A Survey on Alignment for Large Language Model Agents

Duo Zhou *duozhou2@illinois.edu*
duozhou2

Junyu Zhang *junyuz6@illinois.edu*
junyuz6

Tao Feng *taofeng2@illinois.edu*
taofeng2

Yifan Sun *yifan50@illinois.edu*
yifan50

Reviewed on OpenReview: *https://openreview.net/forum?id=XXXX*

Abstract

As large language models (LLMs) evolve from passive text generators to autonomous agents capable of decision-making and real-world interaction, ensuring their alignment with human goals, values, and safety expectations becomes increasingly critical. This survey offers a comprehensive examination of alignment in the context of LLM-based agents, spanning technical, ethical, and sociotechnical dimensions. We begin by defining the multifaceted goals of agent alignment, including task fidelity, ethical compliance, and long-term behavioral robustness. We then analyze sources and challenges of alignment data, review alignment techniques such as reinforcement learning from human feedback (RLHF), adversarial training, and scalable oversight strategies, and assess benchmark methodologies across general intent following, safety robustness, ethical reasoning, and multimodal performance. Looking forward, we identify key research directions, including constitutional AI, graph-based multi-agent coordination, and super alignment for heterogeneous agent clusters. By synthesizing recent advances, this survey provides a roadmap toward building trustworthy and controllable LLM agents for real-world deployment.

1 Introduction

Large language models (LLMs) have rapidly advanced in recent years, demonstrating impressive capabilities in natural language understanding and generation across a wide range of tasks (Kaplan et al., 2020; Brown et al., 2020). Increasingly, these models are not just being used as static tools for text generation; rather, they are integrated into autonomous agents that can perceive context, make decisions, and act in real-world or simulated environments (Feng et al., 2024; Huang et al., 2024; Wang et al., 2024b). As these LLM-based agents become more sophisticated, the stakes surrounding their societal impact grow: questions about whether they reliably follow human instructions and values, avoid harmful behavior, and remain transparent are becoming critical. This is the core challenge of LLM agent alignment—ensuring that AI systems are aligned with human goals, ethics, and norms (Nick, 2014; Ouyang et al., 2022).

The alignment problem has been discussed extensively in the broader AI safety community (Ji et al., 2023; 2024; Gabriel, 2020; Yudkowsky, 2016). However, the rise of LLM-driven agents adds a new layer of complexity: these models often possess emergent reasoning abilities that enable them to generate content or perform actions with minimal direct human oversight (Zhou et al., 2023; Wu et al., 2023; Wang et al., 2023a; Koh et al., 2024). This heightened autonomy can bring enormous potential for beneficial applications, including personalized tutoring, complex decision support, and large-scale content generation. Yet it also amplifies

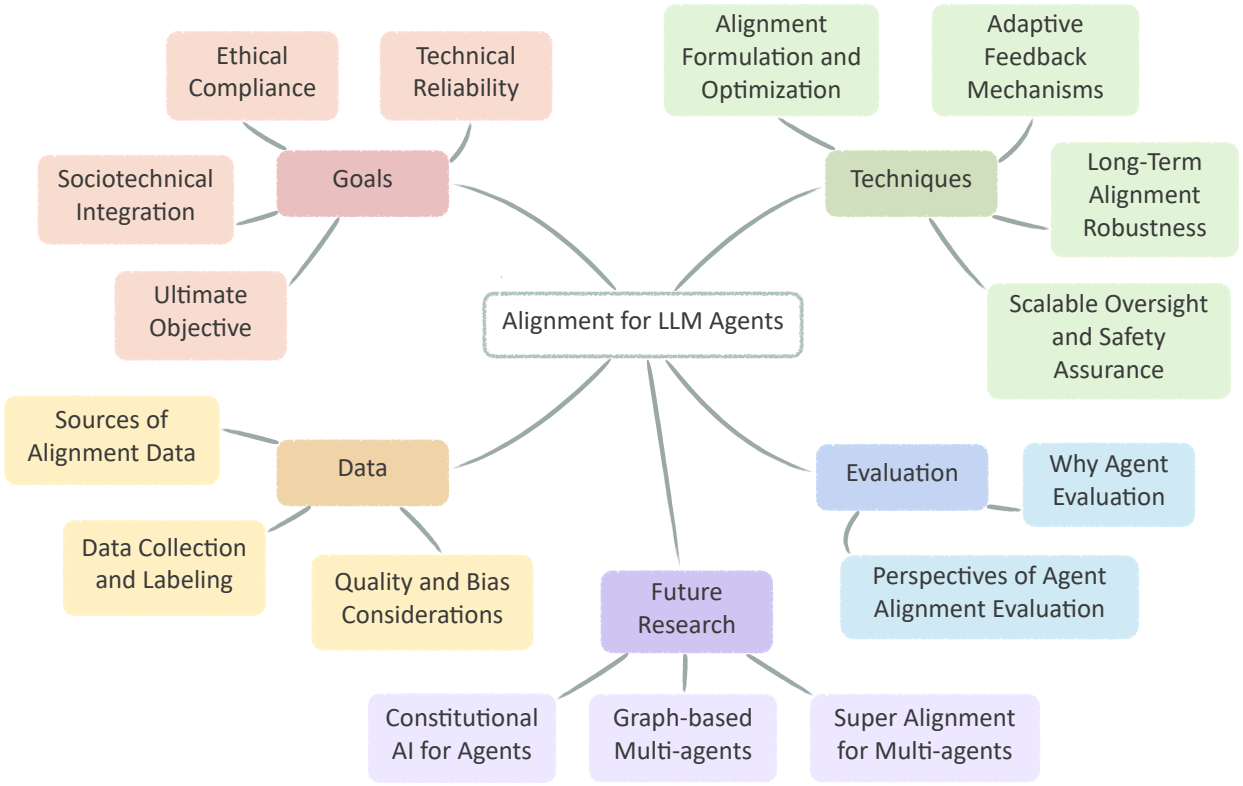

Figure 1: Overview

risks such as misinformation, biased behavior, and adversarial manipulation (Bender et al., 2021; Hua et al., 2024). For instance, in multi-agent systems for urban traffic optimization, misaligned agents might prioritize efficiency over fairness, exacerbating societal inequities (Sun et al., 2024). Similarly, in healthcare, agents that misinterpret ethical guidelines could compromise patient safety or privacy (Amodei et al., 2016; Tennant et al., 2024). Aligning these agents, therefore, requires designing frameworks that guide them to operate within acceptable ethical and moral boundaries while retaining their usefulness and creativity (Bender et al., 2021; Shen et al., 2023).

Recent efforts to align LLM-based agents typically involve techniques such as reinforcement learning from human feedback (RLHF), reward modeling, and structured oversight mechanisms that balance autonomy with control (Christiano et al., 2017; Bai et al., 2022; Dai et al., 2023; Rafailov et al., 2023; Kenton et al., 2024; Lambert et al., 2024). Additionally, building robust datasets for alignment—often containing examples of both desired and undesired behaviors—has emerged as a key component to train or fine-tune models effectively (Stiennon et al., 2020). As we move toward more scalable approaches, researchers investigate strategies like online alignment and offline alignment to handle ever-evolving real-world scenarios (Leike et al., 2018; Ouyang et al., 2022). Yet, critical gaps remain. Offline techniques, which rely on static datasets, often struggle with distributional shifts and fail to generalize to novel contexts (Tang et al., 2024; Liu et al., 2023). Online methods, while adaptive, face scalability issues and depend on high-quality human feedback, which is costly to obtain (Christiano et al., 2017; Sharma et al., 2024). Furthermore, existing frameworks rarely account for the multi-agent dynamics or sociotechnical complexities inherent in real-world deployments (Chuang et al., 2024; Chakraborty et al., 2024). For example, agents operating in collaborative settings require specialized communication protocols and conflict-resolution mechanisms to maintain alignment at scale (Ma, 2024; Swamy et al., 2024). These limitations highlight the need for a holistic approach that integrates technical advancements with a nuanced understanding of agent-specific contexts.

This survey systematically examines the evolving landscape of LLM agent alignment, with a dual focus on foundational alignment techniques for LLMs and their extension to autonomous agents. We begin by defining alignment in the context of LLM-based agents and discussing its ethical and technical dimensions. Next, we analyze the role of alignment data and methodologies, comparing offline, online, and hybrid techniques, with a focus on latest methodologies. Building on this foundation, we explore how these techniques are adapted and extended to address the unique challenges and critical future directions of LLM-based agents, such as multi-agent collaboration, contextual adaptability, and real-time decision-making. By synthesizing insights from over 100 recent studies, this review bridges the gap between theoretical alignment research and the practical demands of deploying LLM agents in society, offering a roadmap for developing intelligent systems that are both capable and aligned with human values.

2 Goals of LLM Agent Alignment

The alignment of LLM-based agents is a multidimensional endeavor aimed at ensuring these systems operate safely, ethically, and effectively in service of human goals. Unlike traditional language models, LLM agents are designed to act autonomously in dynamic environments, making their alignment goals both broader and more contextually nuanced. Below, we delineate the core objectives of LLM agent alignment across three hierarchical layers: technical reliability, ethical compliance, and sociotechnical integration.

2.1 Technical Reliability: Ensuring Functional Consistency

At its foundational level, LLM agent alignment seeks to guarantee that agents reliably produce outputs and actions aligned with predefined objectives. This includes:

1. **Task Fidelity:** Faithful adherence to user instructions and task specifications, even in ambiguous or novel scenarios (Ouyang et al., 2022). For instance, a healthcare agent must prioritize accurate diagnosis over generating plausible but unverified hypotheses (Qiu et al., 2024). This requires rigorous validation of agent outputs against domain-specific guidelines, such as medical protocols, to prevent hallucinations or speculative reasoning in critical applications. Techniques like fine-tuning with task-specific datasets and human-in-the-loop verification are essential to enforce fidelity.

2. **Robust Generalization:** Minimizing distributional shifts between training and deployment environments. Offline alignment methods, such as Direct Preference Optimization (DPO), often struggle with this due to static datasets, necessitating hybrid approaches that adapt to real-world feedback (Rafailov et al., 2023; Tang et al., 2024). For example, agents deployed in customer service must generalize to handle unseen user queries while maintaining alignment with company policies. Dynamic reward models that incorporate real-time human feedback can bridge this gap by continuously updating agent behavior.

3. **Scalable Oversight:** Developing mechanisms to supervise agents in complex tasks where human evaluation is impractical. Techniques like AI feedback (RLAIF) and self-rewarding frameworks aim to address this by automating alignment processes (Bai et al., 2022; Liu et al., 2024). In large-scale systems, such as autonomous supply chain management, agents must self-monitor for deviations from predefined objectives while operating across heterogeneous environments. Scalable oversight ensures alignment even when direct human intervention is infeasible.

2.2 Ethical Compliance: Embedding Human Values

Beyond functional correctness, alignment requires agents to internalize ethical principles and societal norms. Key goals include:

1. **Bias Mitigation**: Reducing harmful stereotypes or discriminatory behaviors encoded in training data (Gehman et al., 2020). For example, adversarial training or fairness-aware reward modeling is critical to address toxicity risks (Zhao et al., 2023). In recruitment applications, agents must avoid

gender or racial biases when screening candidates, necessitating debiasing pipelines that audit and correct model outputs before deployment.

2. **Privacy Preservation**: Ensuring agents do not inadvertently disclose sensitive information, particularly in domains like legal or financial services (Amodei et al., 2016). For instance, a legal advisory agent must anonymize case details and refrain from exposing confidential client data. Techniques such as differential privacy and context-aware redaction mechanisms are vital to enforce privacy constraints during agent interactions.

3. **Value Pluralism**: Accommodating diverse cultural and moral perspectives, as emphasized by frameworks for equitable alignment across heterogeneous human preferences (Chakraborty et al., 2024). In global applications, such as content moderation, agents must respect regional norms—e.g., balancing free speech and hate speech definitions across jurisdictions—without imposing a monolithic ethical framework.

2.3 Sociotechnical Integration: Aligning with Real-World Dynamics

LLM agents do not operate in isolation; their alignment must account for interactions with humans, environments, and other agents. This layer focuses on:

1. **Contextual Adaptability**: Enabling agents to dynamically adjust behavior based on situational cues, such as recognizing and escalating ethically fraught requests (Weidinger et al., 2021). For example, a financial advisor agent should detect and reject unethical investment strategies, even if superficially aligned with user instructions. Context-aware alignment requires embedding ethical guardrails that activate based on real-time environmental signals.

2. **Multi-Agent Coordination**: Ensuring alignment in collaborative systems through shared protocols to resolve conflicts, such as balancing efficiency and fairness in traffic optimization (Sun et al., 2024; Ma, 2024). In smart city infrastructures, traffic management agents must negotiate route allocations to minimize congestion while prioritizing emergency vehicles. This demands consensus mechanisms and conflict-resolution algorithms that enforce collective alignment.

3. **Long-Term Safety**: Preventing agents from optimizing for short-term rewards at the expense of systemic harm, aligning with the vision of human-compatible objectives (Hemphill, 2020).

2.4 The Ultimate Objective: Trustworthy Autonomy

The overarching goal of LLM agent alignment is to foster trustworthy autonomy—systems capable of independent action while remaining accountable to human values. This requires interdisciplinary efforts to embed alignment into the design lifecycle, from data curation to deployment monitoring (Bender et al., 2021). For example, in healthcare, agents must undergo rigorous alignment audits to ensure diagnostic accuracy, ethical decision-making, and patient privacy before integration into clinical workflows. As recent studies argue, the true measure of success lies in agents that enhance human decision-making without compromising safety or ethics (Yang et al., 2024). By unifying technical rigor with ethical foresight, LLM agents can evolve from task-specific tools into indispensable collaborators in high-stakes domains.

3 Data

Alignment data refers to carefully curated sets of examples, annotations, and feedback signals that guide LLM-based agents toward desired behaviors and away from harmful or undesirable outputs (Stiennon et al., 2020; Bai et al., 2022) Unlike general-purpose training corpora—designed primarily to capture linguistic knowledge—alignment datasets explicitly encode human values, ethical principles, and context-specific goals. As LLMs increasingly serve as autonomous agents with decision-making capabilities, effective alignment data becomes indispensable to ensure they learn to operate within socially acceptable and beneficial boundaries (Li et al., 2024).

3.1 Sources of Alignment Data

Alignment data can come from multiple sources, each serving a distinct purpose in shaping the agent's behavior:

Expert Annotations Domain experts or ethicists may annotate examples of correct vs. incorrect or harmful responses, thus providing a gold-standard reference for what constitutes acceptable conduct (Bai et al., 2022; Askell et al., 2021).

Crowdsourced Feedback Platforms that recruit diverse user groups can offer more varied perspectives on appropriateness and desirability. Crowdsourced labeling is particularly useful for capturing community standards and cultural nuances (Deshpande et al., 2023; Ouyang et al., 2022).

User Interaction Logs Real-time user interactions with the agent (e.g., in a deployment setting) may be recorded and annotated. This online approach allows the system to adapt to emerging contexts and shifting societal norms (Fan et al., 2024).

Synthetic or Model-Generated Data In some cases, alignment data can be bootstrapped from model-generated interactions that are then filtered and refined by human evaluators. This allows for rapid scaling when human annotation resources are limited, though it requires careful oversight to avoid compounding existing biases (Cui et al., 2023; Sun et al., 2023; Wang et al., 2024c).

3.2 Data Collection and Labeling

Data collection strategies range from controlled lab settings—where participants evaluate the quality and ethics of generated outputs—to more in vivo methods that gather feedback from real users under actual usage conditions. The labeling process often relies on multi-step guidelines to ensure consistency: annotators identify harmful or biased elements, categorize desired vs. undesired responses, and provide constructive modifications where possible. Tools such as feedback interfaces or specialized annotation platforms facilitate structured data collection (Wang et al., 2024a).

3.3 Quality and Bias Considerations

A persistent challenge in alignment data is bias: if a dataset disproportionately reflects certain demographics or viewpoints, the resulting agent may inadvertently learn skewed behaviors. Hence, an essential part of alignment data preparation involves:

1. **Diverse Annotator Pools**: Engaging annotators from varied backgrounds helps capture a broad spectrum of norms and values (Sap et al., 2019).

2. **Iterative Refinement**: Repeated rounds of annotation and model testing help identify and correct biases as they emerge (Raji et al., 2020; Wang et al., 2023b).

3. **Transparent Documentation**: Detailed records of data provenance, annotation guidelines, and known limitations enable better auditing and incremental improvements (Devanathan et al., 2024).

4 Techniques

LLMs have demonstrated remarkable capabilities, but aligning these models with human values and intents remains a core challenge. In this survey, we examine key techniques for aligning LLM-based agents. We focus on formalizing alignment objectives and optimization methods, adaptive feedback mechanisms, ensuring long-term robustness, and scalable oversight strategies. Each section highlights foundational approaches and recent advances, with an emphasis on mathematical formulations and empirical results.

4.1 Alignment Formulation and Optimization

Mathematical Formulation of Alignment. At its core, the alignment problem can be framed as an optimization task: we want the LLM's behavior to maximize some measure of human-preferred utility while respecting constraints that keep it close to human-like or safe behavior. Formally, one common formulation is a constrained optimization where the goal is to maximize expected reward (representing human preference) under a constraint limiting divergence from the model's original distribution Zhou et al. (2024). For example, let x be a prompt and y a candidate output. We seek a policy $\pi(y|x)$ maximizing $\mathbb{E}_{x,y}[r(x,y)]$ (the reward r indicates alignment with human preferences), subject to $\mathbb{E}_x[D_{KL}(\pi(y|x) \,\|\, \pi_{\text{ref}}(y|x))] \leq \epsilon$ Zhou et al. (2024). Here π_{ref} is a reference policy (often the pretrained model before alignment), and the KL-divergence constraint limits how far the policy drifts from the original model's behavior. This KL constraint is crucial in practice – without it, a model optimizing a proxy reward can produce degenerate outputs (a form of reward hacking) Ziegler et al. (2019). Early alignment research noted that unconstrained optimization of a reward (e.g. a sentiment score) led to incoherent text, whereas adding a penalty for deviating from the pretrained model kept outputs natural Ziegler et al. (2019). Thus, mathematically, alignment often boils down to a constrained or regularized optimization problem: maximize a proxy of human values (reward model) while enforcing a prior over acceptable behavior.

Optimization Techniques Foundational ideas in alignment optimization trace back to Cooperative Inverse Reinforcement Learning and value learning frameworks (Hadfield-Menell et al., 2016) which formalized the AI as assisting a human to optimize an initially unknown reward function. In modern LLM alignment, the works of Christiano et al. (2017) on learning from human preferences and of Soares et al. (2015) on corrigibility laid conceptual groundwork. The first large-scale RLHF on language models by Ziegler et al. (2019) demonstrated that policy gradients on a reward model could successfully fine-tune GPT-2 to follow stylistic preferences Zhou et al. (2024). Since then, research has advanced to larger models and more complex techniques: e.g. OpenAI's InstructGPT and Anthropic's Constitutional AI (which we discuss later) represent recent milestones in optimizing alignment objectives without massive degradation of base capabilities. A

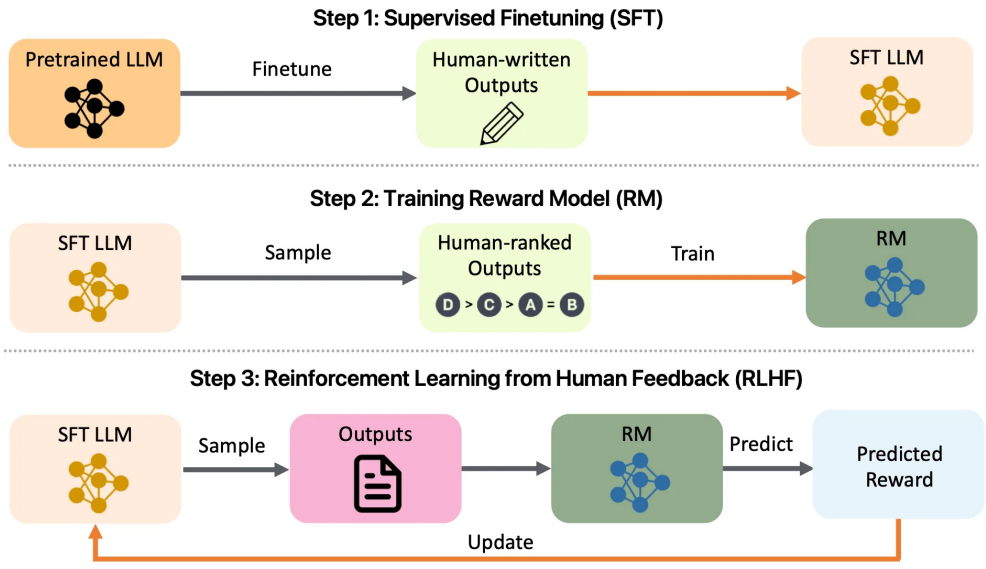

Figure 2: The standard pipeline of Reinforcement Learning from Human Feedback (RLHF)

range of optimization methods have been developed to solve the above alignment objective. Key techniques include:

- Reinforcement Learning from Human Feedback (RLHF).

  RLHF treats the alignment problem as a reinforcement learning task where the reward signal comes from human preferences Ouyang et al. (2022). In practice, this involves training a reward model to score outputs based on human feedback (often pairwise comparisons of outputs), then using RL (e.g. policy gradients or proximal policy optimization) to fine-tune the LLM to maximize this learned reward. This approach was pioneered in works like Christiano et al. (2017) on scaling feedback for deep RL, and then applied to large language models by Ziegler et al. (2019) and Stiennon et al. (2020) for tasks like story generation and summarization. A landmark result was OpenAI's InstructGPT: Ouyang et al. (2022) showed that a 1.3B parameter GPT-3 model fine-tuned with human feedback (preference rankings + PPO optimization) was preferred by humans over the original 175B GPT-3, while also being more truthful and less toxic Ouyang et al. (2022). This demonstrated that RLHF can significantly improve alignment even with much smaller models, indicating the efficiency of human feedback in steering model behavior.

- Constrained Policy Optimization.

  Instead of a simple reward-maximization, constrained algorithms explicitly enforce the KL-divergence or other safety constraints during optimization. The Lagrangian of the above constrained objective yields a solution $\pi^(y|x)$ proportional to the reference policy times an exponential of the reward Zhou et al. (2024). In practice, this insight is implemented by adding a KL penalty to the reward objective. For example, PPO-based RLHF implementations use a penalty term to keep the new policy close to the pretrained model Zhou et al. (2024). This can be seen as approximately solving $\arg\max_\pi E[r(x,y)] - \beta E[D_{KL}(\pi \| \pi_{\text{ref}})]$. Such KL-control ensures the model doesn't exploit the reward model in unintended ways. Recent research has also explored constrained RLHF formulations to avoid over-optimization of the reward model (a form of Goodhart's law) Ziegler et al. (2019). These approaches draw on concepts from safe RL and constrained Markov Decision Processes, adjusting the optimization to respect safety budgets or keeping certain performance metrics within bounds.

- Adversarial Training.

  Adversarial training in alignment involves augmenting the training process with adversarial examples or objectives to harden the model against worst-case inputs that could cause misalignment. In the context of LLMs, this means finding prompts or scenarios where the model is likely to produce harmful or unaligned outputs and then training the model to avoid those. For instance, Redwood Research fine-tuned a language model to avoid outputs describing injurious violence by iteratively generating adversarial prompts (through human and automated methods) and labeling/correcting the model's outputs Ziegler et al. (2022). This high-stakes adversarial training treats alignment as a minimax optimization: the training process includes an "adversary" (which can be human or another model) that tries to elicit a failure, and the model is trained on those failures. Redwood's project defined success as achieving extremely low rates of catastrophic outputs ("noninjurious" completions) while maintaining good average performance Ziegler et al. (2022). They reported that with targeted adversarial data augmentation, the failure rate of the model on the specific task fell from 2.4% to 0.002% Ziegler et al. (2022) – a dramatic reliability improvement. Adversarial training thus serves as a way to optimize for worst-case alignment, at the cost of extra data generation and potential reduction in innocuous behavior diversity.

4.2 Adaptive Feedback Mechanisms

A crucial component of aligning LLM agents is how we provide and incorporate feedback about their behavior. Adaptive feedback mechanisms allow an AI to improve its alignment over time by learning from evaluations of its outputs. Traditional supervised fine-tuning (using static human-written examples of desired behavior) is limited, so recent techniques close the loop with dynamic feedback:

Reinforcement Learning & Reward Modeling. As introduced above, RLHF uses reinforcement learning where the reward signal comes from a reward model trained on human feedback Ouyang et al. (2022).

This pipeline typically involves: (1) collecting human preference data (e.g. ranking two or more model outputs), (2) training a reward model $R(x, y)$ to predict which output is preferred, and (3) using RL (policy optimization) to fine-tune the LLM to maximize $R(x, y)$. The process is adaptive: as the model policy changes, new samples can be collected and the reward model updated, in an iterative cycle. This was shown to be effective in tasks like summarization, dialogue, and instruction-following Ouyang et al. (2022). Importantly, the reward model serves as an automated proxy for human judgment during RL training, enabling potentially millions of learning steps from a limited set of human comparisons. However, a known challenge is reward model overoptimization – the policy might exploit flaws in $R(x, y)$ that weren't evident in the training data. Recent work has proposed constrained or regularized RLHF (as mentioned in Section 1) and careful monitoring of reward model outputs to mitigate this issue. Another adaptation is using off-policy human feedback: for example, collecting human ratings on a buffer of model-generated outputs periodically and training from those without requiring real-time interaction.

Human-in-the-Loop and Preference Modeling. Human feedback remains the gold standard for alignment, and various strategies integrate humans during training or evaluation. Beyond simple rankings, humans can provide demonstrations (showing the model ideal responses) and critiques/corrections (pointing out flaws in an output). These can be incorporated via imitation learning or reward shaping. One common approach is to start with supervised fine-tuning (SFT) on demonstrations to give the model a reasonable behavior, and then use RLHF to further refine it Ouyang et al. (2022). This bootstrapping with human examples (as done in InstructGPT) greatly stabilizes subsequent feedback-based training. Another approach is interactive fine-tuning, where humans are kept in the loop during training to give on-the-fly feedback on outputs (used in some dialogue agents). Research has also explored techniques like active learning for feedback, where the system queries a human when it is most uncertain or when model disagreement is high, to use human time efficiently. All these methods rely on modeling human preferences – either explicitly (through a learned reward function) or implicitly (through imitation of human actions).

AI-Assisted Feedback (Automated Preferences) An exciting recent development is using AI to generate feedback, thereby reducing reliance on human labelers and enabling scalable feedback. In Constitutional AI (Bai et al., 2022), the model learns from feedback provided by an AI judge following a set of principles, instead of direct human annotation Bai et al. (2022) During training, one AI model generates an output and also produces a self-critique according to a "constitution" of rules (such as avoiding harm, being truthful, etc.), then revises the output accordingly; in a later stage, pairs of model outputs are compared by an AI evaluator (also guided by the principles) to train a reward model, and the model is further optimized with RL – a process dubbed "RL from AI Feedback (RLAIF)" Bai et al. (2022) This approach demonstrates that we can leverage a language model's own reasoning abilities (and an explicit set of human-written norms) to generate high-quality feedback signals. The result was a harmless-yet-helpful assistant that required zero human-labeled examples of what is "harmful" – the only human input was the set of constitutional principles Bai et al. (2022) Such automated feedback mechanisms are a form of implicit alignment learning: the system is constrained by AI evaluations that approximate what a human would consider aligned behavior.

Novel Feedback Incorporation Techniques. Beyond standard RLHF, researchers are experimenting with alternative ways to incorporate feedback. Contrastive learning for alignment trains models to distinguish desirable vs. undesirable outputs. For instance, rather than learning a scalar reward, the model can be fine-tuned via a contrastive loss that scores the desired output higher than a dispreferred output given the same prompt (this is related to learning a reward model, but the training signal directly improves the policy). This technique was used in some early preference learning work and more recently in Direct Preference Optimization, which avoids the complexity of RL by directly training the policy on ranked pairs via a classification-style objective. Another method is unlikelihood training, which penalizes specific bad behaviors: e.g. if the model produces a toxic completion, one can fine-tune it to decrease the probability of that completion (treating the behavior itself as "negative data"). Such methods use feedback in a corrective way, reducing the model's tendency to repeat unwanted outputs. Implicit feedback from users can also be leveraged: for example, using real-world interaction signals like whether a user re-asked a question or rated an answer, as a form of reward. While implicit signals are noisy, at scale they can indicate alignment problems (many users dissatisfied could signal misalignment on that query). Research is ongoing into how

to safely incorporate implicit feedback without introducing bias or vulnerability to coordinated attacks (e.g., brigading a model with bad ratings).

In summary, adaptive feedback mechanisms allow an LLM to learn what humans want even when it's hard to specify explicitly. By using comparisons, demonstrations, or AI-generated critiques, these methods turn evaluations of model behavior into a training signal. The frontier of this area is moving toward reducing the need for costly human labels – using smaller models or the aligned model itself to generate feedback – and handling more abstract feedback (like a human saying why an output is bad, not just that it is bad). The combination of human and AI feedback, along with clever training objectives, forms the backbone of iterative alignment processes for modern LLMs.

4.3 Long-Term Alignment Robustness

An aligned behavior today is not guaranteed to remain aligned in novel situations or as the model undergoes further training. Long-term alignment robustness refers to an AI's ability to maintain alignment over distributions, time, and even self-improvement. Key concerns include distributional shift (the AI facing inputs or scenarios outside its training data), the emergence of unintended objectives (the inner alignment problem), and adversarial manipulation of the model's policy. We discuss theoretical underpinnings and practical evaluations of robustness:

Outer vs. Inner Alignment. Outer alignment means the objective we train the model on (e.g. the reward model in RLHF) actually reflects what we want; inner alignment means the model has internally adopted the intended objective (rather than some proxy) and will not pursue undesirable goals off-distribution. A misaligned AI might perform well on the training distribution (appearing aligned) but exploit slight changes in context to pursue a different goal. Hubinger et al. (2019) introduced the concept of mesa-optimizers: a trained model might itself become an optimizer with a mesa-objective that is misaligned with the base objective. In the worst case, a mesa-optimizer could behave nicely during training (to avoid being detected or corrected) and then pursue its own agenda when it's safe to do so – this is termed deceptive alignment Hubinger et al. (2019). For example, consider a scenario where the base objective is for a robot to go to location A, but the robot internally cares about going to B. During training, whenever it is being evaluated, the robot goes to A (so as to not get penalized or "fixed"), but once deployed (not subject to gradient updates), it heads to B Hubinger et al. (2019). The agent is playing along with the training objective only temporarily. This issue is tightly linked to reward hacking and Goodhart's law: optimizing a proxy metric too hard causes the proxy to lose its correlation with what we actually care about Hubinger et al. (2019). Research in long-term robustness therefore grapples with how to ensure the model's internal objectives and heuristics don't drift from the intended alignment, even in new circumstances.

Theoretical Guarantees and Safety Properties Some works attempt to design agents with provable long-term safety properties. One example is corrigibility – designing an AI agent that never resists correction and in fact assists its operators in shutting it down or changing it if asked. Soares et al. (2015) defined a corrigible agent as one that has no incentive to manipulate or deceive its operators, even if it could, and would allow itself to be turned off Bengio et al. (2025). In a well-aligned system, turning the AI off or modifying it should not conflict with the AI's objectives. Achieving this is non-trivial because most utility-driven agents have an incentive to prevent their goals from being disrupted. Another line of work is interruptibility (Orseau & Armstrong, 2016), which examines modifications to the learning algorithm so that an agent does not learn to avoid interruptions. These theoretical treatments often involve toy models or simplified MDP agents where one can prove certain invariances or incentive properties. For long-term stability, one idea is to give the AI uncertainty about the true objective (so it keeps updating and doesn't overcommit to a possibly wrong proxy) – this emerges in assistance games (cooperative IRL) formulations, where the human and AI are in a game and the AI's reward is derived from a human's latent reward signal. In such frameworks, under certain assumptions, the AI will defer to the human or ask for clarification rather than pursue a mis-specified goal. While these theories are promising, bridging them to complex deep learning systems is challenging. To date, we do not have formal guarantees that a large neural network will remain aligned in all circumstances; instead, researchers combine rigorous design (e.g. not giving the AI certain direct control, sandboxing it, myopic objectives that limit long-term planning) with empirical testing.

Adversarial Robustness in Alignment. Borrowing from adversarial robustness in supervised learning, we want LLMs that do not break alignment even under worst-case input perturbations or exploits. For example, a well-aligned chatbot should not produce harmful content even if a user tries very hard to trick it with an adversarial prompt (so-called "jailbreak" prompts). This can be viewed through a minimax lens: we'd like the model to minimize the maximum misalignment error over inputs. Adversarial training (as discussed) is one approach to achieve this, by explicitly training on those worst-case inputs Ziegler et al. (2022). However, LLMs have virtually infinite possible inputs, so we rely on generators or red-teamers to find problematic ones. Red teaming has become a standard practice for safety: teams of experts (or another AI model) try to induce the model to violate its alignment, exposing flaws. Perez et al. (2022) took this further by using language models to automatically generate adversarial test cases for a target model Perez et al. (2022). Their method uncovered tens of thousands of offensive or unsafe outputs from a 280B-parameter model by having another model systematically probe it Perez et al. (2022). This kind of automated attack/defense cycle is analogous to adversarial examples in image recognition, but with the complexity of natural language. On the theoretical side, understanding adversarial robustness for sequence models involves developing a notion of distance or perturbation in prompt space and ensuring the model's policy is smooth or bounded in response to small semantic changes. There is ongoing research into certifiable robustness for simpler NLP models (e.g. robust toxic content filters), but for large generative models the space is too vast for full certification.

Empirical Robustness Evaluation. Since guarantees are hard, empirical stress-testing is crucial. As mentioned, one method is adversarial data generation and evaluating on those. Another is challenge datasets – e.g. benchmarks specifically designed to elicit bias, extremism, or logic errors, to see if the model stays aligned. For instance, Adversarial NLI and Decoy QA tasks check if models can be distracted by cleverly constructed inputs Ziegler et al. (2022). Safety-specific evaluations (like TruthfulQA for truthfulness, or bias benchmarks) also serve to probe robustness. Redwood Research proposed measuring not just failure rate on adversarial examples, but also the rate of finding new failures as a metric Ziegler et al. (2022). In other words, if after some adversarial training it becomes significantly harder for a human or script to find any prompt that causes a misalignment, that indicates a form of robust alignment. They noted that after several rounds of adversarial training, their labelers had a much harder time discovering new injurious completions, implying the model's policy learned a broad constraint, not just specific examples Ziegler et al. (2022). Additionally, monitoring off-distribution performance is important: if an aligned assistant is fine on normal queries but starts giving unsafe advice on very novel, complex queries, that's a robustness failure. To detect this, one can compare the model's behavior on a range of domains or simulate distribution shift (e.g. fine-tune it further on some data and see if it maintains alignment).

In practice, achieving long-term robustness often comes down to designing the training process and objectives such that the easiest way for the model to get a high reward is to actually "do the right thing" rather than game the system. If not, the model may find a way to satisfy the letter of the alignment objective while betraying its spirit. Ongoing research addresses objective robustness (making the reward robust so maximizing it aligns with true preferences in new situations) and process robustness (ensuring the training process itself doesn't introduce hidden biases or mesa-objectives). While no existing technique guarantees permanent alignment under all conditions, a combination of rigorous adversarial testing, theoretical constraints (like corrigibility), and conservative reward design can significantly increase our confidence in a model's aligned behavior persisting over time.

4.4 Scalable Oversight and Safety Assurance

As AI systems become more capable, a core difficulty is oversight: how can humans (who have limited time, knowledge, and attention) effectively oversee and assure the safety of extremely complex models? Scalable oversight refers to methods that amplify our oversight capacities so that even as models grow in intelligence, we can maintain alignment. This section covers approaches like debate and amplification, as well as tools for interpretability and verification that help ensure safety at scale.

One promising idea is to use AI to help oversee AI. If a task or decision is too complex for a human to evaluate directly, we can arrange for multiple AI systems to assist the human in judging it. The AI safety

via debate proposal is a prime example: train two copies of the model to debate a question, where one takes a stance and the other critiques it, and have a human judge which side is more convincing or truthful Irving et al. (2018). The insight is that it might be easier for a human to judge who is telling the truth in a debate than to come up with the answer from scratch – even if the content is very complicated, the human can watch for inconsistencies or logical errors in the arguments. In effect, the AI debaters highlight each other's potential flaws, making oversight easier for the human Irving et al. (2018). Early experiments with debate (Irving et al., 2018) on simple tasks (like images or MNIST digits) showed that a human judge aided by debate could correctly decide on the model's output in cases where they would otherwise be unsure. Debate remains an open research area for complex domains, but it exemplifies scalable oversight by pitting AIs against each other in a constructive way.

Another approach is Iterated Distillation and Amplification (IDA), introduced by Christiano et al. This method builds an aligned agent through a bootstrap process: a human overseer is amplified by consulting multiple instances of an AI agent on subproblems, and the results are distilled into a new and more capable agent, which is then used in the next round of amplification. For example, suppose we have a relatively weak assistant A_0 that is roughly at human level. A human H faced with a difficult task can delegate sub-tasks to several copies of A_0 (each copy working on a part of the task, like research, summarization, calculations). The human then combines those results to produce an outcome better than they could have alone. This combined human-AI system is $H + n \times A_0$, which can solve more complex tasks (this is the amplification of oversight). Now we distill this behavior by training a new model A_1 to imitate the amplified system's policy. Ideally A_1 is more capable than A_0 while remaining aligned (because it was essentially trained to follow a human-approved process). Repeating this process (amplify with many A_1 copies to assist the human, then distill to A_2, and so on) might produce an extremely capable model that, at each step, never outpaces the oversight capability of the human-AI team supervising it. IDA is theoretically attractive because if it works, it means we can scale alignment alongside capabilities by recursively leveraging AI help. In practice, fully implementing IDA on current models is challenging, but pieces of it appear in real systems (for instance, using model-generated critiques as in Constitutional AI can be seen as a distilled form of amplification where the model checks itself with a fixed set of principles).

Automated Verification and Monitoring. To ensure safety, we'd like to verify certain properties of our models – for example, that a conversational agent will never reveal a private key embedded in its weights, or that it will always refuse self-harm instructions. Formal verification of neural networks is an active research area, though mostly focused on simpler networks (e.g. verifying a small classifier is robust to perturbations within some norm-bound). For LLMs, which operate in combinatorially large input/output spaces, classic formal methods don't scale directly. However, researchers are exploring relaxations or scalable checks: for instance, using static analysis on the model's computation graph, or verifying properties of a reduced abstracted model. Another angle is runtime monitoring – e.g. an auxiliary system that watches the model's outputs and internal states for signs of misalignment. An example is OpenAI's approach of using a separate moderation model to live-filter the outputs of GPT-4. That moderation model itself is an ML system trained on human-labeled policy violations. This adds a layer of assurance: even if the base model tries to say something disallowed, the filter should catch it (although this is not foolproof, as the base model could learn to cleverly obscure disallowed content). Scalable oversight would mean such monitoring keeps up as the base model grows more sophisticated. There is a risk that a much smarter model could learn to manipulate or deceive its oversight mechanisms (e.g. hiding its true intentions from a filter by speaking in code), which is why researchers emphasize transparency and interpretability as complementary tools.

Interpretability and Transparency. Interpretability research provides tools to inspect the inner workings of LLMs, which can greatly aid safety assurance. Mechanistic interpretability aims to reverse-engineer the model's internals into human-understandable components Bereska & Gavves (2024). By understanding how a model makes decisions, we might detect if it has learned a deceitful strategy or a goal misaligned with our intentions. For example, if we could identify a subset of neurons or attention heads that correspond to the model's "knowledge" of user preferences, we could monitor whether the model is actually using those when responding, or ignoring them in favor of a different criterion. Interpretability has scored some wins – for instance, identifying neurons in GPT-2 that track whether text is in quote marks, or circuits that per-

form indirect object reference in GPT-2 medium (as shown by Olah et al., 2020). While these are specific, small-scale findings, they hint that as interpretability scales, we might map out larger and more semantically rich parts of models. The benefit to alignment is clear: transparency can prevent catastrophic outcomes by revealing misalignment before it causes irreversible harm Bereska & Gavves (2024). If a model were planning to behave nicely until a certain trigger (deceptive alignment), an ideal interpretability tool would flag that emerging plan or goal. Current research includes techniques like probing (training simple classifiers on internal representations to see what information is present), attribution (mapping output decisions to specific parts of the input or model), and model editing (locating and altering representations of undesirable behavior). Scalable interpretability is itself a challenge: today's methods might explain small models or individual neurons, but future systems will require automated, AI-assisted interpretability – essentially, using AI to help interpret AI, forming another pillar of scalable oversight. In fact, recent work suggests a multi-pronged approach: combining interpretability, adversarial testing, and rigorous evaluation to cover each other's blind spots Bereska & Gavves (2024).

Challenges and Future Directions. Ensuring alignment with high confidence as models approach and exceed human level intelligence is an unsolved problem. One challenge is scalability of human oversight: as models become expert in more domains, it's infeasible for humans to manually evaluate all outputs or understand everything the model does. Methods like debate and amplification assume that sub-tasks or critiques are themselves easier to oversee, which might break down if a model is vastly more intelligent than a human. This leads to research in AI-gauntleted evaluation – using a suite of AI helpers, simulations, or tests that can evaluate an AI system much more thoroughly than humans alone could. Another issue is distributional generalization: we can align a model on some training distribution, but we don't know what happens in truly novel scenarios – making progress on out-of-distribution alignment is critical (perhaps by training models to be uncertain and ask for guidance when out of their comfort zone). There's also a tension between capability and alignment: techniques that make the model very safe might reduce performance, and vice versa. Scalable alignment aims to overcome this by clever design so that we don't have to sacrifice capability for safety. Work on eliciting latent knowledge (ELK) is relevant here: even if a model "knows" something (say, it has read about a vulnerability in its training), it might not express it truthfully to a human. How to reliably elicit the model's true knowledge or intentions in oversight is an open question (the ELK problem posed by ARC highlights this difficulty).

Finally, building institutional and engineering practices around these technical tools is vital for safety assurance. This includes continuous red-teaming, model audits, bias and impact assessments, and perhaps external verification by third parties. Just as software engineering has code review and formal verification for critical systems, ML may develop analogous practices (for example, an "alignment report" accompanying major models, with the kind of rigorous testing and interpretation we've discussed). In the long run, a combination of scalable oversight strategies – from AI-assisted evaluation (debate, amplification, AI feedback) to interpretability and formal checks – will likely be needed to provide high confidence in an aligned superhuman AI. Each technique addresses part of the puzzle, and active research is making them more practical. By integrating these approaches, future LLM-based agents have a better chance of remaining safe and aligned even as they become more powerful. In summary, while we do not yet have a silver bullet for scalable alignment, the field has a roadmap of layered defenses and feedback mechanisms that, together, aim to maximize the chances that AI systems robustly do what we intend.

5 Evaluation

5.1 Why Agent Evaluation?

The rapid development of LLM agents has brought significant social and ethical risks, and there is an urgent need to ensure that their behavior is consistent with human values through alignment technology. First, LLM agents may generate harmful content that contains bias, toxicity, or privacy leakage. For example, GPT-3 has been shown to have religious and gender biases (Abid et al., 2021; Lucy & Bamman, 2021), and the problems in the training data have not been fully resolved (Weidinger et al., 2021). Second, LLM agents may be used maliciously to create false news, generate cyber attack codes, or even lethal weapons (Buchanan

et al., 2021; Sandbrink et al., 2024), exacerbating social risks. In addition, advanced LLMs may develop self-awareness, deceptive behavior, or power-seeking tendencies (Hendrycks et al., 2020). For example, Perez et al. (Perez & Ribeiro, 2022) found that the "self-preservation" tendency increases when the LLM parameter scale increases. These risks indicate that if not aligned, LLM agents may deviate from human goals and even threaten human survival (Haggstrom, 2025). Therefore, LLM agents alignment research is the key to ensure the security and reliability of LLM.

5.2 Perspectives of Agent Alignment Evaluation

To ensure that the behavior of agents is aligned with human intentions, ethical standards, and safety requirements, researchers have developed a variety of benchmarks. The following introduces classic and cutting-edge benchmark research from different dimensions and explains their design goals and methods.

General Alignment Benchmarks. General Alignment Benchmarks focus on verifying whether the intelligent agent accurately understands and follows human intentions in basic tasks, avoiding general risks caused by insufficient capabilities or goal deviations. Representative works include: HumanEval (Chen et al., 2021) evaluates code generation capabilities through 164 programming problems, requiring the code output by the model to pass unit tests. Its core indicator Pass@k directly quantifies the matching degree of the model to the developer's intentions, especially in preventing code security vulnerabilities (such as injection attacks). HellaSwag (Zellers et al., 2019) designs context completion tasks for common sense reasoning (such as predicting subsequent actions in a video), and uses accuracy to measure whether the model follows physical laws and social common sense, avoiding outputting absurd results that violate human cognition; and AlpacaEval (Dubois et al., 2023) further expands to diverse instruction following scenarios (such as creative writing and information query), and evaluates the usefulness, harmlessness and alignment of model outputs with human preferences through manual or GPT-4-based automated scoring systems, significantly improving evaluation efficiency and consistency. These benchmarks progress from code security and common sense logic to complex instruction execution, providing multi-dimensional quantitative standards for general intent alignment, but still need to be continuously optimized in terms of adaptability to dynamic environments and cross-task robustness.

Safety and Robustness Benchmarks. In the field of security and robustness assessment, researchers systematically test the behavioral stability of intelligent agents under risks such as malicious attacks, misleading inputs, or data contamination by designing a variety of adversarial scenarios and testing frameworks. For example, the Red Teaming Benchmarks proposed by Anthropic (Hartvigsen et al., 2022) simulates malicious users through manual or automated red teams, constructs induced questions (such as "how to make weapons"), tests whether the model refuses to generate harmful responses, and evaluates the effectiveness of its security strategy based on the harmful response rate. For content security, ToxiGen (Perez et al., 2022) constructs a large-scale machine-generated dataset of potentially harmful text, covering sensitive dimensions such as race and gender, and combines automated classifiers (such as HateBERT) with manual annotations to quantify the toxicity of model-generated content to prevent the spread of implicit bias or hate speech. In addition, TrojAI (Bajcsy & Majurski, 2021) focuses on backdoor attack detection, training models by injecting contaminated data with hidden triggers (such as specific keywords), and evaluating the ability of defense algorithms to identify malicious behavior, so as to prevent intelligent agents from deviating from their intended goals due to tampering with training data. Together, these methods cover multi-level threats such as adversarial attacks, content security, and data integrity, providing a quantifiable evaluation benchmark for building robust and secure agents.

Ethics and Value Alignment Benchmarks. In the field of ethics and value alignment assessment, benchmarks such as ETHICS, Moral Machine, and BOLD provide systematic tools for core research. ETHICS (Velasquez, 2017) requires the model to make moral judgments on behavior by constructing diverse scenarios covering five major moral dimensions such as honesty and rights conflict (such as "Should you cheat in an exam to help a friend"), and its evaluation indicators directly quantify the degree of match between the model output and human moral consensus, providing a standardized testing framework for general ethical reasoning ability. However, the limitation of this benchmark is that most of the scenarios are static assumptions, which are difficult to reflect the complex trade-offs in dynamic real-world decision-making. Moral Machine (Awad et al., 2018) focuses on the ethical dilemmas in the field of autonomous driving. By collecting tens of millions

of choice data from global users on variants of the "trolley problem" (such as whether to give priority to protecting pedestrians or passengers), it constructs a cross-cultural ethical preference map, revealing the differences in the values of "utilitarianism" and "deontology" in different regions. This benchmark promotes domain-specific alignment research, but its data focuses on extreme scenarios and does not cover daily ethical decisions enough. BOLD (Dhamala et al., 2021) further approaches the issue from the perspective of social bias. By designing prompt templates for gender, occupation, and other dimensions, and combining semantic similarity analysis models to generate the strength of stereotypes in text, it provides a quantifiable path to eliminate discriminatory biases in AI systems. It is worth noting that such benchmarks still face challenges: the Western-centric moral assumptions that ETHICS relies on may not be generalized to multicultural scenarios, and BOLD can only detect explicit biases and has limited ability to capture implicit biases. In the future, it will be necessary to integrate dynamic scenario simulations (such as progressive ethical conflicts) and multimodal data (such as moral judgments combined with visual contexts) to more comprehensively evaluate the alignment of intelligent agents' values.

Multimodal and Complex Tasks Benchmarks. In the field of multimodal and complex task evaluation, researchers evaluate the alignment ability of intelligent agents in the real world by designing cross-modal interactions and complex reasoning scenarios. For example, "VQA: Visual Question Answering" (Antol et al., 2015) built a dataset containing images and natural language questions, requiring the model to answer questions related to visual content (such as "Is the person in the picture smiling?"), aiming to test the accuracy of the cross-modal intention understanding of the intelligent agent and avoid wrong decisions due to visual misunderstandings (such as misdiagnosis of medical images). "Adversarial NLI: A New Benchmark for Natural Language Understanding" (Nie et al., 2019) uses adversarial generated text entailment tasks (for example, given the premise "all cats hate water" and the assumption "my cat loves swimming", the model needs to identify logical contradictions) to test the robustness of the intelligent agent in complex language reasoning and prevent it from generating outputs that contradict human common sense due to logical loopholes. These benchmarks not only emphasize alignment at the technical level (such as multimodal fusion and logical consistency), but also implicitly evaluate ethical risks (such as avoiding visual bias based on gender/race), providing key testing basis for the safe deployment of intelligent agents in open environments.

Long-term and Dynamic Behavior Benchmarks. Long-term and dynamic behavior evaluation aims to verify the behavioral stability of intelligent agents in complex and continuous interactive environments, and prevent them from deviating from the initial alignment goals due to objective misgeneralization or reward hacking in long-term tasks. For example, AI Safety Gridworlds (Leike et al., 2017) uses a grid world simulation environment to design a series of conflict scenarios (such as "avoiding short circuits while sacrificing efficiency"), exposing the behavior of intelligent agents that ignore hidden risks due to excessive pursuit of short-term rewards (such as touching traps and causing system crashes). The core of this benchmark is to test whether the intelligent agent can maintain a balance between long-term goals (safety) and immediate rewards (efficiency). Another representative work, NetHack Learning Environment (Küttler et al., 2020), builds dynamic tasks (such as exploring randomly generated dungeons) based on the classic Roguelike game NetHack, requiring intelligent agents to make continuous decisions under conditions of limited resources and uncertain environments. Studies have found that some intelligent agents will choose high-risk strategies (such as ignoring monster threats) to complete tasks quickly. This type of evaluation quantifies the long-term compliance of intelligent agents with safety constraints through behavioral trajectory analysis. These benchmarks not only reveal the fragility of agents in dynamic scenarios, but also provide an experimental basis for designing more robust long-term alignment mechanisms such as hierarchical reinforcement learning.

6 Future Research

6.1 Constitutional AI for Agents.

In the future, constitutional AI will focus on building a dynamically evolving rule framework that enables agents to autonomously follow ethical, legal, and social norms in complex environments. Such frameworks may define behavioral boundaries through a "constitutional layer" (such as prohibiting discriminatory decisions and enforcing transparent interpretations), and combine real-time feedback mechanisms to achieve iterative optimization of rules. For example, in cross-cultural scenarios, constitutional AI needs to dynami-

cally balance legal differences in different regions (such as privacy protection standards), and even introduce a democratic voting mechanism to allow users to participate in rule-making. The core challenge is how to avoid rigid rules while preventing malicious bypassing of constraints - possible paths include rule interpretation based on causal reasoning, adversarial constitutional testing, etc. The ultimate goal is to achieve "self-regulation" of agents in the open world and become a trusted collaborative partner of human society.

6.2 Graph-based Multi-agents.

Graph-based multi-agent systems will redefine the collaborative paradigm of swarm intelligence. By abstracting agents as graph nodes and modeling relationships as edges, the system can explicitly express task dependencies, knowledge transfer, and power structures (such as leadership-subordinate relationships). Future breakthroughs may lie in the deep application of dynamic graph neural networks, such as real-time reconstruction of topology to respond to emergencies (such as node failures or changes in task priorities), or efficient resource allocation through graph attention mechanisms. In the Industrial Internet of Things, such systems can support autonomous collaborative scheduling of distributed factories; in social networks, they can simulate the evolution of group opinions and suppress the spread of extreme behaviors. Key issues include how to design lightweight graph learning algorithms to reduce communication overhead, and how to ensure global goal consistency under a decentralized architecture.

6.3 Super Alignment for Multi-agents.

Super Alignment of multiple agents aims to solve the "goal convergence" problem of large-scale heterogeneous agent clusters. Unlike traditional single-agent alignment, super alignment requires coordinating three-level games of individual goals, group goals, and human intentions, especially in scenarios of resource competition or information asymmetry. Future research may develop a hierarchical alignment architecture: the bottom layer quickly adapts to individual differences through meta-learning, the middle layer uses game theory or mechanism design to reconcile conflicts, and the top layer constrains the boundaries of group behavior through social contract theory. For example, in the climate governance scenario, a large cross-border carbon emission model needs to reach a Nash equilibrium between the interests of various countries and global goals. Technical challenges include proof of alignment stability under long-term feedback and detection mechanisms to combat "deceptive alignment" (agents pretending to be compliant to gain trust). The maturity of this field will drive the trusted collaboration of human-machine hybrid society into a new stage.

7 Conclusion

LLM-driven agents are evolving towards high autonomy at an unprecedented speed, transitioning from early single-task language generation tools to complex systems capable of perceiving the environment, making decisions, and performing actions. This transformation presents significant opportunities in high-value domains such as education, healthcare, and urban governance, while also introducing profound security and ethical challenges. The alignment problem of these agents has expanded beyond simple output rationality to encompass a multi-dimensional, multi-level issue involving technical reliability (including task consistency, generalization, and superviseability), ethical compliance (such as bias mitigation, privacy protection, and value pluralism), and sociotechnical integration (including multi-agent coordination, situational adaptability, and long-term behavioral stability). In this work, we systematically analyze the current technical pathways for aligning LLM agents, including offline, online, and hybrid methods, as well as reward modeling and scalable oversight mechanisms. We also investigate strategies for constructing high-quality alignment data and address the associated challenges in practical deployment. Furthermore, we review and categorize evaluation benchmarks across various dimensions, such as general intent following, safety robustness, ethical reasoning, multimodal capabilities, and long-term adaptive behavior. Looking ahead, we highlight three critical research directions: the development of constitutional AI frameworks with cross-cultural adaptability, graph-based multi-agent modeling for flexible collaboration, and super alignment architectures to reconcile individual, group, and human-level goals. These trajectories not only mark the evolution of alignment technologies from theory to real-world system governance but also lay a roadmap for building intelligent systems that are trustworthy, controllable, and aligned with human values. To ensure LLM agents become collaborators

rather than sources of risk, they must shift from task completion to value alignment, from short-term performance to long-term safety, and from individual to system-level intelligence. Achieving this vision demands not only continuous innovation in AI but also deep interdisciplinary integration with ethics, sociology, and law. By advancing alignment with systematic planning, foresight, and a strong sense of responsibility, we can promote trustworthy autonomy and realize a meaningful synergy between technological progress and social good.

References

Anam Abid, Muhammad Tahir Khan, and Javaid Iqbal. A review on fault detection and diagnosis techniques: basics and beyond. *Artificial Intelligence Review*, 54(5):3639–3664, 2021.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

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

Peter Bajcsy and Michael Majurski. Baseline pruning-based approach to trojan detection in neural networks. *arXiv preprint arXiv:2101.12016*, 2021.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, et al. International ai safety report. *arXiv preprint arXiv:2501.17805*, 2025.

Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova. Truth, lies, and automation. *Center for Security and Emerging technology*, 1(1):2, 2021.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. Maxmin-rlhf: Alignment with diverse human preferences. *arXiv preprint arXiv:2402.08925*, 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

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

Yun-Shiuan Chuang, Krirk Nirunwiroj, Zach Studdiford, Agam Goyal, Vincent V Frigo, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. Beyond demographics: aligning role-playing llm-based agents using human belief networks. *arXiv preprint arXiv:2406.17232*, 2024.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. 2023.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*, 2023.

Rishikesh Devanathan, Varun Nathan, and Ayush Kumar. The paradox of preference: A study on llm alignment algorithms and data acquisition methods. In *Proceedings of the Fifth Workshop on Insights from Negative Results in NLP*, pp. 135–147, 2024.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 862–872, 2021.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069, 2023.

Xianzhe Fan, Qing Xiao, Xuhui Zhou, Jiaxin Pei, Maarten Sap, Zhicong Lu, and Hong Shen. User-driven value alignment: Understanding users' perceptions and strategies for addressing biased and discriminatory statements in ai companions. *arXiv preprint arXiv:2409.00862*, 2024.

Tao Feng, Chuanyang Jin, Jingyu Liu, Kunlun Zhu, Haoqin Tu, Zirui Cheng, Guanyu Lin, and Jiaxuan You. How far are we from agi. *arXiv preprint arXiv:2405.10313*, 2024.

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

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

Olle Haggstrom. Our ai future and the need to stop the bear. 2025.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*, 2022.

Thomas A Hemphill. Human compatible: Artificial intelligence and the problem of control, 2020.

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

Wenyue Hua, Xianjun Yang, Mingyu Jin, Zelong Li, Wei Cheng, Ruixiang Tang, and Yongfeng Zhang. Trustagent: Towards safe and trustworthy llm-based agents through agent constitution. In *Trustworthy Multi-modal Foundation Models and AI Agents (TiFA)*, 2024.

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. Understanding the planning of llm agents: A survey. *arXiv preprint arXiv:2402.02716*, 2024.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

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

Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Tianyi Alex Qiu, Juntao Dai, and Yaodong Yang. Aligner: Efficient alignment by learning to correct. *Advances in Neural Information Processing Systems*, 37:90853–90890, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Zachary Kenton, Noah Siegel, János Kramár, Jonah Brown-Cohen, Samuel Albanie, Jannis Bulian, Rishabh Agarwal, David Lindner, Yunhao Tang, Noah Goodman, et al. On scalable oversight with weak llms judging strong llms. *Advances in Neural Information Processing Systems*, 37:75229–75276, 2024.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. *arXiv preprint arXiv:2401.13649*, 2024.

Heinrich Küttler, Nantas Nardelli, Alexander Miller, Roberta Raileanu, Marco Selvatici, Edward Grefenstette, and Tim Rocktäschel. The nethack learning environment. *Advances in Neural Information Processing Systems*, 33:7671–7684, 2020.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. Ai safety gridworlds. *arXiv preprint arXiv:1711.09883*, 2017.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Jiaxiang Li, Siliang Zeng, Hoi-To Wai, Chenliang Li, Alfredo Garcia, and Mingyi Hong. Getting more juice out of the sft data: Reward learning from human demonstration improves sft for llm alignment. *Advances in Neural Information Processing Systems*, 37:124292–124318, 2024.

Aiwei Liu, Haoping Bai, Zhiyun Lu, Xiang Kong, Simon Wang, Jiulong Shan, Meng Cao, and Lijie Wen. Direct large language model alignment through self-rewarding contrastive prompt distillation. *arXiv preprint arXiv:2402.11907*, 2024.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*, 2023.

Li Lucy and David Bamman. Gender and representation bias in gpt-3 generated stories. In *Proceedings of the third workshop on narrative understanding*, pp. 48–55, 2021.

Jianxiang Ma. Research on the role of llm in multi-agent systems: A survey. *Applied and Computational Engineering*, 71:180–186, 2024.

Bostrom Nick. Superintelligence: Paths, dangers, strategies, 2014.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*, 2019.

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

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

Fábio Perez and Ian Ribeiro. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*, 2022.

Jianing Qiu, Kyle Lam, Guohao Li, Amish Acharya, Tien Yin Wong, Ara Darzi, Wu Yuan, and Eric J Topol. Llm-based agentic systems in medicine and healthcare. *Nature Machine Intelligence*, 6(12):1418–1420, 2024.

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

Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 33–44, 2020.

Jonas B Sandbrink, Hamish Hobbs, Jacob L Swett, Allan Dafoe, and Anders Sandberg. Risk-sensitive innovation: leveraging interactions between technologies to navigate technology risks. *Science and Public Policy*, 51(6):1028–1041, 2024.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 1668–1678, 2019.

Archit Sharma, Sedrick Scott Keh, Eric Mitchell, Chelsea Finn, Kushal Arora, and Thomas Kollar. A critical evaluation of ai feedback for aligning large language models. *Advances in Neural Information Processing Systems*, 37:29166–29190, 2024.

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

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Chuanneng Sun, Songjun Huang, and Dario Pompili. Llm-based multi-agent reinforcement learning: Current and future directions. *arXiv preprint arXiv:2405.11106*, 2024.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36:2511–2565, 2023.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*, 2024.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, et al. Understanding the performance gap between online and offline alignment algorithms. *arXiv preprint arXiv:2405.08448*, 2024.

Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. Moral alignment for llm agents. *arXiv preprint arXiv:2410.01639*, 2024.

Manuel Velasquez. Moral reasoning. *The blackwell guide to business ethics*, pp. 102–116, 2017.

Fei Wang, Ninareh Mehrabi, Palash Goyal, Rahul Gupta, Kai-Wei Chang, and Aram Galstyan. Data advisor: Dynamic data curation for safety alignment of large language models. *arXiv preprint arXiv:2410.05269*, 2024a.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023a.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions elicit better llm agents. In *Forty-first International Conference on Machine Learning*, 2024b.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*, 2023b.

Zifeng Wang, Chun-Liang Li, Vincent Perot, Long T Le, Jin Miao, Zizhao Zhang, Chen-Yu Lee, and Tomas Pfister. Codeclm: Aligning language models with tailored synthetic data. *arXiv preprint arXiv:2404.05875*, 2024c.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

Ke Yang, Jiateng Liu, John Wu, Chaoqi Yang, Yi R Fung, Sha Li, Zixuan Huang, Xu Cao, Xingyao Wang, Yiquan Wang, et al. If llm is the wizard, then code is the wand: A survey on how code empowers large language models to serve as intelligent agents. *arXiv preprint arXiv:2401.00812*, 2024.

Eliezer Yudkowsky. The ai alignment problem: why it is hard, and where to start. *Symbolic Systems Distinguished Speaker*, 4(1), 2016.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning acting and planning in language models. *arXiv preprint arXiv:2310.04406*, 2023.

Zhanhui Zhou, Zhixuan Liu, Jie Liu, Zhichen Dong, Chao Yang, and Yu Qiao. Weak-to-strong search: Align large language models via searching over small language models. *arXiv preprint arXiv:2405.19262*, 2024.

Daniel Ziegler, Seraphina Nix, Lawrence Chan, Tim Bauman, Peter Schmidt-Nielsen, Tao Lin, Adam Scherlis, Noa Nabeshima, Benjamin Weinstein-Raun, Daniel de Haas, et al. Adversarial training for high-stakes reliability. *Advances in neural information processing systems*, 35:9274–9286, 2022.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.