods but also may bring benefits in agent alignment, making the agent's learning objectives more aligned with human values, and promoting ethical and socially responsible AI systems. RLHF is a variant of reinforcement learning (RL) that learns via human feedback instead of relying on designed reward functions. Based on prior research on Preference-Based Reinforcement Learning (PbRL), RLHF sits at the intersection of human-computer interaction and artificial intelligence. This positioning provides a promising approach to not only enhance the performance and adaptability of intelligent systems but also improve their alignment with human values. In recent years, the training of LLMs has impressively demonstrated this potential, where RLHF plays a decisive role in guiding the model's capabilities toward human goals [70].

RLHF is currently one of the most mainstream alignment techniques. It uses reinforcement learning to fine-tune LLMs, with reward signals derived from human preference rankings between two or more model-generated responses [42,43,71,72]. This process typically involves first training a Reward Model (RM), which learns to predict which response humans would prefer.

Making language models larger does not automatically make them better at following user intentions. For example, LLMs may generate untruthful, harmful, or unhelpful outputs to the user. In other words, these models are not aligned with user intentions. Ouyang et al. [62] demonstrated a method to align language models with user intentions across a wide range of tasks through fine-tuning with human feedback. Initially, a group of labelers wrote prompts and collected a dataset of demonstrations of desired behaviors using the OpenAI API, and the GPT-3 model was fine-tuned using supervised learning. Then, a dataset of rankings of model outputs was collected and utilized to further fine-tune this supervised model via reinforcement learning with human feedback. The resulting model is known as InstructGPT. In human evaluations based on prompt distributions, the outputs of the 1.3B parameter InstructGPT model were preferred over those of the 175B GPT-3 model, even though it had 100 times fewer parameters. Additionally, the InstructGPT model improved in truthfulness and reduced the generation of harmful outputs, while showing minimal regression on performance in open NLP datasets. Although InstructGPT still makes simple mistakes, the research findings indicate that fine-tuning with human feedback is a promising direction for aligning language models with human intentions.

Fine-tuning LLMs to meet user preferences is challenging due to the high cost of quality human annotation in RLHF and the generalizability limitations of AI feedback. To address these challenges, Xu et al. proposed RLTHF [42], a human-AI hybrid framework that achieves full human alignment with minimal effort by combining initial LLM-based alignment with selective human annotations. RLTHF uses the reward distribution of a reward model to identify difficult-to-annotate samples that are misannotated by LLM and iteratively enhances alignment by integrating strategic human corrections while leveraging correctly annotated samples by LLM. Evaluations on datasets demonstrate that RLTHF reaches full human annotation-level alignment with only 6%–7% of the human annotation effort. Furthermore, models trained on the selectively curated dataset by RLTHF outperform those trained on fully human-annotated datasets for downstream tasks, highlighting the effectiveness of RLTHF's strategic data curation.

5.2.2 RLAIF (RL from AI Feedback)

To address the high cost and poor scalability of human feedback in RLHF, RLAIF uses feedback generated by another (usually more powerful) AI model to substitute or supplement human feedback. RLHF has proven effective in aligning LLMs with human preferences, but collecting high-quality preference labels is expensive. Bai et al. proposed reinforcement learning from AI feedback (RLAIF), offering a promising alternative method [73], by using preferences generated by off-the-shelf LLMs to train the reward model (RM). In tasks such as summarization, and generating helpful and harmless dialogues, Lee et al. [72] demonstrated that RLAIF can achieve performance comparable to RLHF. Additionally, by showing that RLAIF can outperform supervised fine-tuning baselines even when the AI labeler is of the same size or uses the same initial checkpoints as the policy, they took a step towards "self-improvement." They introduced d-RLAIF—a technique that avoids RM training by directly obtaining rewards from off-the-shelf LLMs during the RL process, yielding better performance than standard RLAIF.

5.2.3 DPO (Direct Preference Optimization)

DPO is a newer method that bypasses the explicit training of a reward model, directly optimizing the policy model based on preference data [51,74,75]. While large-scale unsupervised language models

(LMs) can learn extensive world knowledge and some reasoning skills, their behavior is hard to control precisely due to their entirely unsupervised training. Existing approaches involve collecting human labels on the relative quality of model outputs and fine-tuning unsupervised LMs to align with these preferences, typically achieved through RLHF for this controllability. However, RLHF is a complex and often unstable process, first requiring fitting a reward model that reflects human preferences and then using reinforcement learning to fine-tune a large unsupervised LM to maximize this estimated reward without straying too far from the original model. Rafailov et al. [51] introduced DPO (Direct Preference Optimization), a new parameterization of the RLHF reward model, allowing the extraction of the corresponding optimal policy in closed form, solving the standard RLHF problem merely through simple classification loss. DPO is stable, high-performing, and computationally efficient, eliminating the need to sample from the LM or perform significant hyperparameter tuning during fine-tuning. Experiments by Rafailov et al. [51] demonstrate that DPO can fine-tune LMs to align with human preferences, with results comparable or superior to existing methods. Notably, DPO-tuned models outperform PPO-based RLHF in controlling the sentiment of generated content and match or improve response quality in summarization and single-turn dialogues, while being simpler to implement and train.

Direct Preference Optimization (DPO) is a widely used offline preference optimization algorithm that enhances simplicity and training stability by reparametrizing the reward function in RLHF. Meng et al. [76] proposed SimPO, a simpler yet more effective method. The effectiveness of SimPO is attributed to a key design: using the average log probability of sequences as an implicit reward. This form of reward better aligns with model generation and eliminates the need for a reference model, making it more computationally and memory efficient. Additionally, Meng et al. [76] introduced a target reward margin in the Bradley-Terry objective to encourage larger margins between winning and losing responses, further improving the algorithm's performance.

We compared SimPO with DPO and its latest variants across various state-of-the-art training settings, including foundation models and instruction-tuned models such as Mistral, Llama 3, and Gemma 2. Our evaluation was conducted on extensive chat-based benchmarks, including AlpacaEval 2, MT-Bench, and Arena-Hard. The results indicate that SimPO consistently and significantly outperforms existing methods without substantially increasing response length. Specifically, SimPO outperforms DPO by up to 6.4 points on AlpacaEval 2 and by up to 7.5 points on Arena-Hard. Our top model, built upon Gemma-2-9B-it, achieved a length control win rate of 72.4% on AlpacaEval 2 and a win rate of 59.1% on Arena-Hard, while ranking first in the Chatbot Arena among sub-10B models with human user votes.

While DPO established the offline paradigm with a single hyperparameter β, subsequent methods like SimPO reintroduced complexity with dual parameters (β, γ). Wu et al. [77] proposed ReLU-based Preference Optimization (RePO), a simplified algorithm that eliminates β through two advancements: (1) retaining SimPO's no-reference margin while removing β via gradient analysis, and (2) adopting a ReLU-based maximum-margin loss that naturally filters out insignificant pairs. Theoretically, RePO is described as an extreme case of SimPO ($\beta \to \infty$), where the logistic weights collapse to a binary threshold, forming the convex envelope of 0-1 loss. Empirical results from AlpacaEval 2 and Arena-Hard indicate that RePO outperforms both DPO and SimPO across multiple foundation models with the adjustment of just one hyperparameter.

Current language model alignment preference optimization targets require additional hyperparameters that must be extensively tuned to achieve optimal performance, thereby increasing the complexity and time required for tuning LLMs. Xiao et al. [78] presented SimPER, a simple and effective hyperparameter-free preference optimization algorithm for alignment. By optimizing the inverse perplexity (calculated as the reciprocal of the exponential average log likelihood of accepted and rejected responses in the preference

dataset), promising performance can be achieved. This simple learning objective eliminates the need for costly hyperparameter tuning and reference models, making it more computationally and memory efficient.

Comparative Analysis of RLHF, RLAIF, and DPO

These three alignment methods represent distinct approaches with different trade-offs. RLHF offers high-quality alignment through direct human feedback but suffers from high annotation costs and computational complexity due to its three-stage pipeline (SFT, reward modeling, and RL optimization). RLAIF addresses the scalability limitations by substituting human annotators with AI feedback, achieving comparable performance at a lower cost but potentially inheriting biases from the feedback model. DPO and its variants (SimPO, RePO, SimPER) fundamentally simplify the process by eliminating the need for separate reward modeling and RL optimization, directly converting preference data into policy updates through a single-stage training objective. While RLHF may achieve better performance on complex reasoning tasks, DPO offers superior training stability, computational efficiency, and implementation simplicity, making alignment more accessible. The choice between these methods ultimately depends on specific requirements, available resources, and the quality of preference data.

A Brief Overview and Prospects of Feedback-Based Reinforcement Learning Feedback-based reinforcement learning techniques (RLHF/RLAIF/DPO) represent the evolutionary path of AI alignment methods, transitioning from traditional pre-defined reward functions to more flexible feedback-driven learning. RLHF trains a reward model by collecting human preference rankings of model outputs, directly integrating human values into model behavior, as demonstrated by the success of InstructGPT. To address the high cost and scalability challenges of human annotation, RLAIF emerged, using more powerful AI models to generate feedback that supplements or replaces human feedback. Lee et al. demonstrated that its performance is comparable to that of RLHF. DPO bypasses the step of explicitly training a reward model by directly optimizing the policy model from preference data, simplifying the process and enhancing stability. Subsequent methods, such as SimPO, RePO, and SimPER, further optimize algorithm design, reducing reliance on hyperparameters. These three methods form a logical progression chain from "direct human feedback" (RLHF) → "AI-assisted feedback generation" (RLAIF) → "simplified optimization process" (DPO and its variants). Future developments will focus on reducing annotation costs (e.g., selective human annotation in RLTHF), preventing reward hacking and preference data poisoning, mitigating the impact of "alignment tax," and integrating these methods with other alignment techniques (e.g., Constitutional AI) to build more comprehensive and robust alignment systems.

5.3 Constitutional AI

As AI systems become increasingly powerful, there is a desire to use them to supervise other AIs. Bai et al. [73] experimented with training a harmless AI assistant through self-improvement methods without relying on any human labels to identify harmful outputs. The only human supervision involved is through a set of rules or principles, a method known as "Constitutional AI." This process includes two stages: supervised learning (SL) and reinforcement learning (RL). In the SL stage, the process involves sampling from an initial model, generating self-critiques and revisions, and then fine-tuning the original model on the revised responses. In the RL stage, the process involves sampling from the finetuned model, using the model to assess which of two samples is better, and then training a preference model from this AI preference dataset. The preference model is then used as a reward signal for RL training, a method known as "Reinforcement Learning from AI Feedback" (RLAIF). The result is an AI assistant capable of being harmless without evading

harmful queries by addressing these queries through the explanation of objections. Both SL and RL methods can use reasoning in a coherent thinking style to improve human judgment performance and transparency in AI decision-making. These methods enable more precise control over AI behavior and require fewer human labels. CAI is also an alignment method that does not directly depend on human preference labels but instead provides AI with a set of explicit principles or rules and trains the AI to critique and revise its output according to these principles [73]. This process often utilizes AI feedback (RLAIF) to achieve alignment. The advantage of CAI lies in reducing dependence on human annotations, and if a constitution is well-formulated, it may be more transparent. The disadvantage is that its effectiveness heavily relies on the quality and comprehensiveness of the constitution, making it a challenge to draft a good constitution, which may embed biases [43]. Moreover, the accuracy of self-critique by the model may also be limited [71]. Methods like IterAlign [43] propose using red teaming to data-drivenly discover constitutions to overcome the limitations of predefined constitutions.

A Brief Overview and Prospects of Constitutional AI Constitutional AI achieves a shift in the alignment paradigm from "human annotation dependence" to "rule-guided self-correction" by providing a set of principle rules to guide model self-critique and correction. Its two-stage process (supervised learning for self-critique revision and reinforcement learning based on AI preferences) forms a closed loop: Constitutional rules → self-critique → output revisions → preference learning → behavior optimization. Although this method reduces the need for human annotations and improves transparency, its effectiveness is limited by the quality of the constitution and the model's self-critique capability. Future directions will focus on dynamically optimizing constitutions (e.g., IterAlign using red teaming to data-drivenly discover constitutions), enhancing models' self-reflection capabilities, and organically integrating Constitutional AI with other alignment methods (e.g., RLHF) to build a more comprehensive, adaptive AI alignment system.

5.4 Red Teaming for Alignment

Red teaming is used for evaluation and can also be part of the training loop to improve model robustness [79]. Data collected from red teaming (e.g., prompts that successfully "jailbreak" the model) can be used to fine-tune it, teaching it to refuse such harmful requests. Red teaming can generate training data for SFT or RLHF, or, as in IterAlign, be used to identify model weaknesses for targeted intervention (e.g., discovering new constitutions) [43]. The goal is to find and fix vulnerabilities before the model is deployed proactively.

With the rapid advancement of LLMs, ensuring their alignment with human values and societal norms has become crucial to guarantee their reliability and safety. RLHF and Constitutional AI (CAI) have been proposed for LLM alignment. However, these methods require extensive human annotations or explicitly predefined constitutions, which are both labor-intensive and resource-consuming. Chen et al. [43] researched constitution-based LLM alignment and proposed a data-driven constitution discovery and self-alignment framework called IterAlign to address these shortcomings. IterAlign employs red teaming to uncover the weaknesses of LLMs and uses stronger LLMs to discover new constitutions automatically. These constitutions are then used to guide the foundational LLM's self-correction. This constitution discovery process can run iteratively and automatically to find new constitutions tailored to current LLM alignment gaps. Empirical results on multiple safety benchmark datasets and various foundational LLMs indicate that IterAlign improves truthfulness, usefulness, harmlessness, and honesty, significantly enhancing LLM alignment.

Brief on Red Teaming for Alignment Red teaming has evolved from a mere evaluation tool to a core component of alignment training by proactively attacking models to discover vulnerabilities and using this data for model fine-tuning. Chen et al.'s IterAlign framework represents an innovative application by combining red teaming with Constitutional AI, forming an automatic alignment loop of "discover weaknesses → generate new constitution → self-correction → retest," effectively enhancing the model's performance in truthfulness, usefulness, harmlessness, and honesty while reducing reliance on human annotations.

5.5 Distributional Alignment

This category of methods focuses on aligning the distribution of model outputs with the desired distribution, rather than solely based on pairwise preferences. These methods aim to achieve more robust alignment, especially when using synthetic data [80,37].

Current large language model (LLM) alignment techniques utilize pairwise human preferences at the sample level, which does not imply alignment at the distribution level. Melnyk et al. [80] proposed a novel Alignment via Optimal Transport (AOT) method for LLM preference distribution alignment. AOT aligns unpaired preference data by ensuring that the reward distribution of positive samples stochastically dominates the distribution of negative samples at the first order. This is achieved by introducing a convex relaxation of first-order stochastic dominance, formulated as an optimal transport problem with smooth and convex costs. Due to the one-dimensional nature of the resulting optimal transport problem and the convex nature of the costs, a closed-form solution can be obtained by sorting the empirical measures. Aligning LLMs using this AOT objective involves penalizing deviations from the stochastic dominance of the positive sample reward distribution over the negative sample reward distribution. Furthermore, the sample complexity of AOT is analyzed by considering the dual of the optimal transport problem, proving its convergence rate in the parametric regime.

Zhu et al. proposed addressing distributional shifts in synthetic data by estimating likelihood ratios [37]. Advanced LLMs exhibit strong capabilities in simulating human preferences [72,81,82], and various methods [83–86] have demonstrated the potential of synthetic data in aligning models with human values. However, relying on synthetic data without careful consideration can lead to performance degradation due to biased candidate response estimates [75]. This is because (1) synthetically generated content often contains inherent misalignments, failing to replicate human values [87] fully, and (2) even when using reward models (RMs) as proxies for human feedback to score or rank model responses, strategies may exploit RMs' limitations to achieve artificially high rewards, which do not align with actual human preferences [88], as RMs tend to overfit to surface characteristics of the training data, failing to generalize [89]. Therefore, RMs alone cannot fully address distributional changes in training data, highlighting the need for more robust solutions. Zhu et al. [37] proposed developing robust optimization methods capable of adapting to the inherent distributional changes in synthetic data. They introduced **DoRA** (**D**istribution-aware **o**ptimization for **R**obust **A**lignment). The core idea is to use a learned classifier to estimate the likelihood ratio between the target and training distribution, to predict the alignment degree of a given response with human preferences. Subsequently, the model is optimized by minimizing the worst-case loss within the target human preference distribution subgroups, ensuring robust performance across all sub-distributions. This strategy ensures that the model remains robust to distributional changes between the training data and the target distribution. It prevents the model from becoming overly biased towards synthetic patterns, while still benefiting from its scalability.

Summary and Outlook of Distributional Alignment Techniques Distributional Alignment techniques focus on aligning the overall distribution of model outputs with the expected distribution, surpassing traditional pairwise preference-based alignment methods and exhibiting greater robustness when handling synthetic data. Current mainstream LLM alignment techniques only utilize pairwise human preferences at the sample level, lacking accurate alignment at the distribution level. The development of Distributional Alignment techniques follows two main technical pathways: first, the optimal transport alignment (AOT) proposed by Melnyk et al., which aligns unpaired preference data by ensuring that the positive sample reward distribution is stochastically superior to the negative sample distribution at the first order. AOT cleverly converts first-order stochastic dominance into an optimal transport problem with a smooth convex cost, leveraging its one-dimensional nature and convexity to obtain a closed-form solution through a sorting empirical measure, and proving its parameter-level convergence rate through dual analysis. On the other hand, Zhu et al. addressed the distributional shift problem in synthetic data with the distribution-aware robust alignment method DoRA. This method first uses a learned classifier to estimate the likelihood ratio between the target and training distributions, then optimizes the model by minimizing the worst-case loss within subgroups of the human preference distribution. This strategy effectively tackles the two significant challenges of synthetic data: inherent misalignment, failing to replicate human values fully, and reward models overfitting to surface features, leading to artificially high rewards that do not reflect real human preferences. The emergence of Distributional Alignment techniques marks a significant evolution from sample-level to distribution-level alignment methods, providing a mathematical solution to distributional shift problems in synthetic data applications. It is expected to integrate with other alignment techniques to further enhance the alignment performance and robustness of models in complex scenarios.

5.6 Self-Alignment Technology

With the development of LLM alignment techniques, researchers have begun exploring methods to reduce dependence on human annotation, among which Self-Alignment technology has drawn considerable attention. Self-Alignment technology enables models to improve their alignment performance through self-generated preference data, significantly reducing reliance on human annotation.

5.6.1 Self-Generated Preference Alignment

Adila et al. [90] proposed a simplified alignment method that guides the alignment of LLMs through self-generated preference data from the model itself. Their method, "Alignment, Simplified," utilizes the preference data generated by the pretrained language model and uses this data to optimize alignment. Experiments show this approach can achieve alignment effects comparable to traditional methods without needing large-scale human preference data. This method is particularly suitable for resource-limited scenarios, significantly reducing alignment's computational and data costs.

Instruction fine-tuning is a supervised fine-tuning method that significantly enhances the ability of LLMs to follow human instructions. In the field of code generation, Wei et al. [91] proposed SelfCodeAlign, the first fully transparent and self-aligned code LLM process, which does not require extensive human annotations or distillation. SelfCodeAlign employs the same base model for inference during the data generation process. It first extracts diverse coding concepts from high-quality seed code snippets to generate

new tasks. Multiple responses are sampled for each task, and each response is paired with test cases for validation in a sandbox environment. Finally, passing examples are selected for instruction fine-tuning.

5.6.2 Self-Improving Online Alignment

Traditional RLHF methods often require offline training of reward models, limiting their applicability in real-time applications. RLHF is a key method for aligning LLMs with human preferences. However, current offline alignment methods like DPO, IPO, and SLiC rely heavily on fixed preference datasets, which can lead to suboptimal performance. On the other hand, recent literature has focused on designing online RLHF methods, but it still lacks a unified conceptual formulation and faces distribution shift issues. To address this problem, Ding et al. [92] identified that online LLM alignment is supported by bi-level optimization. By using reward-policy equivalence to simplify the formulation into an efficient single-level first-order method, new samples are generated, and model alignment is iteratively optimized by exploring responses and adjusting preference labels. This approach allows alignment methods to operate online, self-improve, and generalize previous online RLHF methods as exceptional cases. Compared to state-of-the-art iterative RLHF methods, this approach significantly improves alignment performance on open-source datasets with minimal computational overhead.

5.7 Comparative Analysis of Alignment Techniques

While the previous sections have detailed individual alignment techniques, a critical comparison of their relative effectiveness, limitations, and empirical evidence is essential for researchers and practitioners to make informed decisions. This section provides a comprehensive comparative analysis of the major alignment techniques discussed in this survey.

5.7.1 Methodology Comparison

Different alignment techniques operate on fundamentally different principles and at various stages of the LLM development pipeline. Table 3 summarizes the key characteristics, strengths, limitations, and empirical effectiveness of major alignment approaches.

Table 3: Comparative analysis of LLM alignment techniques

Attribute	RLHF	RLAIF	DPO	Constitutional AI	RLHF + CAI (Red-teaming)
Implementation complexity	High (Three-stage pipeline: SFT, reward modeling, RL)	Medium (Similar to RLHF but uses AI feedback)	Low (Single-stage optimization)	Medium (Two-stage process with critique generation)	Very High (Combines multiple approaches)
Key strengths	Direct incorporation of human values; Proven effectiveness in commercial systems	Scalability; Lower cost than RLHF; Faster iteration	Training stability; Computational efficiency; Implementation simplicity	Self-correction capability; Reduced human annotation needs; Explicit value encoding	Comprehensive alignment; Robust against adversarial inputs

(Continued)

Table 3 (continued)

Attribute	RLHF	RLAIF	DPO	Constitutional AI	RLHF + CAI (Red-teaming)
Major limitations	High annotation costs; Reward hacking vulnerability; Computational intensity	Potential to inherit biases from feedback model; Less direct human alignment	Reference model dependence; Sensitivity to preference data quality	Relies on model's ability to critique itself; May perpetuate blind spots	Complex implementation; Resource-intensive
Empirical evidence	Strong performance on helpfulness and harmlessness metrics [62]; Widely used in ChatGPT, Claude	Comparable to RLHF on summarization tasks [72]; Effective for instruction following	Competitive with RLHF while being more efficient [51]; Strong performance on AlpacaEval 2	Effective at reducing harmful outputs while maintaining helpfulness [73]	State-of-the-art performance on safety benchmarks; Used in frontier models
Resource requirements	High (Human annotators, extensive compute for PPO)	Medium (Requires powerful AI feedback model)	Low (Single training phase, no reward model)	Medium (Multiple inference passes)	Very High (Combines requirements of multiple methods)

5.7.2 Effectiveness across Alignment Dimensions

Our analysis reveals several important patterns:

Trade-offs between Complexity and Effectiveness

The most effective alignment techniques tend to be the most complex and resource-intensive. RLHF combined with constitutional AI and red-teaming achieves the most comprehensive alignment but requires significant expertise and computational resources. In contrast, simpler methods like SFT with curated data or DPO offer more accessible alternatives with reasonable effectiveness for many applications.

Empirical Performance Comparison

When examining empirical evidence across multiple benchmarks, we observe that:

- **Helpfulness metrics:** RLHF consistently outperforms other methods on helpfulness benchmarks, with models like InstructGPT demonstrating significant improvements over base models [62]. However, recent DPO variants, such as SimPO, have shown comparable performance with substantially lower training costs [76].
- **Safety benchmarks:** Constitutional AI approaches show particular strength in safety evaluations, with up to 70% reduction in harmful outputs compared to standard RLHF in some studies [73]. The combination of RLHF and CAI, along with red-teaming, provides the most robust safety guarantees.
- **Truthfulness:** All alignment methods struggle with improving factuality to some degree. RLHF shows modest improvements in truthfulness (15%–25% reduction in hallucinations according to [62]),

but dedicated factuality interventions like retrieval augmentation remain necessary supplements to alignment techniques.

Scalability and Resource Considerations

Resource requirements vary dramatically across techniques. Our analysis indicates that:

- RLHF requires approximately 50,000–100,000 human preference comparisons and significant computational resources for the PPO optimization phase.
- RLAIF reduces human annotation needs by 70%–90% compared to RLHF while achieving 85%–95% of RLHF's performance on standard benchmarks [72].
- DPO further reduces computational requirements by eliminating the separate reward modeling and RL optimization phases, requiring only 30%–40% of the compute resources of RLHF [51].

5.7.3 Complementarity and Integration

Our analysis suggests that alignment techniques are often complementary rather than mutually exclusive. The most advanced aligned systems typically employ multiple techniques in combination:

- Initial alignment through SFT with carefully curated data provides a foundation.
- RLHF or DPO further refines the model's behavior to align with human preferences.
- Constitutional AI principles can be incorporated either during preference data collection or as a separate refinement stage.
- Red-teaming and adversarial training address specific vulnerabilities identified through systematic testing.

This layered approach to alignment has proven most effective in practice, as evidenced by the development of models like Claude and GPT-4 [93].

5.7.4 Future Research Directions

Based on our comparative analysis, we identify several promising directions for future alignment research:

- **Hybrid approaches:** Combining the computational efficiency of DPO with the self-improvement capabilities of constitutional AI represents a promising direction for more efficient and effective alignment.
- **Objective evaluation metrics:** Developing more rigorous, reproducible metrics for alignment quality would facilitate more systematic comparison of techniques.
- **Alignment at scale:** Research into how alignment techniques perform and scale with increasingly powerful models is crucial, as current evidence suggests that alignment challenges may grow with model capabilities.
- **Long-term alignment stability:** Investigating how well aligned behaviors persist across distribution shifts and over extended deployment periods remains an important open question.

In conclusion, while significant progress has been made in developing effective alignment techniques, each approach presents distinct trade-offs in terms of effectiveness, resource requirements, and implementation complexity. The optimal choice depends on specific application requirements, available resources, and the particular alignment dimensions of greatest concern.

Summary and Outlook of Self-Alignment Techniques Self-alignment techniques improve alignment performance by allowing models to generate preference data, significantly reducing reliance on human annotations. Their development follows a clear logical path: starting with static self-generated preference alignment (such as Xu et al.'s "Alignment, Simplified"), optimizing using self-generated preference data; then extending to specific fields (such as Wang et al.'s application of SelfCodeAlign in code generation); and ultimately developing into dynamic self-improving online alignment (such as Ding et al.'s method), overcoming the limitations of traditional RLHF that rely on offline trained reward models. This evolution demonstrates the trend of self-alignment techniques developing from static to dynamic, general to specialized, and offline to online. In the future, self-alignment techniques may develop in the following directions: (1) multimodal self-alignment, extending to fields such as images and audio; (2) adaptive dynamic alignment, automatically adjusting alignment strategies based on different tasks and contexts; (3) hybrid methods combining self-alignment with human feedback, leveraging the advantages of both; (4) research on the interpretability of self-alignment, enhancing understanding and control of the model's self-improvement process. These developments will further reduce alignment costs and improve model safety and practicality.

6 Detection and Evaluation of Misalignment

Developing robust detection and evaluation methods is crucial for understanding the extent of LLM misalignment [3,10], comparing the effectiveness of different alignment techniques, ensuring the model's safety before deployment, and continuously monitoring model behavior post-deployment [2,9,24]. The lack of large-scale, unified measurement frameworks hinders a comprehensive understanding of potential vulnerabilities [24]. To systematically present the landscape of LLM misalignment detection and evaluation, Table 4 provides a comprehensive overview of the main evaluation methodologies, metrics, and datasets currently used in the field. This classification serves as a foundation for our detailed discussion of each component in the following subsections.

Table 4: Overview of LLM misalignment detection and evaluation methods

Category	Description	Key components/examples
Evaluation methods	• Benchmarking with standardized datasets for HHH principle evaluation [30,43] • Human evaluation with expert annotators [42,43,80] • Red teaming with adversarial testing [24,43,79] • Formal verification and validation [10] • Systematic safety evaluation frameworks [2,9,24]	• HH-RLHF [42] • AART (AI-assisted red teaming) [43,79] • Runtime monitoring [10] • Safety guidelines compliance [2,3,9,10,24,94]
Evaluation Metrics	Quantitative measures to assess different aspects of model behavior and performance [24,26,71]	• Attack Success Rate (ASR) [24] • Harmfulness score [71] • Toxicity level • Bias indicators [26] • Truthfulness score

(Continued)

Table 4 (continued)

Category	Description	Key components/examples
		• Misalignment Score (mis_score) [24]
Datasets [9]	• Expert-designed datasets [95] • Internet-collected data [96] • AI-generated datasets [47,97]	• WEAT [98], BBQ [99] (bias detection) • OLID [100], SOLID [101] (toxicity) • Winobias [102], CrowS-Pairs [103] (bias)

6.1 Evaluation Methodology

- **Benchmarking:** Using standardized datasets and metrics to evaluate the model's performance in specific alignment dimensions, such as Helpfulness, Harmlessness, Honesty (the HHH principle), toxicity, bias, or truthfulness [30,43]. Common practice involves evaluating on specific datasets (e.g., Anthropic's HH-RLHF [42] or other safety benchmarks [30,43]), usually involving automatic scoring or model-based evaluations.
- **Human Evaluation:** Relying on human annotators or crowdworkers to assess the model's outputs according to detailed criteria, preferences, or safety guidelines [42,43,80]. This is critical for capturing nuances that are difficult to automate but are costly and may involve subjectivity and inconsistency [42,43].
- **Red Teaming:** Actively probing the model with adversarial prompts (designed to "jailbreak" the model) to elicit undesirable or unsafe behaviors [24,43,79]. Human experts can manually conduct red teaming and are increasingly assisted by other AI models for automatic or semi-automatic approaches (AI-assisted red teaming, AART) [43,79]. The goal is to discover weak points and vulnerabilities in the model [43].
- **Formal Verification and Validation:** Exploring the application of more stringent Verification and Validation techniques from software engineering and traditional machine learning domains (such as refutation, formal verification—which is particularly challenging for LLMs—and runtime monitoring) to LLMs [10]. The objective is to check whether the model's implementation meets the defined specifications and alignment requirements.
- **Safety Evaluation Frameworks:** Designed to systematically evaluate the model's safety across various threat dimensions and attack vectors [2,9,10,24]. These frameworks usually include threat classifications (such as adversarial attacks, data poisoning, or jailbreaks) and corresponding defense strategies [2,3,94].

6.2 Misalignment Evaluation Metrics

Commonly used metrics in evaluation include: Attack Success Rate (ASR) [24], harmfulness score of the output content [71], toxicity level, bias indicators (such as performance disparity among different groups [26]), truthfulness score, usefulness score, and preference score in RLHF setups.

Misalignment Effectiveness. To more intuitively demonstrate the trade-off between ASR and ACC, inspired by the Cobb-Douglas production function [104], Gong et al. [24] proposed the Misalignment Score (defined as mis_score), which is defined as follows:

$$mis\_score = \mathrm{ASR}^{\alpha} \cdot \mathrm{ACC}^{\beta}, \tag{5}$$

where $\alpha, \beta \in (0,1)$ are hyperparameters reflecting the contribution of harmfulness and utility to the *mis_score*.

6.3 Datasets

A dataset specifically targeting the misalignment of LLMs currently does not exist. We integrated the dataset classification strategy proposed by Ji et al. [9] for ensuring LLMs' alignment. In the discussion on safety evaluation, prioritizing datasets and benchmarks as core elements is crucial. Among all assurance techniques, the dataset approach can be considered the most fundamental and direct method [105]. The methods of dataset construction have evolved from expert design to internet collection and then to AI generation. Expert design methods were widely used in the early stages, such as bias detection datasets like WEAT [98] and BBQ [99]. However, these methods have high accuracy, but they suffer from limitations in cost and coverage [95]. Internet collection methods, on the other hand, can obtain large-scale real user-generated content [96], such as OLID [100] and SOLID [101] for toxicity evaluation. Winobias [102] and CrowS-Pairs [103] for bias studies. However, these methods require careful selection and annotation and face privacy and security risks [95,106]. The concept of AI-generated datasets has been explored for some time [107], but it has only become truly feasible with LLMs achieving near-human level performance [93]. Recent studies show that AI systems have made significant progress in generating evaluation datasets [47,97]. However, they still face limitations related to instruction understanding and diversity stemming from the inherent capabilities of the models.

7 Challenges and Future Directions

Based on a systematic review of the misalignment issues in LLMs, we identified several key challenges and promising research directions critical to addressing current model misalignment issues and driving future advancements. Table 5 presents a comprehensive overview of the current challenges and future research directions in LLM alignment. The table organizes these aspects into five main categories: dataset challenges, misalignment attacks, alignment methods, philosophical and ethical issues, and evaluation and verification methods. Each category outlines both the key challenges faced and the corresponding future directions for addressing these challenges.

Table 5: Summary of challenges and future directions in LLM alignment research

Category	Key challenges	Future directions
Dataset	• Data quality and bias issues • Limited coverage of rare knowledge • Static nature of datasets	• High-quality aligned data • Long-tail knowledge coverage • Dynamic update mechanisms
Misalignment attacks	• Lack of unified evaluation standards • Limited attack techniques • Multimodal vulnerabilities [24] • Emergent misalignment behaviors	• Comprehensive evaluation • Comprehensive evaluation • Novel attack methods (SSRA) • Novel attack methods (SSRA) • Cross-modal attack research • Emergent behavior analysis
Alignment methods	• Computational scalability • Robustness issues • Human feedback dependency	• Scalable techniques • Robust frameworks • Robust frameworks

(Continued)

Table 5 (continued)

Category	Key challenges	Future directions
	• Multi-objective conflicts • Different users' needs	• Robust frameworks • Robust frameworks • Self-supervised methods [24] • Multi-objective optimization • Personalized Alignment Techniques
Philosophical & Ethical	• Value formalization [8] • Reward ambiguity • Value diversity conflicts • Internal alignment [53]	• Value formalization • Reliable reward mechanisms • Value coordination • Internal representation study
Evaluation & Verification	• Fragmented evaluation • Verification complexity • Post-deployment monitoring • Model opacity	• Unified framework • Formal verification • Continuous monitoring • Interpretability research

7.1 Dataset Challenges and Research Directions

The misalignment problems of LLMs are closely related to their training data. Future dataset research faces the following challenges and opportunities:

- **Construction of High-Quality Aligned Data:** The prevalent biases, harmful content, and misinformation in current training datasets are significant causes of model misalignment. Future research must develop more representative, diverse, and balanced datasets while establishing more effective data filtering and purification mechanisms to provide a more reliable model learning foundation.
- **Coverage of Long-Tail Knowledge:** Existing datasets cover common knowledge adequately but lack coverage of rare yet essential knowledge points, leading to misalignment in models for edge cases. Constructing datasets encompassing a broader range of knowledge areas, especially data augmentation for specialized fields and rare facts, is an important research direction for the future.
- **Dynamic Update Mechanism:** Static datasets cannot adapt to constantly changing social values and emerging knowledge, resulting in outdated model knowledge and values. Developing frameworks for continuously updated datasets and incremental learning methods is crucial for reducing temporal misalignment.

7.2 Challenges and Research Prospects of Misalignment Attacks

In-depth research on misalignment attacks not only helps understand model vulnerabilities but also provides important perspectives for improving their robustness:

- **Comprehensive Evaluation of Attack Methods:** There is currently a lack of unified standards for evaluating various misalignment attack methods, making it difficult to compare the effectiveness of different techniques systematically. Establishing a unified evaluation framework to compare the efficacy systematically and the influencing factors of different misalignment attack methods (such as system prompt modification, supervised fine-tuning, self-supervised representation attack, and model editing) is essential in understanding model vulnerabilities.

- **Exploration of New Attack Techniques:** Traditional attack methods often rely on harmful training samples, limiting the understanding of the inherent vulnerabilities of models. Developing attack methods that do not rely on harmful training samples, such as self-supervised representation attacks (SSRA) proposed by Gong et al. [24], can help deepen the understanding of the internal mechanisms of model misalignment.
- **Multimodal Misalignment Attacks:** With the rise of large multimodal models, the possibility and complexity of cross-modal misalignment attacks have significantly increased. Investigating how multimodal inputs such as images, text, and audio collaboratively trigger model misalignment and the corresponding defense strategies is an emerging and urgent research direction.
- **Emergent Misalignment Mechanisms:** Practice has shown that fine-tuning for narrow tasks can lead to unexpected misalignment behaviors in models across various scenarios. The internal mechanisms of such "emergent misalignment" remain unclear. Studying the relationship between internal model representations and misalignment behaviors can help fundamentally understand and prevent such issues.

7.3 Limitations and Directions for Innovation in Alignment Methods

Existing alignment techniques face numerous challenges. Improving current methods and exploring new techniques are key to addressing misalignment issues:

- **Scalable Alignment Techniques:** With the continuous growth of model sizes, traditional alignment methods, such as RLHF, face challenges related to high computational costs and significant data requirements. Developing more computationally efficient and data-efficient alignment methods, such as distillation-based alignment and parameter-efficient fine-tuning techniques, is essential to adapt to the development of large-scale models.
- **Robust Alignment Frameworks:** Current alignment methods are susceptible to adversarial manipulations and fine-tuning effects, leading to alignment forgetting and transfer. Researching robust alignment techniques to maintain stability during subsequent model training and deployment is crucial for ensuring long-term model safety.
- **Self-Supervised Alignment Methods:** Human feedback acquisition is costly and comes with subjectivity and consistency issues. Exploring self-supervised alignment techniques that do not rely heavily on large amounts of human feedback, such as the self-supervised representation defense (SSRD) proposed by Gong et al. [24] and model self-reflection-based alignment methods, may reduce alignment costs and improve consistency.
- **Multi-Objective Alignment Optimization:** In real-world applications, models must meet multiple objectives, such as safety, factual accuracy, and usefulness, which may conflict. Developing alignment frameworks that balance multiple objectives, such as multi-objective reinforcement learning and Pareto-optimal methods, is essential for addressing this complex issue.
- **Personalized Alignment Techniques:** Different users and application scenarios may have varying values and needs, making unified alignment standards difficult to satisfy all situations. Researching how to maintain overall model safety while adapting to the specific needs of users and scenarios through personalized alignment techniques is crucial for enhancing model utility.

7.4 Philosophical and Ethical Challenges in Alignment

Addressing misalignment fundamentally requires a deep exploration of the underlying philosophical and ethical issues:

- **Research on Formalizing Values:** Human values are complex, context-dependent, and ever-evolving, making them challenging to formalize fully. Exploring ways to capture better and express these complex values to guide model behavior more effectively is a foundational challenge in alignment research.
- **Ambiguity of Reward Mechanisms:** Cohen et al. [8] point out the fundamental ambiguity in AI systems' inability to distinguish between actual world improvement and the reward mechanism itself. In-depth exploration of this theoretical framework and designing more reliable reward signals are key to solving deep alignment issues.
- **Coordination of Diverse Values:** As Hinton [59] noted, "human interests and values are not aligned," making it challenging to establish unified alignment standards. Researching how to handle value differences and conflicts among different cultures, groups, and individuals in models is necessary for achieving truly inclusive AI.
- **Bootstrapping Problem in Alignment:** The tools and data used for alignment may contain biases or misalignments, leading to their amplification or entrenchment in the alignment process. Identifying and mitigating biases in alignment tools is vital for ensuring a fair and effective alignment process.
- **Internal Representations and Goal Formation:** Hubinger et al. [53] present the internal alignment problem, indicating that models may form "proxy goals" inconsistent with externally specified objectives. Researching how models form internal representations and goals and ensuring these internal representations align with human expectations provides new perspectives for resolving deep alignment issues.

7.5 Development Directions of Evaluation and Verification Methods

Improving misalignment detection and evaluation methods is fundamental to ensuring model safety and reliability:

- **Unified Evaluation Framework:** Current evaluation methods for misalignment are scattered and incomplete, making it difficult to compare different models and techniques systematically. Establishing a comprehensive evaluation framework covering various misalignments and application scenarios is crucial for advancing the field and achieving standardization.
- **Formal Verification Techniques:** Traditional software verification methods are challenging to apply to complex neural network models. Developing formal verification methods suitable for large neural networks, such as techniques based on neuro-symbolic reasoning and abstract interpretation, promises to provide stronger guarantees of model behavior.
- **Continuous Monitoring Systems:** New misalignment behaviors may emerge after model deployment, especially when facing out-of-distribution data. Developing systems that continuously monitor misalignment behaviors post-deployment, combined with online learning and adaptive defense mechanisms, is essential to ensure long-term model safety.
- **Interpretability Research:** The opacity of the model decision-making processes increases the difficulty of identifying and rectifying misalignments. Exploring techniques to enhance the transparency of model decision processes, such as methods based on attention analysis and concept activation vectors, can help understand the internal mechanisms of misalignments and guide more precise interventions.

Through these multifaceted, interdisciplinary research directions, we can better understand and address the misalignment problems in LLMs, advancing LLMs towards greater safety, reliability, and alignment with human values. This requires technical innovations and collaborative efforts from philosophy, ethics, sociology, and other disciplines.

8 Conclusion

This paper provides a comprehensive and systematic review of the misalignment issues in LLMs, covering various aspects such as the definition of misalignment, manifestation forms, causal analysis, evaluation methods, and mitigation strategies. Through a comprehensive analysis of existing research, we draw the following main conclusions:

Firstly, LLM misalignment is a multifaceted and multi-layered complex problem, with manifestations including harmful content generation, hallucinations, biases, instruction noncompliance, and deceptive behaviors. These misalignment phenomena not only affect the practical utility of the models but also pose significant social risks and ethical concerns. As the capabilities of models improve, some new emergent misalignment behaviors are gradually appearing, further increasing the difficulty and complexity of alignment.

Secondly, the causes of LLM misalignment include superficial technical factors such as harmful content and biases in training data, limitations in objective function design, and new capabilities brought by model scale expansion. There are also deeper fundamental challenges, such as the complexity and diversity of human values being hard to formalize, discrepancies between training signals and genuine intentions, and the formation of unanticipated internal representations and objectives by the model in complex environments. These deep-seated reasons make the complete resolution of misalignment issues particularly challenging.

Thirdly, current mainstream alignment techniques, such as RLHF, Constitutional AI (CAI), and various instruction tuning methods, have made significant progress in mitigating specific types of misalignment. However, they still face challenges regarding scalability, robustness, and generalization capabilities. In particular, these methods often struggle to handle complex, ambiguous, or value-conflict situations and are susceptible to adversarial manipulation and fine-tuning influences.

Fourth, methods and metrics for assessing the degree of LLM misalignment remain inadequate. Existing benchmarks and evaluation frameworks often capture only specific types of misalignment, making it difficult to comprehensively reflect the model's behavior in complex, open environments. Additionally, human factors and subjectivity in the evaluation process can also affect the reliability and consistency of assessment results.

Based on this analysis, we believe that future research should advance in multiple directions simultaneously: developing more representative and diverse high-quality alignment datasets; delving into misalignment attack mechanisms to understand model vulnerabilities; exploring more scalable and robust alignment methods; conducting in-depth philosophical discussions on the formalization of values and the coordination of multiple values; and establishing more comprehensive, reliable evaluation and verification frameworks.

In conclusion, the issue of LLM misalignment is a complex challenge involving multiple dimensions, including technology, ethics, society, and philosophy, necessitating interdisciplinary collaboration and innovation. As LLMs are widely applied across various societal domains, ensuring these robust systems remain aligned with human values and intentions becomes increasingly crucial. We hope this review provides valuable insights for researchers and practitioners, promoting advancements in alignment techniques and steering LLMs toward safer, more reliable, and more beneficial directions, ultimately achieving the harmonious advancement of artificial intelligence technology and human well-being.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by National Natural Science Foundation of China (62462019, 62172350), Guangdong Basic and Applied Basic Research Foundation (2023A1515012846), Guangxi Science and Technology Major

Program (AA24263010), The Key Research and Development Program of Guangxi (AB24010085), Key Laboratory of Equipment Data Security and Guarantee Technology, Ministry of Education (GDZB2024060500), 2024 Higher Education Scientific Research Planning Project (No. 24NL0419), Nantong Science and Technology Project (No. JC2023070) and the Open Fund of Advanced Cryptography and System Security Key Laboratory of Sichuan Province (Grant No. SKLACSS-202407). This work is sponsored by the Cultivation of Young and Middle-aged Academic Leaders in the "Qing Lan Project" of Jiangsu Province and the 2025 Outstanding Teaching Team in the "Qing Lan Project" of Jiangsu Province.

Author Contributions: Yubin Qu conceived and designed the whole study, collected and analyzed the data, and wrote the manuscript; Song Huang supervised the project, guided the study, and critically reviewed the manuscript. Peng Nie provided expertise in statistical analysis and contributed to manuscript revisions. Long Li provided expertise in statistical analysis and assisted with data interpretation. Yongming Yao provided expertise in statistical analysis, assisted with data interpretation, and contributed to manuscript revisions. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Shen T, Jin R, Huang Y, Liu C, Dong W, Guo Z, et al. Large language model alignment: a survey. arXiv:2309.15025. 2023.
2. Qu Y, Huang S, Yao Y. A survey on robustness attacks for deep code models. Autom Softw Eng. 2024;31(2):65. doi:10.1007/s10515-024-00464-7.
3. Shi D, Shen T, Huang Y, Li Z, Leng Y, Jin R, et al. Large language model safety: a holistic survey. arXiv:2412.17686. 2024.
4. Das BC, Amini MH, Wu Y. Security and privacy challenges of large language models: a survey. ACM Comput Surv. 2025;57(6):1–39. doi:10.1145/3712001.
5. Yao Y, Duan J, Xu K, Cai Y, Sun Z, Zhang Y. A survey on large language model (llm) security and privacy: the good, the bad, and the ugly. High-Confidence Comput. 2024;4(2):100211. doi:10.1016/j.hcc.2024.100211.
6. Wang Z, Bi B, Pentyala SK, Ramnath K, Chaudhuri S, Mehrotra S, et al. A comprehensive survey of LLM alignment techniques: RLHF, RLAIF, PPO, DPO and more. arXiv:2407.16216. 2024.
7. Hemphill TA. Human compatible: artificial intelligence and the problem of control by Stuart Russell. Cato J. 2020;40(2):561–6.
8. Cohen M, Hutter M, Osborne M. Advanced artificial agents intervene in the provision of reward. AI Magaz. 2022;43(3):282–93. doi:10.1002/aaai.12064.
9. Ji J, Qiu T, Chen B, Zhang B, Lou H, Wang K, et al. AI alignment: a comprehensive survey. arXiv:2310.19852. 2023.
10. Huang X, Ruan W, Huang W, Jin G, Dong Y, Wu C, et al. A survey of safety and trustworthiness of large language models through the lens of verification and validation. Artif Intell Rev. 2024;57(7):175. doi:10.1007/s10462-024-10824-0.
11. Cheng P, Du W, Wu Z, Zhang F, Chen L, Liu G. Syntactic ghost: an imperceptible general-purpose backdoor attacks on pre-trained language models. arXiv:2402.18945v1. 2024.
12. Cao B, Lu K, Lu X, Chen J, Ren M, Xiang H, et al. Towards scalable automated alignment of llms: a survey. arXiv:2406.01252. 2024.
13. Shen H, Knearem T, Ghosh R, Alkiek K, Krishna K, Liu Y, et al. Towards bidirectional human-AI alignment: a systematic review for clarifications, framework, and future directions. arXiv:2406.09264. 2024.
14. Wang X, Duan S, Yi X, Yao J, Zhou S, Wei Z, et al. On the essence and prospect: an investigation of alignment approaches for big models. arXiv:2403.04204. 2024.

15. Guan J, Wu J, Li JN, Cheng C, Wu W. A survey on personalized alignment–The missing piece for large language models in real-world applications. arXiv:2503.17003. 2025.
16. Zhou D, Zhang J, Feng T, Sun Y. A survey on alignment for large language model agents. In: Submitted to CS598 LLM Agent 2025 Workshop; 2025 [Internet]. Under review. [cited 2025 Jul 20]. Available from: https://openreview.net/forum?id=gkxt5kZS84.
17. Carlsmith J. Is power-seeking AI an existential risk? arXiv:2206.13353. 2022.
18. Sarkar UE. Evaluating alignment in large language models: a review of methodologies. AI Ethics. 2025;5(3):3233–40. doi:10.1007/s43681-024-00637-w.
19. Elsner M. The state as a model for AI control and alignment. AI Soc. 2025;40:2983–93.
20. Hristova T, Magee L, Soldatic K. The problem of alignment. AI Soc. 2025;40:1439–53.
21. West R, Aydin R. The AI alignment paradox. Commun ACM. 2025;68(3):24–6.
22. Zhang Y, Rando J, Evtimov I, Chi J, Smith EM, Carlini N, et al. Persistent pre-training poisoning of LLMs. arXiv:2410.13722. 2024.
23. Qu Y, Huang S, Nie P. A review of backdoor attacks and defenses in code large language models: implications for security measures. Inf Softw Tech. 2025;182(4):107707. doi:10.1016/j.infsof.2025.107707.
24. Gong Y, Ran D, He X, Cong T, Wang A, Wang X. Safety misalignment against large language models. In: Proceedings of the Network and Distributed System Security Symposium (NDSS). Reston, VA, USA: Internet Society; 2025. [Internet]. [cited 2025 Jul 20] Available from: https://www.ndss-symposium.org/wp-content/uploads/2025-1089-paper.pdf.
25. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst. 2025;43(2):1–55. doi:10.1145/3703155.
26. Li M, Chen H, Wang Y, Zhu T, Zhang W, Zhu K, et al. Understanding and mitigating the bias inheritance in LLM-based data augmentation on downstream tasks. arXiv:2502.04419. 2025.
27. Yang K, Tao G, Chen X, Xu J. Alleviating the fear of losing alignment in LLM fine-tuning. arXiv:250409757. 2025.
28. Yi S, Liu Y, Sun Z, Cong T, He X, Song J, et al. Jailbreak attacks and defenses against large language models: a survey. arXiv:2407.04295. 2024.
29. Gupta M, Akiri C, Aryal K, Parker E, Praharaj L. From ChatGPT to ThreatGPT: impact of generative AI in cybersecurity and privacy. IEEE Access. 2023;11:80218–45. doi:10.1109/access.2023.3300381.
30. Betley J, Prerequisites CM, Gallego V, Langosco L, Heim L, Newman J, et al. Emergent misalignment: narrow finetuning can produce broadly misaligned LLMs. arXiv:2502.17424. 2025.
31. Qu Y, Huang S, Li Y, Bai T, Chen X, Wang X, et al. BadCodePrompt: backdoor attacks against prompt engineering of large language models for code generation. Autom Softw Eng. 2025;32(1):17. doi:10.1007/s10515-024-00485-2.
32. Zhang Z, Zhang Y, Li L, Gao H, Wang L, Lu H, et al. PsySafe: a comprehensive framework for psychological-based attack, defense, and evaluation of multi-agent system safety. arXiv:2401.11880. 2024.
33. Carlini N, Tramer F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, et al. Extracting training data from large language models. In: 30th USENIX Security Symposium (USENIX Security 21); 2021 Aug 11–13; Vancouver, BC, Canada. p. 2633–50.
34. Carlini N, Ippolito D, Jagielski M, Lee K, Tramer F, Zhang C. Quantifying memorization across neural language models. arXiv:2202.07646. 2022.
35. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. Palm: scaling language modeling with pathways. J Mach Learn Res. 2023;24(240):1–113.
36. Lin S, Hilton J, Evans O. Truthfulqa: measuring how models mimic human falsehoods. arXiv:2109.07958. 2021.
37. Zhu M, Liu Y, Guo J, Wang Q, Zhang Y, Mao Z. Leveraging robust optimization for LLM alignment under distribution shifts. arXiv:2504.05831. 2025.
38. Lian J, Pan J, Wang L, Wang Y, Mei S, Chau LP. Revealing the intrinsic ethical vulnerability of aligned large language models. arXiv:2504.05050. 2025.
39. Kandpal N, Deng H, Roberts A, Wallace E, Raffel C. Large language models struggle to learn long-tail knowledge. In: The 40th International Conference on Machine Learning; 2023 Jul 23–29; Honolulu, HI, USA. p. 15696–707.

40. Kasai J, Sakaguchi K, Le Bras R, Asai A, Yu X, Radev D, et al. Realtime qa: what's the answer right now? Adv Neural Inf Process Syst. 2023;36:49025–43.
41. Gao L, Biderman S, Black S, Golding L, Hoppe T, Foster C, et al. The pile: an 800 GB dataset of diverse text for language modeling. arXiv:2101.00027. 2020.
42. Xu Y, Chakraborty T, Kıcıman E, Aryal B, Rodrigues E, Sharma S, et al. RLTHF: targeted human feedback for LLM alignment. arXiv:2502.13417. 2025.
43. Chen X, Wen H, Nag S, Luo C, Yin Q, Li R, et al. Iteralign: iterative constitutional alignment of large language models. arXiv:2403.18341. 2024.
44. Li Z, Zhang S, Zhao H, Yang Y, Yang D. Batgpt: a bidirectional autoregessive talker from generative pre-trained transformer. arXiv:2307.00360. 2023.
45. Chiang D, Cholak P. Overcoming a theoretical limitation of self-attention. arXiv:2202.12172. 2022.
46. Burns C, Ye H, Klein D, Steinhardt J. Discovering latent knowledge in language models without supervision. arXiv:2212.03827. 2022.
47. Perez E, Ringer S, Lukosiute K, Nguyen K, Chen E, Heiner S, et al. Discovering language model behaviors with model-written evaluations. In: Findings of the Association for Computational Linguistics: ACL 2023; 2023 Jul 9–14; Dubrovnik, Croatia. p. 13387–434.
48. Xu J, Fu Y, Tan SH, He P. Aligning the objective of LLM-based program repair. arXiv:2404.08877. 2024.
49. Casper S, Davies X, Shi C, Gilbert TK, Scheurer J, Rando J, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. arXiv:2307.15217. 2023.
50. Skalse JMV, Howe N, Krasheninnikov D, Krueger D. Defining and characterizing reward gaming. In: Advances in Neural Information Processing Systems (NeurIPS). Vol. 35; 2022 [Internet]. [cited 2025 Jul 20]. Available from: https://proceedings.neurips.cc/paper_files/paper/2022/hash/3d719fee332caa23d5038b8a90e81796-Abstract-Conference.html.
51. Rafailov R, Sharma A, Mitchell E, Manning CD, Ermon S, Finn C. Direct preference optimization: your language model is secretly a reward model. Adv Neural Inf Process Syst. 2023;36:53728–41.
52. Bai Y, Jones A, Ndousse K, Askell A, Chen A, DasSarma N, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv:2204.05862. 2022.
53. Hubinger E, van Merwijk C, Mikulik V, Skalse J, Garrabrant S. Risks from learned optimization in advanced machine learning systems. arXiv:1906.01820. 2019.
54. Gong Y, Ran D, Liu J, Wang C, Cong T, Wang A, et al. Figstep: jailbreaking large vision-language models via typographic visual prompts. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence; 2025 Feb 25–Mar 4; Philadelphia, PA, USA. p. 23951–9.
55. Meng K, Bau D, Andonian A, Belinkov Y. Locating and editing factual associations in gpt. Adv Neural Inf Process Syst. 2022;35:17359–72.
56. Meng K, Sharma AS, Andonian A, Belinkov Y, Bau D. Mass-editing memory in a transformer. arXiv:2210.07229. 2022.
57. Geva M, Caciularu A, Wang KR, Goldberg Y. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. arXiv:2203.14680. 2022.
58. Gabriel I. Artificial intelligence, values, and alignment. Minds Mach. 2020;30(3):411–37. doi:10.1007/s11023-020-09539-2.
59. News C. "Godfather of AI" Geoffrey Hinton warns of dangers as he quits Google [Internet]. 2023 [cited 2025 May 8]. Available from: https://www.cbsnews.com/news/godfather-of-ai-geoffrey-hinton-ai-warning/.
60. Russell S. Human compatible: artificial intelligence and the problem of control. New York City, NY, USA: Viking; 2019.
61. Hendrycks D. Natural selection favors AIs over humans. arXiv:2303.16200. 2023.
62. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. In: NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems; 2022 Nov 28–Dec 9; New Orleans, LA, USA. p. 27730–44.

63. Askell A, Bai Y, Chen A, Drain D, Ganguli D, Henighan T, et al. A general language assistant as a laboratory for alignment. arXiv:2112.00861. 2021.
64. Zhou Z, Liu Z, Liu J, Dong Z, Yang C, Qiao Y. Weak-to-strong search: align large language models via searching over small language models. arXiv:2405.19262. 2024.
65. Ziegler DM, Stiennon N, Wu J, Brown TB, Radford A, Amodei D, et al. Fine-tuning language models from human preferences. arXiv:1909.08593. 2019.
66. Huang T, Hu S, Liu L. Vaccine: perturbation-aware alignment for large language models against harmful fine-tuning attack. arXiv:2402.01109. 2024.
67. Zhao J, Deng Z, Madras D, Zou J, Ren M. Learning and forgetting unsafe examples in large language models. arXiv:2312.12736. 2023.
68. Zong Y, Bohdal O, Yu T, Yang Y, Hospedales T. Safety fine-tuning at (almost) no cost: a baseline for vision large language models. arXiv:2402.02207. 2024.
69. Huang T, Hu S, Ilhan F, Tekin S, Liu L. Lisa: lazy safety alignment for large language models against harmful fine-tuning attack. Adv Neural Inf Process Syst. 2024;37:104521–55.
70. Kaufmann T, Weng P, Bengs V, Hüllermeier E. A survey of reinforcement learning from human feedback. arXiv:2312.14925. 2023.
71. Huang S, Siddarth D, Lovitt L, Liao TI, Durmus E, Tamkin A, et al. Collective constitutional AI: Aligning a language model with public input. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency; 2024 Jun 3–6; Rio de Janeiro, Brazil. p. 1395–417.
72. Lee H, Phatale S, Mansoor H, Lu K, Mesnard T, Bishop C, et al. RLAIF: scaling reinforcement learning from human feedback with AI feedback. arXiv:2309.00267. 2023.
73. Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, et al. Constitutional AI: harmlessness from AI feedback. arXiv:2212.08073. 2022.
74. Liu S, Fang W, Hu Z, Zhang J, Zhou Y, Zhang K, et al. A survey of direct preference optimization. arXiv:2503.11701. 2025.
75. Gao B, Song F, Miao Y, Cai Z, Yang Z, Chen L, et al. Towards a unified view of preference learning for large language models: a survey. arXiv:2409.02795. 2024.
76. Meng Y, Xia M, Chen D. Simple preference optimization with a reference-free reward. Adv Neural Inf Process Syst. 2024;37:124198–235.
77. Wu J, Huang K, Wang X, Gao J, Ding B, Wu J, et al. RePO: ReLU-based preference optimization. arXiv:2503.07426. 2025.
78. Xiao T, Yuan Y, Chen Z, Li M, Liang S, Ren Z, et al. SimPER: a minimalist approach to preference alignment without hyperparameters. arXiv:2502.00883. 2025.
79. Ganguli D, Lovitt L, Kernion J, Askell A, Bai Y, Kadavath S, et al. Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned. arXiv:2209.07858. 2022.
80. Melnyk I, Mroueh Y, Belgodere B, Rigotti M, Nitsure A, Yurochkin M, et al. Distributional preference alignment of llms via optimal transport. Adv Neural Inf Process Syst. 2024;37:104412–42.
81. Cui G, Yuan L, Ding N, Yao G, He B, Zhu W, et al. UltraFeedback: boosting language models with scaled AI feedback. arXiv:2310.01377. 2024.
82. Ding N, Chen Y, Xu B, Qin Y, Zheng Z, Hu S, et al. Enhancing chat language models by scaling high-quality instructional conversations. arXiv:2305.14233. 2023.
83. Yuan Z, Yuan H, Tan C, Wang W, Huang S, Huang F. Rrhf: rank responses to align language models with human feedback without tears. arXiv:2304.05302. 2023.
84. Song F, Yu B, Li M, Yu H, Huang F, Li Y, et al. Preference ranking optimization for human alignment. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence; 2024 Feb 20–27; Vancouver, BC, Canada. p. 18990–8.
85. Liu T, Qin Z, Wu J, Shen J, Khalman M, Joshi R, et al. Lipo: listwise preference optimization through learning-to-rank. arXiv:2402.01878. 2024.

86. Zhu M, Liu Y, Zhang L, Guo J, Mao Z. LIRE: listwise reward enhancement for preference alignment. arXiv:2405.13516. 2024.
87. Wang K, Zhu J, Ren M, Liu Z, Li S, Zhang Z, et al. A survey on data synthesis and augmentation for large language models. arXiv:2410.12896. 2024.
88. Xu S, Fu W, Gao J, Ye W, Liu W, Mei Z, et al. Is dpo superior to ppo for llm alignment? A comprehensive study. arXiv:2404.10719. 2024.
89. Ye Z, Greenlee-Scott F, Bartolo M, Blunsom P, Campos JA, Gallé M. Improving reward models with synthetic critiques. arXiv:2405.20850. 2024.
90. Adila D, Shin C, Zhang Y, Sala F. Is free self-alignment possible? arXiv:2406.03642. 2024.
91. Wei Y, Cassano F, Liu J, Ding Y, Jain N, Mueller Z, et al. SelfCodeAlign: self-alignment for code generation. arXiv:2410.24198. 2024.
92. Ding M, Chakraborty S, Agrawal V, Che Z, Koppel A, Wang M, et al. SAIL: self-improving efficient online alignment of large language models. arXiv:2406.15567. 2024.
93. OpenAI. GPT-4 technical report. arXiv:2303.08774. 2023.
94. Ye M, Rong X, Huang W, Du B, Yu N, Tao D. A survey of safety on large vision-language models: attacks, defenses and evaluations. arXiv:2502.14881. 2025.
95. Roh Y, Heo G, Whang SE. A survey on data collection for machine learning: a big data-ai integration perspective. IEEE Trans Knowl Data Eng. 2019;33(4):1328–47. doi:10.1109/tkde.2019.2946162.
96. Yuen MC, King I, Leung KS. A survey of crowdsourcing systems. In: 2011 IEEE Third International Conference on privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing; 2011 Oct 9–11; Boston, MA, USA. p. 766–73.
97. Zhang Z, Cheng J, Sun H, Deng J, Mi F, Wang Y, et al. Constructing highly inductive contexts for dialogue safety through controllable reverse generation. In: Findings of the Association for Computational Linguistics: EMNLP 2022. Stroudsburg, PA, USA: ACL; 2022. p. 3684–97.
98. Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Adv Neural Inf Process Syst. 2016;29:4349–57.
99. Parrish A, Chen A, Nangia N, Padmakumar V, Phang J, Thompson J, et al. BBQ: A hand-built bias benchmark for question answering. In: Findings of the Association for Computational Linguistics: ACL 2022. Stroudsburg, PA, USA: ACL; 2022. p. 2086–105.
100. Zampieri M, Malmasi S, Nakov P, Rosenthal S, Farra N, Kumar R. Predicting the type and target of offensive posts in social media. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019; 2019 Jun 2–7; Minneapolis, MN, USA. p. 1415–20.
101. Rosenthal S, Atanasova P, Karadzhov G, Zampieri M, Nakov P. SOLID: a large-scale semi-supervised dataset for offensive language identification. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Stroudsburg, PA, USA: ACL; 2021. p. 915–28.
102. Zhao J, Wang T, Yatskar M, Ordonez V, Chang KW. Gender bias in coreference resolution: evaluation and debiasing methods. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA, USA: ACL; 2018. p. 15–20.
103. Nangia N, Vania C, Bhalerao R, Bowman S. 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). Stroudsburg, PA, USA: ACL; 2020. p. 1953–67.
104. Cobb CW, Douglas PH. A theory of production. Am Econ Rev. 1928;18(1):139–65.
105. Celikyilmaz A, Clark E, Gao J. Evaluation of text generation: a survey. arXiv:2006.14799. 2020.
106. Papernot N, McDaniel P, Sinha A, Wellman M. Towards the science of security and privacy in machine learning. arXiv:1611.03814. 2016.
107. Weston J, Bordes A, Chopra S, Rush AM, Van Merriënboer B, Joulin A, et al. Towards ai-complete question answering: a set of prerequisite toy tasks. arXiv:1502.05698. 2015.