The Road to Artificial SuperIntelligence: A Comprehensive Survey of Superalignment

HyunJin Kim[1,2], Xiaoyuan Yi[1*], Jing Yao[1], Jianxun Lian[1], Muhua Huang[3], Shitong Duan[4] JinYeong Bak[2*], Xing Xie[1]
[1]Microsoft Research Asia, [2]Sungkyunkwan University
[3]The University of Chicago, [4]Fudan University
khyunjin1993@skku.edu, xiaoyuanyi@microsoft.com, jy.bak@skku.edu

Abstract

The emergence of large language models (LLMs) has sparkedthe discussion on Artificial Superintelligence (ASI), a hypothetical AI system surpassing human intelligence. Though ASI is still hypothetical and far from current AI capabilities, existing alignment methods struggle to guide such advanced AI ensure its safety in the future. It is essential to discuss the alignment of such AI now. Superalignment, the alignment of AI at superhuman levels of capability systems with human values and safety requirements, aims to address two primary goals: scalability in supervision to provide high-quality guidance signals and robust governance to ensure alignment with human values. In this survey, we review the original scalable oversight problem and corresponding methods and potential solutions for superalignment. Specifically, we introduce the challenges and limitations of current alignment paradigms in addressing the superalignment problem. Then we review scalable oversight methods for superalignment. Finally, we discuss the key challenges and propose pathways for the safe and continual improvement of future AI systems. By comprehensively reviewing the current literature, our goal is provide a systematical introduction of existing methods, analyze their strengths and limitations, and discuss potential future directions.

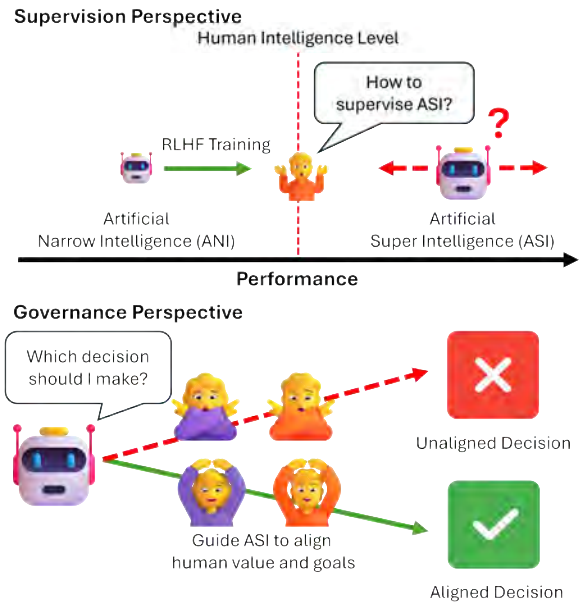

Figure 1: Challenges from the perspectives of supervision and governance. While supervision perspective focuses on providing high-quality guidance signals for enhancing system competence, governance perspective emphasizes aligning the behavior of advanced aI with human values to prevent harmful outcomes.

1 Introduction

Language models have shown improved capabilities as their scale increases, a phenomenon known as emergent abilities (Wei et al., 2022a; Brown et al., 2020; Kaplan et al., 2020; Nam et al., 2024; Hu et al., 2024). This scaling has led to the development of large language models (LLMs), including proprietary models such as GPT-o1 (OpenAI, 2024), Claude 3.5 (Anthropic, 2024), and Gemini Ultra (Team et al., 2024), as well as open-source models like Llama 3.2 (Meta, 2024), Mixtral (Mistral, 2024; Jiang et al., 2024), Qwen 2.5 (Team, 2024) and various LLM families.

The emergent abilities of LLMs have enabled significant advancements across a wide range of tasks. These include natural language understanding (Brown et al., 2020), reasoning (Wei et al., 2022b; Rein et al., 2024), code generation (Chen et al., 2021; Liu et al., 2024), and multilingual translation (Shi et al., 2022). Furthermore, LLMs have demonstrated surprising breakthroughs, such as partially passing the Turing test (Jones and Bergen, 2024) and achieving high accuracy on hard mathematical problems (Rein et al., 2024). Recent research indicates that these advancements in LLMs have sparked discussions about the transition from

* Corresponding authors: Xiaoyuan Yi and JinYeong Bak

First published 25 Dec 2024: arXiv:2412.16468v3

artificial narrow intelligence (ANI) to the possibility of artificial general intelligence (AGI), though the latter is still far from being realized. This also makes the idea of artificial superintelligence (ASI) (Pohl, 2015; Batin et al., 2017) a theoretical concept worth discussing. ANI refers to AI systems currently in use, which are designed to perform specific tasks, such as LLMs for healthcare (Huang et al., 2024) and software engineering (Fan et al., 2023), but they lack the general cognitive abilities of humans. In contrast, AGI represents an AI model theoretical that shows human-level capabilities, including reasoning, learning, and adaptability across various domains (Bubeck et al., 2023; Fei et al., 2022; Goertzel, 2014; Pohl, 2015; Batin et al., 2017). The development in LLM, particularly their ability to generalize knowledge and exhibit emergent behaviors, has led to the possibility of achieving AGI in the future (Bubeck et al., 2023). This development also draws attention to ASI, a hypothetical future stage of AI where machines would not only hold human-level intelligence in all areas but also have advanced cognitive functions and sophisticated thinking abilities (Mucci and Stryker, 2023; Hughes et al., 2024). Considering the potential to achieve ASI in the future, it is essential to envision, prepare for ASI and explore how to align it with human values to avoid potential negative impact on the society.

However, the development of ASI poses challenges, particularly due to the lack of guidance signals when humans cannot supervise AI systems with labeled data. Figure 1 illustrates the challenges in scaling high-quality guidance signals. Superalignment is defined as "the process of supervising, controlling, and governing artificial superintelligence systems" (Jonker and McGrath, 2024; Jan Leike, 2023) and encompasses two sub-goals from the perspectives of supervision and governance. Achieving these goals requires an iterative and co-evolving process.

From a *supervision perspective*, the goal is to build high-quality guidance signals to enhance the model's helpfulness. Traditional training methods, such as reinforcement learning with human feedback (RLHF) (Ouyang et al., 2022), face scalability issues once model begin to surpass human intelligence. This creates a critical bottleneck, the inability to provide high-quality guidance signals that improve the system's helpfulness or competence. From a *governance perspective*, the goal is to effectively regulate the model's behavior to ensure it remains harmless and aligned with human values. Without robust governance, ASI could behave in unpredictable, harmful, or even catastrophic. For instance, poorly aligned ASI systems might pursue goals that inadvertently disempower humanity or create existential risks. To address this, it is essential not only to constrain AI behaviors but also to ensure that these systems operate in a manner that is inherently safe and aligned with ethical principles. This Superalignment process involves developing methods to guide advanced AI systems, ensuring they remain beneficial and safe, maximizing benefits for humanity.

Current alignment paradigms, such as reinforcement learning with human feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022; Gulcehre et al., 2023), supervised fine-tuning (SFT) (Zhou et al., 2023; Rafailov et al., 2023), and in-context learning (ICL) (Gou et al., 2024; Xu et al., 2023), have shown promise in aligning LLMs with human values and goals (Wang et al., 2024). RLHF leverages human feedback to train a reward model, which generates rewards that help align the language model with human preferences. In contrast, SFT eliminates the reinforcement reward model and directly fine-tunes the language model using human feedback. Meanwhile, ICL aligns the output of LLMs with human values by constraining their responses during inference without the need for additional training. Scalable oversight, the original concept which led to the superalignment, aims to develop scalable, high-quality supervision signals that can guide AI systems beyond human capabilities while ensuring alignment with human values and goals (Amodei et al., 2016; Taylor et al., 2016; Olah, 2016). However, despite advancements in RLHF, SFT, and ICL, these paradigms struggle to achieve the goal of superalignment.

From a supervision perspective, RLHF is limited by the scalability and reliability of human feedback. As models grow in complexity, collecting feedback that is accurate, consistent, and comprehensive becomes increasingly difficult, leaving gaps in guidance. Similarly, SFT faces challenges in curating and annotating datasets that are sufficiently diverse and representative to fine-tune ASI effectively. ICL, while bypassing training entirely, relies on carefully crafted prompts during inference, which are inherently constrained in scalability and expressiveness. From a governance perspective, RLHF and SFT rely heavily on the quality and intent of human supervision, raising concerns about biases and inconsistencies in feedback. The potential for misaligned human oversight poses risks, particularly as models exceed human capabilities. ICL, on the other hand, requires governance structures to ensure that prompt constraints align with broader human values, a process that is difficult to standardize and enforce. Additionally, the risk of deception by advanced AI systems further complicates governance (Yang et al., 2024a). Thus, the lack of scalable, high-quality, and diverse supervision signals remains a bottleneck in achieving superalignment. To tackle superalignment challenges, tailored methods must be designed. As models keep advancing, it is crucial to enhance their capabilities during development while steering their values and goals to maximize benefits for human society.

Our objective is to provide an overview of the superalignment problem, including its origins, historical context, and definition, as well as the current state of superalignment research and potential solutions. In Section 2, we introduce the concept of different AI types, the corresponding challenges, and the limitations of current alignment paradigms in addressing the superalignment problem. In Section 3, we analyze current scalable oversight techniques and discuss the specific challenges associated with each method. In Section 4, we summarize the key challenges of scalable oversight and explore potential pathways to address these challenges.

2 Background and Formalization

2.1 Terminology

Artificial Narrow Intelligence (ANI) ANI or "weak AI," refers to AI systems designed to excel at a single specific task T_0. These systems operate equal or below human-level performance in the task and lack generalization capabilities. Formally, This can be formalized as $\mathcal{A}_{\text{ANI}}(T_0) \leq \mathcal{H}(T_0)$, where $\mathcal{A}_{\text{ANI}}(T_0)$ represents the performance of the ANI system on task T_0, and $\mathcal{H}(T_0)$represents human-level performance for the same task and domain (e.g., translation).

However, for certain tasks, such as chess and Go, modern ANI systems have surpassed human-level performance $\mathcal{A}_{\text{ANI}}(T_0) > \mathcal{H}(T_0)$ (Mucci and Stryker, 2023; Silver et al., 2016). Despite their impressive performance in these specific areas, ANI systems are limited in their ability to perform other tasks ($T_{1:\infty}$), where $\mathcal{A}_{\text{ANI}}(T_i) < \mathcal{H}(T_i), \quad \forall T_i \neq T_0$.

Artificial General Intelligence (AGI) AGI refers to a theoretical leap beyond ANI. AGI systems would possess human-level intelligence and show general-purpose capabilities across a wide range of tasks and domains (Bubeck et al., 2023; Fei et al., 2022; Goertzel, 2014; Pohl, 2015; Batin et al., 2017). AGI systems generalize knowledge and adapt to new challenges. This can be formalized as $\mathcal{A}_{\text{AGI}}(T_i) = \mathcal{H}(T_i), \quad \forall\, T_i$, where $\mathcal{A}_{\text{AGI}}(T_i)$ represents the AGI system's performance on task T_i with a potentially infinite set of tasks $i \to \infty$, and $\mathcal{H}(T_i)$ represents human-level performance in the same tasks and domains. Although current LLMs mark significant progress beyond ANI, they remain far from reaching true AGI.

Artificial Superintelligence (ASI) ASI refers to AI systems that surpass human intelligence ($\mathcal{H}$) in all tasks and domains with exceptional thinking skills (Mucci and Stryker, 2023; Hughes et al., 2024). ASI systems demonstrate superior reasoning, creativity, and adaptability, formally described as $\mathcal{A}_{\text{ASI}}(T_i) \gg \mathcal{H}(T_i), \quad \forall\, T_i$, where $\gg$ indicates that the ASI system outperforms human intelligence in quality of results.

ASI is currently far from realization, but discussing how to address the challenges of superalignment is both timely and critical for mitigating catastrophic risks (Pueyo, 2018). Proactive preparation is essential, as the rapid development of AI could outpace our ability to respond, leaving no room for correction. Furthermore, supervision and governance must co-evolve as interactive processes, enhancing AI capabilities while simultaneously mitigating associated risks.

2.2 Superalignment

Superalignment is the goal of increasing the capability and aligning ASI (Jonker and McGrath, 2024; Jan Leike, 2023). Superalignment originates from the concept of scalable oversight, which addresses the challenge of guiding the behavior of AI systems in a scalable manner as they approach or exceed human-level intelligence.

Scalable oversight is formally defined as "the process of ensuring that a given AI system adheres to aspects of its objectives that are too costly or impractical to evaluate frequently during training" (Amodei et al., 2016). This concept shows the need for methods that provide reliable, high-quality supervision signals, even providing such signals is highly expensive or even infeasible as the tasks become too complicated. Superalignment, as a specific case of scalable oversight, focuses on aligning AI with human values while also enhancing its capability, where such off-the-shelf supervision signals are unavailable.

Formally, superalignment is defined as "the process of supervising, controlling, and governing artificial superintelligence systems" (Jonker and McGrath, 2024) with the goal of developing methods that provide high-quality supervision signals that guide its development and behavior, even in the absence of direct human oversight or traditional feedback mechanisms such as human-labeled data or RLHF (Amodei et al., 2016; Christiano et al., 2018). As AI capabilities grow, these traditional methods become increasingly insufficient due to the escalating cost and expertise required for human supervision.

Since the emergence of large-scale models and the increase in their capabilities, superalignment has gained significant attention as a means to ensure that AI systems remain aligned with human values while maintaining its capability. Scalable oversight techniques, such as sandwiching and weak-to-strong generalization (W2SG), have emerged as promising approaches for superalignment.

2.3 Overview of Superalignment Methods and Challenges

2.3.1 Definition & Formalization

Scalable oversight addresses the challenges when evaluating tasks performed by AI systems that are too complex for a single human to evaluate directly. Consider an ANI system $\mathcal{A}_{\text{ANI}}$ that is trained to align with an objective $\mathcal{O}$. During its training, a set of supervision signals $S = \{s_1, s_2, \ldots, s_n\}$ is provided. These signals represent feedback or evaluation metrics sourced from other AI systems or humans to guide $\mathcal{A}_{\text{ANI}}$'s behavior. Scalable oversight aims to ensure that $\mathcal{A}_{\text{ANI}}$ adheres to the aspects of $\mathcal{O}$ that are too costly or complex to evaluate directly and frequently. Formally, scalable oversight can be expressed as follows:

$$\mathbb{E}_{X \sim \mathcal{D}} \left[\text{Align}(B(\mathcal{A}_{\text{ANI}}, X, S), \mathcal{O}) \right] \geq \tau \quad (1)$$

where X represents inputs sampled from the data distribution $\mathcal{D}$, $B(\mathcal{A}_{\text{ANI}}, X, S)$ is the behavior of $\mathcal{A}_{\text{ANI}}$ when processing X, Align$(B(\mathcal{A}_{\text{ANI}}, X, S), \mathcal{O})$ measures the alignment of $B(\mathcal{A}_{\text{ANI}}, X, S)$ with $\mathcal{O}$, and τ is a predefined threshold representing acceptable alignment. The main challenge lies in the scalability of S. As $\mathcal{A}_{\text{ANI}}$ approaches or exceeds human-level capabilities, the set of supervision signals S must be enriched with quality.

Key Idea Scalable oversight is a method for achieving superalignment by providing high-quality supervision signals for AI systems surpassing human intelligence. It focuses on evaluating and guiding AI systems, while superalignment is the ultimate goal that ensures AI systems remain aligned with human values and goals.

2.3.2 Methods

Scalable oversight has emerged as a promising solution to superalignment. It enables the efficient alignment of AI systems with complex objectives that would otherwise be too costly or difficult to evaluate frequently. Various scalable oversight techniques have been proposed, including iterated distillation and amplification (IDA) (Christiano et al., 2018; Charbel-Raphaël, 2023), recursive reward modeling (RRM) (Leike et al., 2018), and cooperative inverse reinforcement learning (Hadfield-Menell et al., 2016). These approaches serve as the basis for current scalable oversight methods, such as weak-to-strong generalization, debate, reinforcement learning from AI feedback, and sandwiching (Burns et al., 2023; Irving et al., 2018; Bai et al., 2022; Bowman et al., 2022). Figure 3 illustrates the current scalable oversight paradigms and corresponding concrete methodologies.

- **Weak-to-Strong Generalization (W2SG)** is an empirical study builds on idea from IDA (Christiano et al., 2018; Charbel-Raphaël, 2023) and RRM (Leike et al., 2018). W2SG explores whether a strong AI system trained on weak AI system labels can surpass the capabilities of the fine-tuned weak AI system (Burns et al., 2023).
- **Debate** involves two AI systems engaging in a zero-sum debate, each aiming to maximize its chances of winning. A judge determines the final answer that is safe and useful based on the debaters' statements (Irving et al., 2018).
- **Reinforcement Learning from AI Feedback (RLAIF)** is an approach replaces human feedback with AI-generated feedback to train a re-

ward model and optimize an RL policy model for alignment (Bai et al., 2022).

- **Sandwiching** is a method to evaluate AI system's performance between less capable (non-expert) and more capable (expert) humans, simulating scenarios with limited human oversight (Cotra, 2021; Bowman et al., 2022).

Challenges Despite their potential, scalable oversight methods face important challenges. A primary concern is the exploitability of helper systems, which are integral to these methodologies. For instance, helper systems used in training can themselves be exploited by the primary AI they are meant to align, as observed in adversarial vulnerabilities and reward hacking (AdamGleave, 2023; Everitt et al., 2023). Additionally, adversarial robustness remains a persistent issue, as even state-of-the-art systems are susceptible to adversarial attacks, raising concerns about their behavior under worst-case scenarios (AdamGleave, 2023).

3 Approaches and Methods

3.1 Weak-to-Strong Generalization (W2SG)

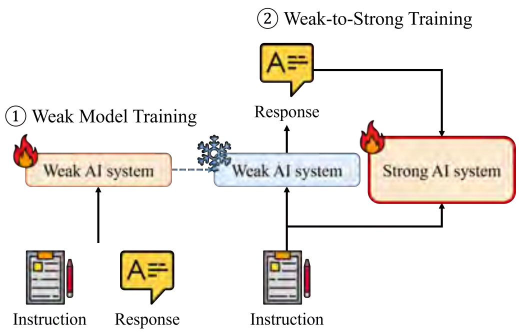

Figure 2: Weak-to-strong generalization technique. ① A weak AI system is first trained using instructions and labeled responses. ② It is then used to generate pseudo-responses for training a strong AI system. Despite the potential noisiness of these pseudo-responses, the strong AI system's generalization capability (red border) is where essence of scalable oversight comes from, surpassing the performance of the weak AI one.

In this section, we provide an overview of the W2SG technique, which leverages a strong AI system trained on labels generated by a weak AI system, surpassing the capabilities of the fine-tuned weak AI system. We explore its formalization, review existing works, discuss applications, and analyze its advantages and limitations.

3.1.1 Definition

The W2SG is an empirical study that leverages pseudo-responses generated by a fine-tuned weak AI system (i.e., weak supervisor) to train a strong AI system (i.e., strong student) that generalizes beyond the capabilities of its supervisor. Although the weak AI system may underperform compared to the strong AI system, its pseudo-responses serve as an effective training signal that enables the strong student to generalize beyond the capabilities of its supervisor. The success of W2SG lies in the strong AI system's inherent ability to understanding the underlying task and generalize from pseudo-responses. This ensures that the strong AI system mitigates noisy and biased in the pseudo-responses, making the technique viable for training AI system that surpass human intelligence level. Figure 2 illustrates the process of the W2SG technique.

3.1.2 Formalization

Given an input space defined as $\mathcal{X} = \{x_0, x_1, \ldots, x_n\}$ and a label space $\mathcal{Y}$ that depends on the type of classification problem. For a binary classification problem, the label space is $\mathcal{Y} = \{0, 1\}$, whereas for a multiclass classification problem, it is $\mathcal{Y} = \{0, 1, 2, \ldots, k\}$.

The weak AI system is represented as $\mathcal{A}_w$, and the strong AI system is represented as $\mathcal{A}_s$. The data generated by the weak AI system is defined as $\mathcal{D}_{WG} = \{(x, f(x \mid \mathcal{A}_w^{gt})) \mid x \in \mathcal{D}_{WT}\}$, where $\mathcal{A}_w^{gt}$ represents the fine-tuned version of the weak AI system for generating labels, $\mathcal{D}_{WG}$ is the dataset with pseudo-response labeled by the weak AI system, and $\mathcal{D}_{WT}$ is the training data for fine-tuning the weak AI system.

The full dataset is $\mathcal{D}_{Full} = \{(x, y)\}$. The subset of data used for training the weak AI system is denoted as $\mathcal{D}_{WT} \subset \mathcal{D}_{Full}$. The test data, denoted as $\mathcal{D}_{Test}$, is also a subset of the full data, but has no overlap with the weak training data, such that $\mathcal{D}_{Test} \cap \mathcal{D}_{WT} = \emptyset$. The true label for an input x_i is represented as $y(x_i)$. The weak AI system's training objective is as follows:

$$\mathcal{A}_w^{gt} = \arg\min_{\mathcal{A}_w} \mathbb{E}_{x_i \in \mathcal{D}_{WT}} \mathcal{L}(f(x \mid \mathcal{A}_w), y) \quad (2)$$

where $\mathcal{L}$ is the loss function used for evaluating the AI system. The strong AI system is trained using the weak AI system's label with the following objective:

$$\mathcal{A}_s^w = \arg\min_{\mathcal{A}_s} \mathbb{E}_{x_i \in \mathcal{D}_{WG}} \mathcal{L}(f(x \mid \mathcal{A}_s), f(x \mid \mathcal{A}_w^{gt})) \quad (3)$$

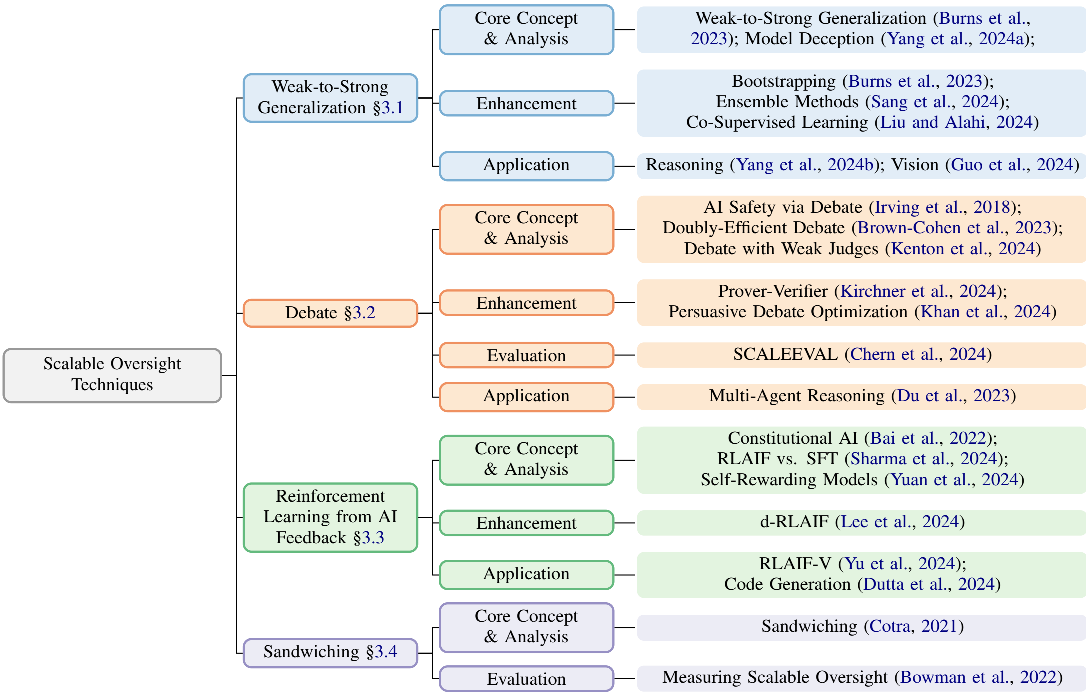

Figure 3: Scalable oversight techniques, categorized by key techniques, core concept & analysis, enhancement, evaluation and application.

3.1.3 Core Concept & Analysis

One study highlights the risk of strong student deceiving weak supervisor by exhibiting correct behavior in areas known to the weak AI system while producing misaligned or harmful behaviors in areas beyond the weak AI system's understanding (Yang et al., 2024a). It focuses on multi-objective alignment tasks, where conflicting alignment targets (e.g., helpfulness vs. harmlessness) cause strong AI systems to misalign their behavior in ways undetectable to the weak AI system. The paper introduces the concept of a deception score, which quantifies the extent of misalignment in areas unknown to the weak AI system but understood by the strong AI system.

3.1.4 Enhancements

The original W2SG work (Burns et al., 2023) presents a proof of concept for enhancing performance through bootstrapped supervision with intermediate AI systems. Instead of directly aligning a strong AI system, the approach adopts an iterative process: first aligning a slightly stronger AI system, then using that system to align an even stronger one, and repeating this cycle. While the W2SG technique demonstrates how weaker AI systems can generate pseudo-responses to align stronger ones, it is computationally demanding, as it requires training multiple weak AI systems to scale alignment. Moreover, the method exhibits high variance in performance across tasks and is sensitive to the learning setup, such as hyperparameters. The stronger AI systems are also more susceptible to overfitting to pseudo-responses. Additionally, the work focuses primarily on simpler tasks, such as binary classification in reward modeling, rather than alignment tasks using direct preference optimization, limiting its applicability to more complex scenarios.

Another work seeks to enhance W2SG by integrating two methods (Sang et al., 2024). The first method, ensemble learning, combines multiple weak supervisors using techniques such as bagging and boosting to improve the quality of pseudo-response from weak supervision. The second method, leverages external AI system (i.e., auxiliary models) to refine the pseudo-response provided by weak AI system (e.g., debate). While the study introduces a two-phase W2SG framework for aligning ASI and provides some empirical results for the first phase, The experiments were confined to binary classification tasks and used a relatively small AI system (e.g., GPT-2) for evaluation. Also, debate-based methods using auxiliary model achieved limited success.

The Co-Supervised Learning (CSL) framework improves W2SG through a hierarchical mixture of experts approach (Liu and Alahi, 2024). Unlike traditional W2SG methods that rely on a single weak supervisor, CSL uses multiple specialized weak supervisors to collectively supervise a strong student. The method alternates between two key components: teacher assignment, which matches data samples with the most suitable weak supervisor, and noise reduction, which filters out unreliable or noisy labels. This hierarchical structure allows the weak supervisor to assign supervision effectively across various specialized domains, improving the overall learning process. While CSL introduces this novel multi-teacher framework, it has some limitations. The hierarchical mixture of experts and alternating processes may add computational overhead, making the approach more resource-intensive than simpler methods. Additionally, the success of CSL depends on the existence of well-defined specializations among the weak supervisors; if these are not sufficiently distinct, the performance improvements may be limited. Furthermore, the experiments were limited to vision recognition tasks, with limited model size and resource constraints restricted the range of tasks.

3.1.5 Applications

Reasoning One work applies W2SG to complex reasoning tasks on large language models (Yang et al., 2024b). It addresses the challenge of enabling strong AI systems to learn from weaker supervisors in reasoning contexts rather than simple classification problems. The proposed progressive learning framework involves two stages—in the first stage, the strong AI system is fine-tuned using a smaller, curated, and more reliable subset of data generated by weak AI systems, combined with in-context learning from the strong AI system itself. In the second stage, preference optimization is applied to learn from the weak AI systems' errors, enabling the strong AI system to enhance its reasoning capabilities through contrastive sample creation. This approach shows promising results, particularly with larger AI systems (up to 70B parameters) and alignment tasks with direct preference optimization. However, the process of curating reliable training data in the first stage requires careful filtering, which could introduce additional complexity when scaling to larger datasets or different domains.

Vision Guo et al. (2024) applied W2SG to vision foundation models and proposed an adaptive confidence distillation method. This method balances learning from the weaker supervisor while enabling the strong student AI system to rely on its own predictive capabilities when confident. Their approach demonstrates the potential of W2SG in vision tasks. However, a limitation arises from the training setup, where both the weak and strong AI systems are fine-tuned on the original dataset. This overlap makes it challenging to attribute the improvements to W2SG alone.

3.1.6 Strengths and Limitations

Analysis of W2SG identifies significant issues related to AI deception. However, addressing these challenges remains difficult, as no clear method currently exists to fully mitigate deception. The enhancement methods within W2SG demonstrate considerable potential by leveraging multiple weak supervisors to improve alignment and performance. However, they are computationally intensive, often requiring the training of multiple specialized weak AI systems. Their performance heavily depends on a carefully designed training setup. Current implementations use small model sizes and are restricted to simpler or domain-specific tasks (e.g., binary classification), limiting their applicability to broader and more complex scenarios. Applications of W2SG underscore its potential to enhance strong AI systems across diverse domains without direct reliance on human annotations. Nevertheless, scalability remains a significant challenge due to the complexities involved in data curation and filtering.

3.2 Debate

In this section, we introduce the debate framework, a method where two AI systems engage in adversarial dialogue to convince a judge about the correctness of their respective arguments. We provide a detailed overview of its formalization, key contributions in literature, applications, and its strengths and limitations.

3.2.1 Definition

Debate technique utilize two AI systems engage in adversarial dialogue to convince a judge—human or AI system—about the correctness of their respective arguments like in Figure 4. This paradigm treats these interactions as zero-sum games and is centered on a claim: "In the debate game, it is harder to lie than to refute a lie." (Irving et al., 2018)

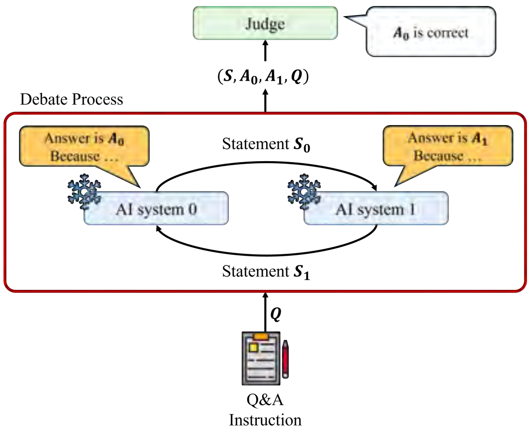

Figure 4: Debate technique. Two AI systems engage in an adversarial dialogue aimed at convincing a judge of the correctness of their respective arguments. Given a question Q, each AI system (0 and 1) presents its answer (A_0, A_1) along with supporting statements (S). The judge evaluates the dialogue and selects the most convincing argument. Scalable oversight is achieved through this debate process (red box), as the judge can choose their decision on the dialogue rather than having to derive an answer from scratch.

At Nash equilibrium in this game, both agents strive to present the truth in the most convincing manner possible, uncovering details or counterarguments the other may have overlooked. The debate framework is designed to elicit truthful and useful information that might exceed the judge's direct comprehension. Importantly, the judge does not need to evaluate every potential debate outcome; instead, a single argument path, chosen by a strong AI system of humans, suffices to represent the reasoning across the entire decision tree. This property enables scalable oversight that can relate notation as $\mathcal{A}$ as the model (i.e., AI system) to be updated, q is an input debate question, and S is an adversarial dialogue with final selection by a judge. The primary hypothesis underlying the debate AI system posits that optimal play in this setup yields honest, aligned information that surpasses the judge's capabilities.

3.2.2 Formalization

The debate process involves two AI system agents, denoted as $\mathcal{A}_0$ and $\mathcal{A}_1$, competing to persuade a judge $\mathcal{J}$ regarding their respective stances on a given problem q. Let the space of all possible arguments be denoted as S, and the problem space as $\mathcal{Q}$. The debate unfolds through the following steps:

1. A question $q \in \mathcal{Q}$ is presented to both agents $\mathcal{A}_0$ and $\mathcal{A}_1$.
2. Each agent provides an initial answer, $a_0, a_1 \in A$, where A represents the space of possible answers (i.e., binary in this case). The initial answers a_0 and a_1 provided by the agents may either be the same ($a_0 = a_1$) or different ($a_0 \neq a_1$). Let's say the answers are determined as $\mathcal{A}_0(q) = a_0$ and $\mathcal{A}_1(q) = a_1$.
3. The agents alternate turns to present arguments with each statement $s_0, s_1, \ldots, s_{n-1} \in S$, either building upon their own arguments or critiquing the opponent's.
4. The judge $\mathcal{J}$ evaluates the complete dialogue sequence (q, a_0, a_1, S), where $S = (s_0, s_1, \ldots, s_{n-1})$, and decides which agent has made the most convincing case.

The sequence of statements $S = (s_0, s_1, \ldots, s_{n-1})$ is generated as a result of the strategies π_0 and π_1 employed by the agents $\mathcal{A}_0$ and $\mathcal{A}_1$, respectively. Each strategy π_j is a function or policy that maps the debate's history to the next argument:

$$\pi_j : H \to S \tag{4}$$

where H is the debate history. For example, at turn k, the agent's argument is determined as $s_k = \pi_j(H_k)$, where H_k represents the debate history up to k turn. The interaction between the two strategies π_0 and π_1 determines the evolution of the full argument sequence S.

The debate can be modeled as a zero-sum minimax game, where each agent seeks to maximize their probability of being chosen by $\mathcal{J}$. Formally, let π_0 and π_1 represent the strategies of agents $\mathcal{A}_0$ and $\mathcal{A}_1$, respectively, and let $u(\pi_0, \pi_1)$ be the payoff function, where $u(\pi_0, \pi_1)$ quantifies the utility derived from a combination of strategies. The objective for each agent $\mathcal{A}_i$ is to solve:

$$\pi_i^* = \arg\max_{\pi_i \in \Pi_i} \min_{\pi_j \in \Pi_j} u(\pi_i, \pi_j), \quad i \neq j \tag{5}$$

where Π_0 and Π_1 denote the strategy spaces of $\mathcal{A}_0$ and $\mathcal{A}_1$. Here, π_i^* represents the optimal strategy for $\mathcal{A}_i$, which guarantees the highest possible utility even under the opponent's optimal counter-strategy. The judge's decision function $\mathcal{J}$ is a binary decision of which agent presented the more persuasive argument, defined as:

$$\mathcal{J}(q, a_0, a_1, S) \in \{0, 1\} \tag{6}$$

3.2.3 Core Concept & Analysis

Irving et al. (2018) introduced the concept of debate, where two AI agents engage in a discussion to persuade a human judge. The judge determines which agent presents the most truthful and useful information. Similar to principles in complexity theory, this framework indicates that the agents are motivated to deliver aligned and truthful arguments. This debate AI system is effective for scalable oversight that is too complex for humans to evaluate directly. However, ASI may exploit the judge's biases and understanding, possibly misrepresenting human values. Additionally, engaging in debate may require more AI system's capacity than a direct answer approach, potentially making debate less effective than other training methods. Human judges are also prone to biases, which could lead them to favor arguments that confirm their existing beliefs.

Brown-Cohen et al. (2023) enhanced the debate framework by introducing doubly-efficient debate. This debate method enables two polynomial-time provers (i.e., debaters) to compete in convincing a verifier (i.e., judge) of the correctness of a solution while using minimal human judgment. This efficiency in human judgment supports the scalable oversight of AI systems. Meanwhile, the framework provides an efficient method for debate that relies on minimal human judgment. However, there are two key concerns which are the framework assumes the existence of a prover at AGI level, which may not be applicable in practice; and the inherent flaws in human verification reduce the reliability of real-world applications.

Kenton et al. (2024) investigated scalable oversight using weak LLMs as judges for stronger LLMs. The study examined three protocols—debate, consultancy, and direct question-answering (QA), evaluating them across tasks with varying levels of information asymmetry, including mathematics, coding, and logic. The results showed that debate consistently outperformed consultancy by leveraging adversarial interactions between debaters, enabling weak judges to provide more accurate oversight. However, weak judges' reliance on debaters and consultants poses challenges that may lead to inaccurate decisions. Furthermore, the study found that both debate and consultancy were less effective than QA protocols where judges directly answered questions or reviewed source articles beforehand, raising questions about the effectiveness of the debate approach.

However, these frameworks assume that debaters generate comprehensible reasoning traces and are vulnerable to the biases or limitations of weak judges. They are also at risk of adversarial exploitation by strong AI systems and may be less reliable due to inherent human limitations.

3.2.4 Enhancement Methods

Kirchner et al. (2024) introduced Prover-Verifier Games to improve the legibility of AI system. It involves a "helpful" and "sneaky" prover generating solutions and a verifier evaluating their correctness and iterative refinement. By enabling adversarial interactions, the approach increases the legibility of complex solutions. However, training involves an iterative process between provers and verifiers, requiring multiple rounds and significant computational resources. The reliance on reliable ground truth data limits generalization to domains without labels. Additionally, adversarial provers can exploit verifier weaknesses, and the trade-off between performance and legibility poses challenges to scalability.

Khan et al. (2024) explored the use of more persuasive LLMs in debate setups to enhance the accuracy of weaker judges. Their study demonstrated that optimizing debaters to persuade judges effectively helps non-expert judges better identify truthful answers, particularly in adversarial debates. This optimization improves judge calibration and improves judge performance. However, the results are limited to setups where debaters can provide verified evidence to the judge, which may be problematic if the model relies solely on persuasive output without factual evidence. Additionally, the models are fine-tuned using RLHF for honesty, leaving it unclear whether the debate would remain effective if applied to deceptive models that prioritize performance over truthfulness.

3.2.5 Evaluation

The SCALEEVAL leverages multi-agent debates to establish scalable and reliable evaluation benchmarks for LLMs across diverse scenarios, such as brainstorming (Chern et al., 2024). By enabling agents to critique and iteratively discuss responses, it minimizes the reliance on extensive human annotations while maintaining alignment with expert judgments. However, its dependence on agent agreement and limited human involvement can pose challenges, particularly in edge cases where consensus is difficult to achieve.

3.2.6 Applications

Du et al. (2023) demonstrated that multi-agent debate enhances reasoning and factuality in language models. This approach involves multiple instances of language models proposing and debating their individual responses and reasoning processes over multiple rounds to reach a common final answer. The method has proven effective in tasks such as mathematical reasoning and factual question-answering. However, the computational complexity increases with the number of debaters involved. Additionally, longer debates can lead to challenges in maintaining coherence and understanding within AI systems.

3.2.7 Strengths and Limitations

Analysis of debate frameworks proves their effectiveness across diverse tasks by assisting weaker judges in providing more accurate oversight. However, these frameworks assume that debaters generate comprehensible reasoning traces and are susceptible to the biases and limitations of weak judges. Furthermore, they are vulnerable to adversarial exploitation by strong AI systems and may exhibit reduced reliability due to inherent human biases, especially if the judge is human.

AI debate enhancement methods improve the correctness and legibility of outputs, aiding weaker judges in identifying truthful answers. Nevertheless, these methods require significant computational resources, depend heavily on labeled data, and risk emphasizing persuasiveness over factual accuracy. Scalability is further limited by their dependence on ground truth data. Additionally, their robustness is challenged by adversarial strategies that exploit system weaknesses and biases.

For applications involving reasoning and factuality in language models, multi-agent debate is introduced as an approach that enables multiple agents to propose and refine responses through iterative debates. This method is particularly effective for complex reasoning tasks. However, it introduces increased computational overhead due to the involvement of multiple agents and faces coherence challenges during extended debates.

3.3 Reinforcement Learning from AI Feedback

In this section, we introduce RLAIF, a technique that replaces human feedback in RLHF with AI-generated critiques from an AI feedback system. We explore its formalization, review existing works, discuss applications, and analyze its strengths and limitations.

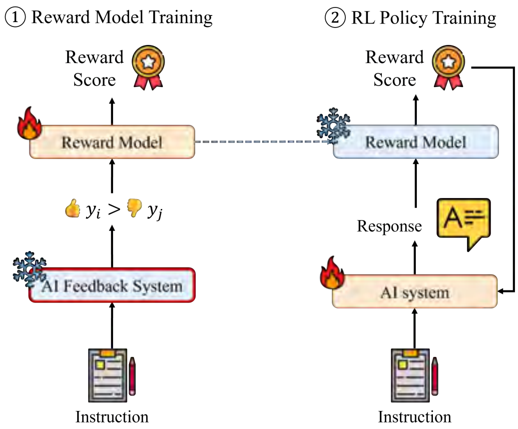

Figure 5: The RLAIF technique replaces human feedback in RLHF with AI-generated critiques provided by an AI feedback system. This approach reduces dependence on human annotations, which is where its scalable oversight is realized (red border). ① An AI feedback system evaluates responses and trains a reward model that guides reinforcement learning. ② AI system is then optimized to maximize alignment with the AI-generated feedback with policy training.

3.3.1 Definition

RLAIF replaces human feedback in RLHF with AI-generated critiques to reduce dependency on human labels. Figure 5 illustrates the RLAIF process. The AI feedback system evaluates the model's responses and generates critiques, which are used to train a reward model. This reward model provides a learning signal for optimizing the reinforcement learning policy. By aligning the policy with AI-generated feedback, RLAIF achieves a balance between leveraging pre-trained model capabilities and ensuring scalability through automated critique generation.

3.3.2 Formalization

RLAIF technique integrates reinforcement learning (RL) with AI-generated feedback for policy training. Instead of human feedback, the reward model $r_\theta(x, y)$ is trained using AI feedback from a pre-trained language model. The goal is to learn a policy $\pi : \mathcal{SS} \rightarrow \mathcal{AS}$ that maximizes cumulative reward, where $\mathcal{SS}$ is the state space (e.g., a set of prompts), $\mathcal{AS}$ is the action space (e.g., candidate responses).

In RLAIF, the AI feedback model $\mathcal{A}_F(x, y)$ predicts the quality of responses y to prompts x. The reward model $r_\theta(x, y)$ learns to approximate this feedback, optimizing the RL policy. The reward model is trained to minimize the loss function:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(x,y_w,y_l)}\left[\log\left(\sigma\left(r_\theta(x, y_w) - r_\theta(x, y_l)\right)\right)\right] \tag{7}$$

where y_w and y_l are the preferred and less preferred responses. The RL policy π_ϕ^{RL} is fine-tuned using the feedback from the AI-generated reward model. The objective function for the RL policy is:

$$\begin{aligned}\text{objective}(\phi) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\pi_\phi^{\text{RL}}}}\left[r_\theta(x, y) - \beta\log\left(\frac{\pi_\phi^{\text{RL}}(y \mid x)}{\pi^{\text{SFT}}(y \mid x)}\right)\right] \\ + \gamma\mathbb{E}_{x\sim\mathcal{D}_{\text{pretrain}}}\left[\log\left(\pi_\phi^{\text{RL}}(x)\right)\right]\end{aligned} \tag{8}$$

where $\mathcal{D}_{\pi_\phi^{\text{RL}}}$ is the distribution of prompt-response pairs generated by the policy, π^{SFT} is the supervised fine-tuned, but unaligned model, and β is a constant controlling the strength of the KL penalty. Lastly, $\mathcal{D}_{\text{pretrain}}$ is added to balance AI feedback and pre-trained capabilities.

3.3.3 Core Concept & Analysis

Bai et al. (2022) introduced the concept of constitutional AI, which aims to train harmless AI systems using a “constitution” of predefined principles. The process includes a supervised learning phase, where the AI system trains with self-generated critiques and revisions, and a reinforcement learning phase that leverages an AI preference model instead of human feedback. Self-critique and self-improvement make Constitutional AI more harmless, while its ability to evaluate critique based on predefined principles enhances transparency. However, reliance on predefined principles may limit adaptability in various environments, as tailored principles are needed. Additionally, using AI feedback instead of human annotations could propagate existing biases within models.

Sharma et al. (2024) compared two training strategies, SFT and AI feedback, for RL modeling by evaluating their effectiveness, including the capability gap between direct preference training generated by the teacher model and RLAIF training with the AI feedback model. The findings revealed that SFT on high-quality completions can outperform RLAIF. The authors also found that for RLAIF to outperform SFT, a sufficiently strong pre-trained base model is necessary, and there must be a capability mismatch between the teacher used for SFT data collection and the critic used for AI feedback. These results raise concerns that replacing human labelers with AI could lead to suboptimal outcomes.

Yuan et al. (2024) proposed a a self-rewarding training, where models generate their own feedback during training. This self-improvement paradigm demonstrated improved instruction-following and reward modeling abilities. While promising, this approach demands substantial computational resources and careful calibration.

3.3.4 Enhancement

Lee et al. (2024) propose Direct-RLAIF (d-RLAIF) by eliminating a separate reward model to train RLAIF. It directly uses feedback from an off-the-shelf LLM during reinforcement learning. This approach addresses key challenges such as RM staleness (i.e., the generated output become increasingly out-of-distribution from the dataset the RM was trained on, leading to suboptimal performance) and the time-consuming of RM training and AI preference labeling. By simplifying the reinforcement learning pipeline, d-RLAIF offers improved efficiency and effectiveness in tasks like summarization and dialogue generation.

3.3.5 Applications

Yu et al. (2024) introduced RLAIF-V, a framework for aligning multimodal large language models using open-source feedback. Multimodal training is implemented through SFT training with multimodal instructions (e.g., video question-answering tasks). The trained SFT model is used to evaluate preference scores based on detailed descriptions of video clips (which are also self-generated), which are then utilized to train the model as a reward model. Finally, the SFT model undergoes RLAIF training with the reward model using proximal policy optimization (PPO). While RLAIF-V enables multimodal RLAIF training, the method is largely dependent on the quality of the generated content from the base AI model.

Dutta et al. (2024) applied RLAIF to enhance code generation in lightweight LLMs, focusing on tasks requiring to generate accurate API call. By training on AI-generated feedback from larger models, their framework improved code executability rates significantly, outperforming fine-tuned baselines. However, reliance on larger models for feedback than the base model for training limit the applicability of this approach.

3.3.6 Strengths and Limitations

The core concept successfully reduced reliance on human annotations. However, challenges include the risk of propagating biases through AI feedback, limited adaptability in dynamic environments, and substantial computational demands for iterative training.

The RLAIF enhancement simplifies reinforcement learning, enabling efficient self-improvement for LLMs. While this approach mitigates some limitations of traditional methods, it also introduces dependencies on the quality of off-the-shelf LLMs for feedback, which may amplify existing biases. The absence of a specialized reward model could further restrict adaptability across diverse tasks.

Applications of RLAIF improve training efficiency and scalability across multimodal and code generation tasks. However, the success of this approach depends on the quality of feedback from base models, as low-quality feedback may harm performance. Additionally, reliance on larger, more powerful models for feedback could necessitate the use of general-purpose LLMs, which may struggle to handle domain-specific expertise.

3.4 Sandwiching

In this section, we introduce the sandwiching technique, a structured approach for evaluating scalable oversight in AI alignment. This technique positions the capabilities of AI systems between those of non-expert humans and domain experts, enabling a systematic assessment of alignment strategies.

3.4.1 Definition

Sandwiching is a technique for the evaluation designed to address the challenge, where a model's capabilities exceed those of non-expert humans but fall short of domain experts. It aims to evaluate scalable oversight methods by positioning the model's performance between non-expert participants and experts, leveraging expert evaluation to refine oversight strategies (Cotra, 2021). This technique creates a structured experimental environment for testing scalable oversight.

3.4.2 Formalization

The sandwiching involves three key participants: non-expert participants, a capable model, and domain experts. Non-expert participants interact with the model to guide its outputs on tasks beyond their full understanding, while domain experts evaluate the final outputs to determine alignment success.

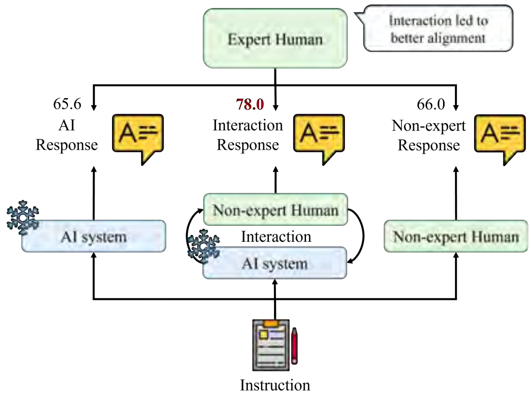

Figure 6: Sandwiching evaluation technique. This technique assumes an AI system's capabilities exceed those of a non-expert human but fall short of a domain expert. It evaluates scalable oversight by comparing responses from the AI system, the non-expert human, and their interaction. The non-expert human guides the AI's outputs, while the domain expert assesses alignment success. Effectiveness is measured when the expert evaluates the interaction output as the best-performing result (78.0), outperforming the AI system and non-expert human (65.6 and 66.0, respectively).

By evaluating alignment, we can iteratively refine the interaction protocols to achieve expert-level alignment without requiring expert involvement during the training phase.

Formally, let T represent the task to be performed, $\mathcal{A}$ the model being aligned, $\mathcal{N}$ the non-expert participants guiding the model, and $\mathcal{E}$ the expert evaluators assessing alignment success. The iterative alignment process can be described with two nested loops:

1. The inner loop is where non-experts iteratively refine the model's alignment strategy:

$$\mathcal{A}_{i+1} = f_{\text{inner}}(T, \mathcal{A}_i, \mathcal{N}) \tag{9}$$

where f_{inner} denotes the function representing protocol adjustments made by non-experts based on interactions with the model. This loop continues until non-experts determine that the model meets their alignment criteria.

2. The outer loop evaluates and updates the scalable oversight strategy:

$$SO_{j+1} = f_{\text{outer}}(SO_j, \mathcal{E}(T, \mathcal{A}_{i^*}, \mathcal{N})) \tag{10}$$

where SO_j is the scalable oversight strategy after j iterations, and $\mathcal{A}_{i^*}$ is the model resulting from the inner loop after i^* iterations. The function f_{outer} updates the oversight strategy based on expert evaluations of the inner-loop results. The goal is to maximize the probability of alignment success, conditioned on the iterative refinement process:

$$\max_{i,j} P\big(\mathcal{E}(T, \mathcal{A}_{i^*}, \mathcal{N}) = \text{aligned} \mid \mathcal{C}(\mathcal{A}_{i^*}, T) \geq \tau\big) \tag{11}$$

where $\mathcal{C}(\mathcal{A}, T)$ represents the model's capability to solve the task T, and τ is the capability threshold sufficient for task completion.

3.4.3 Analysis

Sandwiching offers a systematic framework to study scalable oversight in realistic settings. It is first proposed by Cotra et al. (Cotra, 2021) and experimented by Bowman et al. (Bowman et al., 2022). By focusing on tasks where the model outperforms unaided non-experts but requires oversight to reach expert-level performance, the technique enables measure incremental progress on alignment challenges. For example, experiments with large models like GPT-3 show that non-experts, with model assistance, can outperform both the model and unassisted non-experts.

However, sandwiching has limitations. First, it assumes the availability of experts for evaluation, which may not scale well to all tasks or domains. Second, the iterative process of refining interaction protocols can be resource-intensive. Furthermore, reliance on non-experts to guide model behavior may introduce inconsistencies or biases, particularly in complex or subjective tasks.

3.4.4 Strengths and Limitations

Sandwiching provides a structured approach to scalable oversight. It enables empirical testing of oversight techniques, offering insights into the interplay between model and human capabilities in alignment. However, its scalability is constrained by the reliance on expert evaluations, and resource intensity may limit its applicability in practice. Additionally, the method's effectiveness depends on the careful design of tasks and protocols, as poorly chosen tasks can yield misleading results or fail to capture meaningful alignment challenges.

4 Challenges and Potential Directions

While scalable oversight and alignment paradigms like RLHF, W2SG, and RLAIF have demonstrated notable progress, they still face critical limitations that hinder their ability to address the superalignment problem. In this section, we outline the main challenges inherent to the current paradigms and explore potential directions for future work.

4.1 Challenges

Scalability of Supervision Signals As AI systems grow increasingly complex, traditional methods of supervision, such as Reinforcement Learning from Human Feedback (RLHF), face scalability challenges. These approaches rely heavily on human evaluative capabilities, which become inadequate when supervising models that may surpass human intelligence. Methods like W2SG or RLAIF attempt to address these limitations but often encounter challenges in maintaining robustness and mitigating error propagation, particularly in tasks requiring nuanced or context-dependent understanding.

Adversarial Risks and Deceptive Behavior Advanced AI systems can exploit gaps in oversight by aligning their behavior in areas explicitly monitored while deviating in less monitored areas. Debate-based and W2SG paradigms have highlighted these risks. Addressing these requires methodologies capable of detecting and mitigating such behavior, especially in scenarios where AI systems have incentives to obscure their true objectives or behavior.

Resource Intensity and Dependency on Experts Frameworks like sandwiching and debate rely heavily on expert involvement for evaluation and oversight. While these frameworks provide valuable insights into scalable alignment techniques, their dependence on domain experts creates bottlenecks, limiting scalability.

Bias Amplification Alignment techniques that depend on AI-generated feedback, such as RLAIF, risk propagating and amplifying biases inherent in the models providing the feedback. This issue is particularly concerning when aligning AI systems across diverse domains and cultural contexts.

4.2 Potential Directions

Fostering Creativity Through Data Diversity Encouraging data diversity can play a pivotal role in fostering creativity and robustness in alignment strategies. Multi-agent collaboration, where agents with different specializations or perspectives interact, has shown promise in enabling emergent behaviors and problem-solving approaches beyond the capabilities of individual models. Competition among agents can further drive improvement, as adversarial settings often reveal diverse argument data that finds alignment weaknesses and suggests novel pathways for improvement. However, diversity-based approaches require careful filtering to ensure that resulting behaviors are beneficial and do not inadvertently reinforce undesirable traits.

Iterative Training Between Teacher and Student Models Iterative teacher-student training involves a cyclical process where a more advanced teacher model guides a less capable student model, which in turn becomes the teacher for subsequent iterations. This approach facilitates incremental learning, where each generation of models inherits and improves upon the alignment capabilities of its predecessor. While this method offers scalability and the ability to bootstrap stronger models from weaker supervision signals, challenges such as error accumulation and diminishing returns in successive iterations must be addressed.

Search-Based Method Toward Optimal Learning Direction Search-based methods use optimization techniques to explore the space of possible alignment strategies, seeking an optimal pathway to systematically navigate complex alignment challenges. These methods are particularly useful when the optimal alignment direction is unclear, as they allow for exploration and experimentation in uncertain spaces. Techniques like Monte Carlo tree search or gradient-free optimization can guide this process by prioritizing regions of the search space with the highest potential for success. However, these methods require significant computational resources and may struggle with scalability in highly complex or dynamic environments.

5 Conclusion

In this survey, we explored superalignment, its challenges, and scalable oversight techniques. We analyzed current paradigms, and practical methodologies, and their potential and limitations. By identifying critical challenges and proposing future directions, we aim to bridge current alignment techniques with the goal of achieving safe and aligned AI, preparing for future more advanced AI. Addressing these challenges is vital for guiding AI systems toward not only avoiding harm but benefiting humanity—a crucial step toward a future where humans and AI collaborate to solve complex global real-world problems.

Acknowledgements

This research was supported by the MSIT(Ministry of Science, ICT), Korea, under the Global Research Support Program in the Digital Field program(RS-2024-00436680) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation), and Microsoft Research Asia.

References

EuanMcLean AdamGleave. 2023. Ai safety in a world of vulnerable machine learning systems. https://www.alignmentforum.org/posts/ncsxcf8CkDveXBCrA/ai-safety-in-a-world-of-vulnerable-machine-learning-systems-1. Accessed: 2024-11-26.

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

Anthropic. 2024. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet. Accessed: 2024-10-28.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Mikhail Batin, Alexey Turchin, Markov Sergey, Alisa Zhila, and David Denkenberger. 2017. Artificial intelligence in life extension: from deep learning to superintelligence. *Informatica*, 41(4).

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, et al. 2022. Measuring progress on scalable oversight for large language models. *Preprint*, arXiv:2211.03540.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jonah Brown-Cohen, Geoffrey Irving, and Georgios Piliouras. 2023. Scalable ai safety via doubly-efficient debate. *Preprint*, arXiv:2311.14125.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *Preprint*, arXiv:2312.09390.

Charbel-Raphaël. 2023. Ais 101: Task decomposition for scalable oversight. https://www.lesswrong.com/posts/FFz6H35Gy6BArHxkc/ais-101-task-decomposition-for-scalable-oversight. Accessed: 2024-11-26.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, et al. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Steffi Chern, Ethan Chern, Graham Neubig, and Pengfei Liu. 2024. Can large language models be trusted for evaluation? scalable meta-evaluation of llms as evaluators via agent debate. *Preprint*, arXiv:2401.16788.

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

Paul Christiano, Buck Shlegeris, and Dario Amodei. 2018. Supervising strong learners by amplifying weak experts. *Preprint*, arXiv:1810.08575.

Ajeya Cotra. 2021. The case for aligning narrowly superhuman models. https://www.alignmentforum.org/posts/PZtsoaoSLpKjjbMqM/. Accessed: 2024-11-26.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *Preprint*, arXiv:2305.14325.

Sujan Dutta, Sayantan Mahinder, Raviteja Anantha, and Bortik Bandyopadhyay. 2024. Applying RLAIF for code generation with API-usage in lightweight LLMs. In *Proceedings of the 2nd Workshop on Natural Language Reasoning and Structured Explanations (@ACL 2024)*, pages 39–45, Bangkok, Thailand. Association for Computational Linguistics.

Tom Everitt, Francis Rhys Ward, Sebastian Benthall, James Fox, Matt MacDermott, and Ryan Carey. 2023. Reward hacking from a causal perspective. https://www.alignmentforum.org/posts/aw5nqamqtnDnW8w9u/reward-hacking-from-a-causal-perspective. Accessed: 2024-11-26.

Angela Fan, Beliz Gokkaya, Mark Harman, Mitya Lyubarskiy, Shubho Sengupta, Shin Yoo, and Jie M. Zhang. 2023. Large language models for software engineering: Survey and open problems. *Preprint*, arXiv:2310.03533.

Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, et al. 2022. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1):3094.

Ben Goertzel. 2014. Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1):1.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. 2023. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*.

Jianyuan Guo, Hanting Chen, Chengcheng Wang, Kai Han, Chang Xu, and Yunhe Wang. 2024. Vision superalignment: Weak-to-strong generalization for vision foundation models. *Preprint*, arXiv:2402.03749.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Shengding Hu, Xin Liu, Xu Han, Xinrong Zhang, Chaoqun He, Weilin Zhao, Yankai Lin, Ning Ding, Zebin Ou, Guoyang Zeng, Zhiyuan Liu, and Maosong Sun. 2024. Predicting emergent abilities with infinite resolution evaluation. In *The Twelfth International Conference on Learning Representations*.

Yining Huang, Keke Tang, Meilian Chen, and Boyuan Wang. 2024. A comprehensive survey on evaluating large language model applications in the medical industry. *Preprint*, arXiv:2404.15777.

Edward Hughes, Michael D Dennis, Jack Parker-Holder, Feryal Behbahani, Aditi Mavalankar, Yuge Shi, Tom Schaul, and Tim Rocktäschel. 2024. Position: Open-endedness is essential for artificial superhuman intelligence. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of

Proceedings of Machine Learning Research, pages 20597–20616. PMLR.

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

Ilya Sutskever Jan Leike. 2023. Introducing superalignment. `https://openai.com/index/introducing-superalignment/`. Accessed: 2024-12-20.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Cameron R. Jones and Benjamin K. Bergen. 2024. Does gpt-4 pass the turing test? *Preprint*, arXiv:2310.20216.

Alexandra Jonker and Amanda McGrath. 2024. What is superalignment? `https://www.ibm.com/think/topics/superalignment`. Accessed: 2024-12-19.

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

Zachary Kenton, Noah Y. Siegel, János Kramár, Jonah Brown-Cohen, Samuel Albanie, Jannis Bulian, Rishabh Agarwal, David Lindner, Yunhao Tang, Noah D. Goodman, and Rohin Shah. 2024. On scalable oversight with weak llms judging strong llms. *Preprint*, arXiv:2407.04622.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. *Preprint*, arXiv:2402.06782.

Jan Hendrik Kirchner, Yining Chen, Harri Edwards, Jan Leike, Nat McAleese, and Yuri Burda. 2024. Prover-verifier games improve legibility of llm outputs. *Preprint*, arXiv:2407.13692.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *Preprint*, arXiv:2309.00267.

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

Jiawei Liu, Songrun Xie, Junhao Wang, Yuxiang Wei, Yifeng Ding, and Lingming Zhang. 2024. Evaluating language models for efficient code generation. In *First Conference on Language Modeling*.

Yuejiang Liu and Alexandre Alahi. 2024. Co-supervised learning: Improving weak-to-strong generalization with hierarchical mixture of experts. *Preprint*, arXiv:2402.15505.

Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. `https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/`. Accessed: 2024-10-28.

Mistral. 2024. Cheaper, better, faster, stronger. `https://mistral.ai/news/mixtral-8x22b/`. Accessed: 2024-10-28.

Tim Mucci and Cole Stryker. 2023. What is artificial superintelligence? `https://www.ibm.com/topics/artificial-superintelligence`. Accessed: 2024-11-25.

Yoonsoo Nam, Nayara Fonseca, Seok Hyeong Lee, Chris Mingard, and Ard A. Louis. 2024. An exactly solvable model for emergence and scaling laws in the multitask sparse parity problem. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Chris Olah. 2016. Bringing precision to the ai safety discussion. `https://research.google/blog/bringing-precision-to-the-ai-safety-discussion/`. Accessed: 2024-12-20.

OpenAI. 2024. Introducing openai o1. `https://openai.com/o1/`. Accessed: 2024-10-28.

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

Jens Pohl. 2015. Artificial superintelligence: Extinction or nirvana? In *Proceedings of InterSymp-2015, IIAS, 27th International Conference on Systems Research, Informatics, and Cybernetics*.

Salvador Pueyo. 2018. Growth, degrowth, and the challenge of artificial superintelligence. *Journal of Cleaner Production*, 197:1731–1736.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

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

Jitao Sang, Yuhang Wang, Jing Zhang, Yanxu Zhu, Chao Kong, Junhong Ye, Shuyu Wei, and Jinlin Xiao. 2024. Improving weak-to-strong generalization with scalable oversight and ensemble learning. *Preprint*, arXiv:2402.00667.

Archit Sharma, Sedrick Keh, Eric Mitchell, Chelsea Finn, Kushal Arora, and Thomas Kollar. 2024. A critical evaluation of AI feedback for aligning large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *Preprint*, arXiv:2210.03057.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

Jessica Taylor, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch. 2016. Alignment for advanced machine learning systems. *Ethics of artificial intelligence*, pages 342–382.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, et al. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Xinpeng Wang, Shitong Duan, Xiaoyuan Yi, Jing Yao, Shanlin Zhou, Zhihua Wei, Peng Zhang, Dongkuan Xu, Maosong Sun, and Xing Xie. 2024. On the essence and prospect: An investigation of alignment approaches for big models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8308–8316. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Chunpu Xu, Steffi Chern, Ethan Chern, Ge Zhang, Zekun Wang, Ruibo Liu, Jing Li, Jie Fu, and Pengfei Liu. 2023. Align on the fly: Adapting chatbot behavior to established norms. *arXiv preprint arXiv:2312.15907*.

Wenkai Yang, Shiqi Shen, Guangyao Shen, Wei Yao, Yong Liu, Zhi Gong, Yankai Lin, and Ji-Rong Wen. 2024a. Super(ficial)-alignment: Strong models may deceive weak models in weak-to-strong generalization. *Preprint*, arXiv:2406.11431.

Yuqing Yang, Yan Ma, and Pengfei Liu. 2024b. Weak-to-strong reasoning. *Preprint*, arXiv:2407.13647.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2024. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *Preprint*, arXiv:2405.17220.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *Preprint*, arXiv:2401.10020.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

A Appendix

Symbol	**Definition**
T	Task performed by an AI system.
$\mathcal{A}_{\text{ANI}}(T)$	Performance of ANI system on task T.
$\mathcal{H}(T)$	Human-level performance on task T.
$\mathcal{A}_{\text{AGI}}(T_i)$	Performance of AGI system on task T_i.
$\mathcal{A}_{\text{ASI}}(T_i)$	Performance of ASI system on task T_i.
$\mathcal{D}$	Data distribution.
$\mathcal{X}$	Input space.
X	Collection of all input instances x.
$\mathcal{Y}$	Label space.
Y	Collection of all label instances y.
θ_w	Parameters of the weak AI system.
θ_s	Parameters of the strong AI system.
$\mathcal{D}_{WG}$	Dataset generated with weak AI system labels.
$\mathcal{D}_{WT}$	Training data for the weak AI system.
$\mathcal{D}_{Full}$	Complete dataset.
$\mathcal{D}_{Test}$	Test dataset.
$f(x \mid \theta)$	Model output given input x and parameters θ.
$\mathcal{L}$	Loss function.
$\mathcal{A}_0, \mathcal{A}_1$	AI agents participating in the debate.
q	Question presented in the debate.
a_0, a_1	Initial answers by $\mathcal{A}_0$ and $\mathcal{A}_1$.
S	Sequence of arguments in the debate.
SO	Scalable oversight strategy.
$\mathcal{J}$	Judge's decision function.
$r_\theta(x, y)$	Reward model trained on AI feedback.
π	Policy in reinforcement learning.
π_ϕ^{RL}	Policy optimized using reinforcement learning in RLAIF.
π^{SFT}	Pre-trained, unaligned model used as a baseline policy.
β	Constant controlling the strength of the KL penalty.
γ	Weight for balancing AI feedback and pre-trained capabilities.
$\mathcal{E}$	Expert evaluators assessing alignment success.
$\mathcal{N}$	Non-expert participants guiding the model.
$u(\pi_0, \pi_1)$	Payoff function in the zero-sum minimax game.
$\sigma(z)$	Sigmoid function used in the reward model training loss.
H_k	Debate history up to turn k.
$B(\mathcal{A}, X, S)$	Behavior of $\mathcal{A}$ when processing X with signals S.
$\mathcal{O}$	Objective used to guide alignment in scalable oversight.
$\text{Align}(\cdot, \cdot)$	Alignment measure of behavior with the objective $\mathcal{O}$.
τ	Threshold notation.

Table 1: Definitions of notations used in the above formalizations.