SCIENCE CHINA
Information Sciences

• POSITION PAPER •

June 2025, Vol. 68, Iss. 6, 160107:1–160107:11
https://doi.org/10.1007/s11432-024-4348-6

Special Topic: Enabling Techniques and Cutting-Edge Applications of Foundation Models

The superalignment of superhuman intelligence with large language models

Minlie HUANG[1*], Yingkang WANG[1], Shiyao CUI[1], Pei KE[2] & Jie TANG[3]

[1] *The CoAI Group, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*
[2] *Laboratory of Intelligent Collaborative Computing, University of Electronic Science and Technology of China, Chengdu 611731, China*
[3] *Knowledge Engineering Group, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*

Received 3 December 2024/Revised 7 February 2025/Accepted 15 March 2025/Published online 20 May 2025

Abstract We have witnessed the emergence of superhuman intelligence thanks to the fast development of large language models (LLMs) and multimodal language models. As the application of such superhuman models becomes increasingly popular, a critical question arises: how can we ensure they still remain safe, reliable, and aligned well with human values encompassing moral values, Schwartz's Values, ethics, and many more? In this position paper, we discuss the concept of superalignment from a learning perspective to answer this question by outlining the learning paradigm shift from large-scale pretraining and supervised fine-tuning, to alignment training. We define superalignment as designing effective and efficient alignment algorithms to learn from noisy-labeled data (point-wise samples or pair-wise preference data) in a scalable way when the task is very complex for human experts to annotate and when the model is stronger than human experts. We highlight some key research problems in superalignment, namely, weak-to-strong generalization, scalable oversight, and evaluation. We then present a conceptual framework for superalignment, which comprises three modules: an attacker which generates the adversary queries trying to expose the weaknesses of a learner model, a learner which refines itself by learning from scalable feedbacks generated by a critic model with minimal human experts, and a critic which generates critics or explanations for a given query-response pair, with a target of improving the learner by criticizing. We discuss some important research problems in each component of this framework and highlight some interesting research ideas that are closely related to our proposed framework, for instance, self-alignment, self-play, self-refinement, and more. Last, we highlight some future research directions for superalignment, including the identification of new emergent risks and multi-dimensional alignment.

Keywords superalignment, superhuman intelligence, large language models, scalable feedback, weak-to-strong generalization

Citation Huang M L, Wang Y K, Cui S Y, et al. The superalignment of superhuman intelligence with large language models. Sci China Inf Sci, 2025, 68(6): 160107, https://doi.org/10.1007/s11432-024-4348-6

1 Introduction

The fast development of generative artificial intelligence (AI), typically known as large language models (LLMs) or multimodal language models (MLMs)[1)], has drawn significant attention due to its emerging ability to tackle a large variety of complex tasks, including mathematics, reasoning, coding, visual understanding and generation, and social tasks [1,2]. These models have shown unbelievable competence and human-level or even beyond-human-level performance on many benchmarks. This progress has fueled discussions about the concept of superhuman intelligence or artificial general intelligence (AGI) whose definition has not been widely accepted. To name a few definitions, OpenAI defines AGI as highly autonomous systems that outperform humans at most economically valuable work [3]. While Marcus [4], a cognitive scientist from New York University, defines AGI as any intelligence that is flexible and general, with resourcefulness and reliability comparable to or beyond human intelligence. However, there are also debating opinions, as LeCun [5,6] said, “Human intelligence is NOT general”, what we are discussing is actually advanced machine intelligence (AMI). As a pivotal milestone in artificial intelligence research,

[*] Corresponding author (email: aihuang@tsinghua.edu.cn)
[1)] In this paper, we only focus on LLMs, but many of our claims are also applicable to MLMs.

AGI aspires to emulate human-like cognitive versatility, enabling it to reason, make decisions, and solve problems in dynamic, unpredictable environments, with very generalizable manners.

Along with their tremendous capabilities, these superhuman models also raise critical ethical, safety, and governance concerns which may pose severe threats to human society [7]. In particular, highly intelligent models possess a greater capacity for autonomous decision-making, making them harder to predict and control. This raises significant concerns about unintended behaviors, especially in high-stakes applications such as in finance, healthcare, and critical infrastructures [8]. Given that unaligned LLMs could pose significant risks to humanity [9], substantial efforts have been made to align them with human values[2]. This is achieved through learning from human feedback using alignment algorithms such as proximal policy optimization (PPO) [10], direct preference optimization (DPO) [11], efficient and exact alignment optimization (EXO) [12], and many more. After being pretrained on large-scale corpora, these models are aligned using well-curated human preference data which reflect human values, social norms, and ethical considerations, either implicitly or explicitly.

Since the capabilities of LLMs have grown very fast and we have already witnessed superhuman intelligence in many tasks, two critical questions arise: (1) how can we ensure these systems remain safe and aligned well with human values? and (2) how can we control the behaviors of such superhuman systems? This concern grows even more serious with the potential development of superhuman intelligence, namely, AI systems that exceed human intelligence in nearly all domains [13] and maintain the ability to self-evolve its abilities automatically. In this setting, vanilla alignment techniques relying on human feedback will not be applicable anymore since the tasks have become increasingly complex; in addition, the involved systems are even more intelligent than humans so that even human experts cannot provide scalable and reliable supervision during the learning of superhuman AI systems. In other words, traditional alignment algorithms and human supervision cannot scale further when (1) the involved task becomes extremely difficult (e.g., Olympic competition-level coding), and (2) the system intelligence is beyond even human expert intelligence.

Therefore, ensuring the safety of superhuman models requires superalignment[3], which aims to automatically align these superhuman models with human values in a scalable, reliable, and generalizable manner. Superalignment enables alignments through self-refinement or self-play driven by interactions and collaboration among AI models. Unlike traditional alignment methods, humans in superalignment only play a minimal role in assisting the automatic alignment process, where the alignment is realized through a "human-in-the-loop" paradigm: the superhuman model is learned from automatically scalable feedbacks and human experts only provide supervision on a small proportion of cases.

This paper is structured as follows. In Section 2, we define superalignment from a machine learning perspective, addressing the key learning paradigms of large-scale pretraining, supervised fine-tuning, and LLM alignment. In Section 3, we highlight some key research problems in superalignment, including weak-to-strong generalization, scalable oversight, and evaluation. In Section 4, we present a feasible framework for superalignment, consisting of three core modules, namely attacker, learner, and critic, and discuss critical research issues within each module and some interesting attempts at this framework. Finally, in Section 5, we summarize this paper and also highlight some important future directions.

2 Definition of superalignment from the learning perspective

In this section, we will formally introduce the concept of superalignment. We will start from the learning paradigm of large-scale pretraining, then introduce classical alignment algorithms of large language models, and finally describe the meanings of superalignment from the machine learning perspective. The learning processes of a powerful LLM fall into three major steps: pretraining from trillions of unlabeled data, supervised fine-tuning on human-curated query-response pairs, and alignment from human preference data.

[2] Kindly note that "human value" is a very broad, general, and vague concept in LLM alignment. There is no clear, widely accepted definition yet. In this paper, human values encompass safety, moral values, Schwartz's values, ethics, and many more.

[3] The term superalignment was first introduced by OpenAI; however, in this paper, we provide a precise definition of what it entails.

2.1 Learning paradigm of large-scale pretraining

During pretraining, a model learns a generation distribution P_θ from large-scale text corpora $\mathcal{D}$ sampled from an unknown, underlying data distribution P_{data}, which is well-known as the next-token-prediction learning paradigm:

$$\mathcal{L}_{\text{pretraining}} = -\mathbb{E}_{x \sim P_{\text{data}}(x)} \sum_{i=1}^{T} \log P_\theta(x_i | x_{<i}), \tag{1}$$

where each x means a text segment consisting of T tokens, each x_i denotes a token, and $x_{<i}$ indicates the preceding context of x_i. Given a huge amount of text, the model will learn the generation distribution $P_\theta(x)$ in an unsupervised way. However, next-token-prediction can date back to 2014 since neural generation models [14] have been used for machine translation or other sequence-to-sequence transformation tasks. In such a framework, the model tries to translate a source sequence x to a target sequence y by generating target tokens in an autoregressive way, as follows:

$$\mathcal{L}_{\text{sft}} = -\mathbb{E}_{(x,y) \sim P_{\text{data}}(x,y)} \sum_{i=1}^{T} \log P_\theta(y_i | y_{<i}, x). \tag{2}$$

The model is trained on a corpus of (x, y) pairs, where the supervision signal is derived from the target sequence y, either constructed by human annotation or automatically from unsupervised data.

2.2 Alignment training of large-scale language models

A pre-trained model can demonstrate surprisingly good cross-task, few shot generalization performance, however, it is still not sufficient for generating results that are well aligned with human values. Therefore, alignment training is crucial for improvement, where the model will be further trained on a dataset consisting of high-quality human-curated (x, y) pairs where human values are implicitly or explicitly embedded in the data. The training objective is the same as that in (2), where this process is usually named supervised fine-tuning (SFT). The construction of the data pairs normally considers human values such as safety issues, social norms, and ethical concerns.

During supervised fine-tuning, we only teach the model to learn what is a good generation, namely, negative examples are not used for learning. However, in the human learning process, we are always learning from both positive and negative examples. Thus, we can learn from paired preference data by constructing data triples $D = \{(x, y_w, y_l)\}$, where for a given input x, a winning response y_w with higher quality and a loss response y_l with lower quality are built. On top of such data triples, we can first learn a reward function $r(x, y)$ which rates how well a response y can respond to an input query, and then apply some alignment algorithm to learn from such preference data.

The most popular and effective alignment algorithm is reinforcement learning from human feedback with PPO [10]. The learning objective is presented as follows:

$$\mathcal{L}_{\text{RLHF}} = -\mathbb{E}_{x \sim P_{\text{data}}(x,y)} \left[\mathbb{E}_{y \sim P_\theta(y|x)}[r(x, y)] - \beta \mathbb{D}_{\text{KL}} \left(P_\theta(y \mid x) \| P_{\text{sft}}(y \mid x) \right) \right], \tag{3}$$

where P_{sft} is the generation distribution obtained via supervised fine-tuning, P_θ is the distribution to be optimized during alignment, $\mathbb{D}_{\text{KL}}(p||q)$ is the KL divergence between two distributions p and q, and β is a hyperparameter weighting the regularization term.

The PPO algorithm has been shown very effective and widely used in aligning a pretrained LLM. However, it becomes very slow since it requires online sampling during the training process when the model size and training data are large. Thus, several methods are proposed to stabilize and accelerate the training process by avoiding reinforcement learning. For example, DPO [11] extracts the optimal policy from the standard reinforcement learning from human feedback (RLHF) objective in a closed form, thereby solving RLHF with a simple classification loss:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y_w,y_l) \sim P_{\text{data}}(x,y_w,y_l)} \left[\log \sigma \left(\beta \log \frac{P_\theta(y_w|x)}{P_{\text{ref}}(y_w|x)} - \beta \log \frac{P_\theta(y_l|x)}{P_{\text{ref}}(y_l|x)} \right) \right], \tag{4}$$

where $P_{\text{ref}}(\cdot|\cdot)$ is usually a generation distribution obtained via supervised fine-tuning.

Essentially, DPO is a maximum likelihood estimation method, where the learning objective tries to increase the likelihood of observing the winning response and yet decrease that of observing the loss response [11]. Due to its simplicity and effectiveness, DPO has become popular and many variants have been proposed, which mainly fall into three types: first, leverage preference from single human reference [15]; second, change the preference data distribution using rejection sampling [16], or extend pair-wise preference to ranking preference data [17]; third, modify the learning objective such as maximizing human utility based on prospect theory [18] or substituting the point-wise reward with a pair-wise preference function [19].

2.3 Superalignment of large-scale language models

In alignment training of LLMs, there are underlying assumptions that are usually neglected. As shown in (2) (next-token-prediction), we actually assume that the next token is a golden target without any noise during the supervised fine-tuning phase. In (3), we implicitly assume that the reward model for rating a query-response pair (x, y) is learned from perfect human preference data and we can learn a perfect reward function. However, in superalignment, these assumptions do not hold any more because of two facts: first, the task itself becomes very complex such that even human experts cannot provide reliable annotations, thereby leading to noisy human labels; second, the model becomes super intelligent and is even smarter than our humans, and thus human experts cannot identify the flaws of a generated response, or reliably distinguish the quality difference between two responses. In other words, during superalignment, we only have noisy labels/annotations for training a superhuman model during both supervised fine-tuning (e.g., as in (2)) and learning from human feedback (e.g., as in (3)).

Now, let us come to the definition of superalignment from the learning perspective: superalignment is about designing effective and efficient alignment algorithms to learn from noisy-labeled data (point-wise samples or pair-wise preference data) in a scalable way when the task becomes very complex for human experts to annotate and the model is stronger than human experts. The superalignment setting raises some fundamental research problems which will be detailed in Section 3.

3 Key research problems in superalignment

There are fundamental research problems in superalignment. These problems are closely related to answering these questions: how can we continuously improve a superhuman model that is even more intelligent than our humans, and how can we ensure the superhuman model is still controllable, safe, and well-aligned with human values?

More specifically, we will discuss the below research problems in Subsections 3.1–3.3:

- **Weak-to-strong generalization.** How to align and improve strong models with weak supervisors? In this setting, a stronger model is supervised by a weaker model or a human (weaker than the strong model in superalignment) but we are seeking to align and further improve the stronger model.
- **Scalable oversight.** How to provide scalable and reliable supervision signals to train strong models from human or AI models when the task is overly complex or even human experts cannot make reliable annotations.
- **Evaluation.** How to validate the alignment of superhuman models by automatically searching for problematic behaviors and problematic internals, and how to conduct adversarial tests automatically to expose the weaknesses of strong models?

3.1 Weak-to-strong generalization

Weak-to-strong generalization aims to optimize a stronger model continuously using a weaker supervisor, which was first introduced by OpenAI [20]. In traditional machine learning tasks, a target model (to be optimized) is weaker than the supervisor which is usually human, or a stronger model (well known as knowledge distillation). However, in superalignment, the supervisor is even weaker than the superhuman model to be optimized, which poses new challenges to further improve the superhuman model.

OpenAI made an analogy to this setting [20]. They supervised GPT-4 with a GPT-2-level model on NLP tasks, and found that the resulting model typically performs somewhere between GPT-3 and GPT-3.5. In this manner, they were able to recover much of GPT-4's capabilities with only much weaker supervision. This research manifests that a strong model can generalize beyond weak supervision, solving even hard problems for which the weak supervisor can only give incomplete or flawed training labels.

There are some important research sub-problems in weak-to-strong generalization. Since OpenAI's study is still very preliminary, there is yet much space to explore in this direction. First, since the supervisor is weaker, which information will be useful for supervising the stronger model and how to identify such information? Second, since the supervision signal is noisy, how can the stronger model learn robustly from noise samples? This problem has been studied extensively in machine learning communities, however, it becomes much more complex in the setting of LLMs as the noises may be imposed at the token, span, or response level and the generative learning problem is more difficult than simple classification or regression problems. Third, since in general purposing a stronger model is very hard to learn from weaker supervision, can we assemble multiple specialized weaker models to supervise the learning of a stronger model?

3.2 Scalable oversight

Scalable oversight aims to empower relatively weak overseers to deliver reliable supervision, including training labels, reward signals, or feedback [21], for complex tasks. As superalignment needs to tackle extremely complex and highly intelligent AI models, scalable oversight can provide a technical road to overcome the limitations of human supervision, providing reliable oversight of great quality. There are two feasible paths towards providing scalable oversights: one is to use powerful models to provide scalable feedbacks and the other is to assist human annotators with strong critic models so that humans can easily provide supervision on complex tasks.

Existing proposals for scalable oversight mainly fall into three types. The first type is about decomposition. Task decomposition is a representative paradigm to provide scalable oversight, where the complex task is decomposed into a series of relatively simpler subtasks that can be more easily handled. In this manner, complex tasks can be more easily supervised and annotated. For instance, iterated amplification [22] constructs training signals iteratively by integrating solutions to simpler subtasks. Wen et al. [23] demonstrated that competition-level code generation can be solved more efficiently by decomposing a complex program into sub-functions, which they called human-centric decomposition. Similarly, recursive reward modeling [24] enhances AI models by progressively supervising them using reward models that are iteratively refined through improved human feedback. The second type utilizes a powerful model to generate feedback, critiques, and labels in accordance with human-designed principles to acquire scalable oversight [25]. Anthropic applies this approach during the reinforcement learning (RL) phase with a trained preference model to provide rewards [26], marking a shift from RLHF to "reinforcement learning from AI feedback" (RLAIF). In the third type, scalable oversight can be achieved via debate between multiple AI agents to determine the best answer to a given question [27,28]. During the process, humans play a minimal role by providing the necessary rules to guide the debate and acting as the final arbiter to select the most appropriate response.

Despite the efforts above, there are key research problems unsolved in scalable oversight. First, can we build a universal model to provide critics or feedbacks in a scalable and generalizable manner, which studies for all tasks and settings? Though GPT-4 has shown very general critic ability for all types of tasks, how such ability is acquired is still unclear. Second, how can human experts be assisted by a copilot model (e.g., CriticGPT [29]) to provide reliable feedback or annotation for extremely challenging tasks? Third, how can human and AI models collaborate together to provide scalable feedback for superalignment?

3.3 Evaluation

The evaluation aims to measure the alignment of superhuman models accurately from different dimensions and automatically reveal the weaknesses of superhuman models. Although evaluation has been a long-standing research problem in NLP, existing evaluation metrics cannot reflect the quality of generated texts from superhuman models since their performance has surpassed humans, which poses severe challenges to these important constituents of superalignment.

Existing studies on the evaluation for alignment of AI models fall into three categories. (1) Benchmarks: Most of the existing benchmark datasets aim to measure specific abilities of LLMs on fixed benchmark datasets, including math [30], reasoning [31], code generation [32], and instruction following [33]. However, these static benchmark datasets face severe challenges in data pollution, thereby causing over-estimated performance especially on subsequent LLMs that may use similar data as training data. Thus, considering the evaluation of the superalignment of AI models, the benchmark dataset should be constructed dynamically and updated quickly by including high-quality and diverse samples which can consistently reveal the weaknesses of fast-growing superhuman models. (2) LLM-based evaluation method: Existing studies mostly utilize the ability of current LLMs to measure the generation quality [33]. Specifically, they formulate evaluation as an instruction-following QA task, and use LLMs to generate both evaluation scores and explanations via elaborate prompt design [33]. The ability to generate evaluation results in an unsupervised manner may come from the pre-training data which are similar to the evaluation task such as comments and reviews. To automatically evaluate the alignment of superhuman models, it is important to trace the root of the evaluation ability of AI models and thus fully stimulate this ability for generation quality assessment. (3) Critic model: To achieve superalignment of AI models, it is important to construct a universal critic model which can efficiently provide evaluation results in a large variety of tasks and settings [29, 34]. Such a critic model can provide scalable feedback to further improve AI models in various tasks, thereby assisting the superalignment of AI models. Existing studies have also connected critique generation with reward models [35], which indicates a promising way to collect high-quality reward signals for guiding superhuman models toward stronger generation capabilities.

Despite the rapid development of evaluation, there still exist some essential research problems towards superalignment. First, how to automatically construct adversarial datasets to expose the weaknesses of superhuman models? This problem is under-explored because most of the existing benchmarks are restricted to human-crafted task taxonomies, thereby only revealing the weaknesses in these tasks. Some preliminary studies have shown that well-designed pipelines based on state-of-the-art LLMs (such as GPT-4) can automatically find the weaknesses in LLMs [36]. Second, how to validate the evaluation results of superhuman models? Since human references may not work for judging the evaluation results of superhuman models, it is important to avoid the misleading evaluation results (like reward hacking [37] in RLHF) causing misaligned with human values [38]. Finally, today's evaluation is heavily reliable on static evaluation (i.e., results on benchmarks), but how can we design auto-evaluation methods for superhuman models and how can we conduct adversary tests automatically?

4 Framework to realize superalignment

In this section, we will present a feasible framework to realize superalignment, as presented in Figure 1. There are three modules in this framework: an attacker model, which simulates attacks and generates adversary queries such that a learner model may fail to produce high-quality responses; a learner model which will be continuously improved by learning from scalable feedbacks generated from a critic model or human feedback whenever human intervention is required; and a critic model which generates explanations, feedbacks, or reasons given a query from the attacker and a response from the learner as input. This pipeline can be automatically executed and iterated when it is started from some seed input. Noticeably, the attacker, learner, and critic can be the same foundation model but with different versions.

This is a conceptual framework, which leaves many questions unsolved in implementation. In general purpose, it is very difficult to make the pipeline work smoothly, however, we believe in some specific cases, for instance, mathematic tasks and code generation, this framework is feasible and there are already some research attempts as shown in [36, 39]. In what follows, we will discuss the key challenges and research problems in this framework.

4.1 Attacker: discovering the weaknesses of LLMs automatically

The attack model aims to generate adversary queries that the learner may fail to answer. In this manner, the weaknesses of the learner model can be automatically exposed, and then these weaknesses can be fixed accordingly. Such adversary attacks have been largely studied (known as red teaming methods) in safety issues of generative models [40], image classification [41], or image generation in diffusion models [42].

However, building such an attack model has never been easy. One straightforward way is to use prompt engineering which designs some prompt templates to trigger a model to generate adversary attacks. Unfortunately, this method is sensitive to pre-specified prompts, largely depends on the base capability of the attack model, and may fail in some cases such as LLM's safety since many LLMs have been trained not to generate harmful queries. Another way is to train an attack model to simulate adversary attacks by constructing adversary training data. This can be enhanced by reinforcement learning. For instance, in the context of LLM's safety, Wen et al. [43] presented an RL method for generating implicitly toxic contents with a reward function, which encourages the model to generate subtle, implicitly toxic contents.

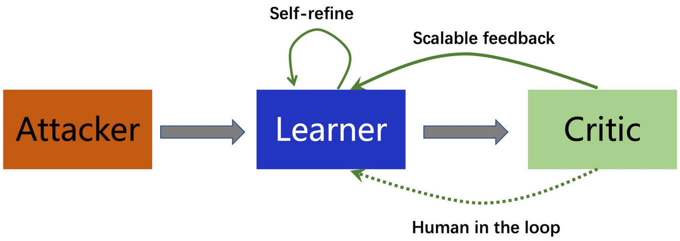

Figure 1 (Color online) Conceptual framework for superalignment. The attacker generates adversary queries so that the learner may fail to produce high-quality responses; the learner will be continuously improved by learning from scalable feedbacks generated from critic or from minimal human feedbacks whenever necessary; and the critic generates explanations, feedbacks, or reasons given a query from the attacker and a response from the learner as input. Starting from some seed input, the pipeline can be automatically iterated.

By this means, the generated contents have very high attack success rates to common toxicity classifiers. In general purpose, we can train the attack model with reinforcement learning, using the reward signal from a critic model, while the objective is to encourage the attack model to generate queries that lead to down-rated responses from a strong model (the learner). Besides, the attacker could also red-team LLM flaws beyond safety issues. For example, Cheng et al. [36] proposed a unified framework "AutoDetect", where three LLM-powered agents work collaboratively to automatically detect potential weaknesses in general-purpose tasks, such as mathematics and coding.

4.2 Learner: learning from scalable feedbacks

The learner model, which is the target model to be optimized in this framework, will make self-refinement by learning from scalable feedbacks from a critic model and also human feedbacks whenever necessary. The core of the learner model is alignment algorithms which enable the model to learn from scalable feedback. In past years, there have been some notable algorithms for this purpose: PPO [10], DPO [11], EXO [12], and many more. Since we have human experts in the loop, there raises a critical research problem in the learning process: how can the learner learn from mixed feedback signals, most times from model-generated feedback and rarely from human experts? In superalignment, designing more efficient and effective alignment algorithms is still the major challenge.

The form of feedback mainly falls into two types: a reward function which is trained on pair-wise preference data, or textual critics generated from a critic model. In our framework, we are more interested in textual critics as feedback since it does not require additional training to obtain a reward function. There are some critical research problems in this form of feedback: how can the learner learn from such textual critics? And which form of critics will be easier for the learner to learn from? Very recently, critic models such as criticGPT [29] and CritiqueLLM [34] have been proposed to generate scalable critics for diverse generation tasks from different dimensions, and with the assistance of such critic models, human experts are easier to provide reliable supervision, however, how such critics can be used to improve the learner model is still an under-explored problem.

A portion of previous work [44, 45] focuses on employing critiques to facilitate more accurate and fine-grained reward estimation, thereby improving the performance of learners in an indirect way. For instance, RELC [44] utilizes critiques to decompose sequence-level rewards into segment-level ones, aiming to alleviate the issue of reward sparsity in PPO optimization. Another line of research directly leverages critiques to refine the generated response through a refinement model. DRC [46] and FENCE [47] demonstrate that fine-grained critiques and refinements are more effective for enhancing the factuality of responses. In summary, due to its unique informational advantage over scalar rewards, natural language feedback holds significant potential for model optimization. However, what are the most efficient and learnable forms of critique is still largely under-explored.

4.3 Critic: generating scalable, faithful, and learnable critics

The critic model aims to generate scalable feedbacks for the learner model. The feedback, in the form of textual description in this paper, may be explanations or reasons why a response to an adversary query is good or bad. There are critical questions in building such critic model. First of all, the central role of the critic model lies in its criticizing ability: can the model generate relevant, informative, and discriminative explanations or reasons for a given query-response pair? Second, how can we ensure and evaluate the faithfulness of a generated critic? This problem is closely related to self-evaluation of a model-generated result [33, 48], probability calibration [49], and confidence estimation [50]. Third, how can the critic model generate critics that will be easily used to optimize the learner model? Such a critic will be called a learnable critic in this paper.

It is not trivial to train a general-purpose critic model that is generalizable across different generation tasks, topics, and evaluation dimensions. To evaluate various generation tasks, GPT-4 has been widely used to generate evaluative critics by prompt engineering. However, this method is faced with high cost, low stability, and low reproducibility. Moreover, the evaluation performance is largely determined by the ability of base models. Therefore, training a specialized critic model has become a common choice [51,52] recently, with the aim of avoiding potential risks of commercial APIs, such as high cost, unstable usage, and data leakage. However, it still faces challenges such as generalization capabilities and hallucinations, which hinder its further applicability. Moreover, several studies attempt to effectively utilize critiques through a new human-in-the-loop approach. OpenAI endeavors to train critic models to generate critiques in summarization [53] and code generation [29], assisting human annotators in identifying mistakes in responses more easily. The results indicate that critiques not only enhance the coverage and accuracy of human annotators in detecting mistakes, but also help generative models refine their own answers to improve their quality further, demonstrating the significant potential of the critic model in research on scalable oversight.

4.4 Realization of the superalignment framework

We proposed a conceptual framework for superalignment in the previous sections and we highlighted some key research problems in each module. We believe that when these problems have been solved, it will be feasible to run the pipeline smoothly. The essence of the superalignment framework lies in self-refinement or self-improvement: the learner can automatically learn and improve himself from scalable feedback.

Interestingly, there have been some notable research attempts similar to the idea of our proposed framework. These studies are highly related to keywords such as bootstrapping, self-alignment, self-play, and self-refine. In the "Self-Taught Reasoner" (STaR) [54], a bootstrapping reasoning technique was proposed. This method relies on a simple loop: generate rationales to answer questions by prompting the model with a few rationale examples; if the generated answers are wrong, try again to generate a rationale given the correct answer; fine-tune the model on all the rationales that ultimately yielded correct answers; and then, iterate the process. A self-alignment framework was proposed by Yuan et al. [15], which consists of two steps. In the self-instruction creation step, some newly created prompts are used to generate candidate responses from an earlier version of the model M_t, which also predicts its own rewards using the LLM-as-a-judge prompting approach. In the instruction following the training step, preference pairs are selected from the generated data based on the reward signals, on which a new model M_{t+1} was trained using the DPO algorithm [11]. The process can be iterated, resulting in both improved instruction following capabilities and enhanced reward modeling ability. Similarly, this idea was explored in self-play fine-tuning (SPIN) [55]: starting from an SFT dataset and an initial model M_0, the method generates synthetic data from an old model M_t ($t = 0$ at the start), and then trains a new version M_{t+1} using the DPO algorithm[4]; in the next iteration, the new version M_{t+1} is used to generate data to train a newer one M_{t+2}. Unlike self-alignment, which selects preference data using self-rewarding signals, SPIN assumes that model-generated data are always worse than the human data in the SFT dataset when constructing preference pairs. Their results show that SPIN can convert a weaker LLM to a stronger LLM and thereby demonstrate the promise of self-play. Another interesting idea, which is largely explored in the community, is self-refine [56]. In this approach, an LLM first generates an initial output and then provides feedback for its output, and subsequently uses the feedback to refine its output. This process can be repeated iteratively. Self-refine uses a single LLM as the generator, refiner, and feedback provider, requiring no additional training. Cheng et al. [39] proposed a self-refinement framework, SPAR, which involves an actor model to be optimized and a refiner model that critiques and generates improved responses through tree-search sampling. This framework effectively scales inference-time computation to construct high-quality training data, enabling continuous self-improvement of the actor and refiner through iterative training.

[4] The preference pairs are automatically constructed, assuming that model-generated responses are always worse than those in the SFT dataset, which largely limits the exploration space.

Some studies attempt to identify the weaknesses in the system automatically and fix them accordingly. Cheng et al. [36] introduced AutoDetect, a framework designed to automatically identify weaknesses in LLMs across various tasks. AutoDetect compromises three key agents: an Examiner, which constructs a detailed task taxonomy; a questioner, which generates queries; and an assessor, which analyzes low-scoring cases to identify potential weaknesses. The questioner's queries are input to the target model, and the responses are scored to identify weak points. This framework has achieved a success rate of over 30% in top models like ChatGPT and Claude. Additionally, the identified weaknesses can help enhance models, such as the LLaMA series, through supervised fine-tuning. Bai et al. [57] introduced the language-model-as-an-examiner framework, designed to automatically benchmark the knowledge of foundation models. This framework employs an LLM as an examiner to generate diverse questions across domains, probe deeper knowledge through follow-up queries, and evaluate the model's responses. Beyond assessing performance, this approach can also serve as a tool for identifying knowledge-related weaknesses in the tested models. Cohen et al. [58] introduced a cross-examination-based framework for evaluating the factuality of language models. This approach involves two interacting LMs: an examinee, which generates claims, and an examiner, which conducts a multi-turn interaction to identify inconsistencies in the examinee's responses. Inspired by legal truth-seeking mechanisms, the examiner crafts targeted questions to uncover contradictions and expose factual inaccuracies in the examinee's claims.

Despite impressive empirical progress, a fundamental understanding of LLM self-improvement remains very limited, thereby requiring much deeper theoretical modeling and empirical analysis. Some studies have reported that recursively training next-generation models on the data generated by previous models can lead to model collapse [59]: a degenerative learning process where models start forgetting improbable events over time, as the models become poisoned with its own generated, biased data. In image generation, Ref. [60] showed that without enough fresh real data in each generation of an autophagous loop, future generative models can have a progressive decrease in output quality or diversity. In [61], the authors discovered a consistent decrease in the diversity of model outputs through iterative training, particularly in highly creative tasks. This finding highlights the potential risks associated with training language models on synthetic text, especially in terms of preserving linguistic richness. Similarly, Ref. [62] showed that the self-refinement training loops can reduce output diversity, with the extent of depending on the proportion of the used generated data. While introducing fresh data can slow this decline, it does not entirely prevent it. In [63], the authors also observed declines in output diversity and out-of-distribution (OOD) generalization during LLM self-refinement training. Notably, Ref. [64] presented a mathematical formulation of self-improvement and formalized the concept of generation-verification gap. The authors reveal that the gap between the verification capability (judging the quality of generations) and the generation capability is the key driving force behind self-refinement. They studied verification mechanisms to improve self-refinement; for instance, an ensemble of different verification methods can enhance self-improvement. We believe that this is the most in-depth theoretical analysis of self-refinement up to now.

5 Conclusion and future directions

In this paper, we discuss the superalignment of superhuman AI systems with LLMs. We give an informal definition of superalignment by outlining the shift of learning paradigms—from pretraining, supervised fine-tuning, alignment, to superalignment. Afterward, we highlight some key research problems in superalignment, including weak-to-strong generalization, scalable oversight, and evaluation of the alignment. Then, we present a conceptual framework for realized superalignment, which comprises three components: an attacker, which aims to automatically discover the weaknesses of LLMs, a learner, which learns from scalable feedbacks (mixture of AI and human feedbacks), and a critic, which produces scalable, faithful, and learnable critics. We highlight some critical research problems in each component, and summarize some major research advancements in these sub-directions. Finally, we also summarize some interesting research attempts that are highly related or may lay a foundation for superalignment: self-alignment, self-play, self-refinement, and others. These studies can be viewed as early attempts towards superalignment, which show promising results, thereby partially verifying the feasibility of the framework proposed in this paper.

Though still in its infancy, superalignment poses new research problems that are worthy of study in the near future.

Identifying new emergent risks of superhuman intelligence. The safety of superhuman AI systems has gained much attention in recent years, leading to the identification and study of many safety issues including discrimination, bias, property and privacy violations, misinformation and disinformation, ethical considerations and social norms, and many more. We refer to these safety issues as low-order safety problems as they can often be detected through superficial cues in the generated content. However, high-order safety problems, such as purposely deception to mislead human and manipulation of human beliefs, may be more subtle, indirect, and complex to identify, and require long-term evaluation. Moreover, unknown risks in specialized domains (e.g., biological threats) are also very dangerous threats to our society. Recognizing, identifying, and evaluating such unknown risks in high-stake fields are very critical to AI safety.

Providing reliable and scalable oversight to superhuman models. We have discussed some studies on self-alignment, self-play, and self-refinement, which share a common principle: iteratively refining a model with synthetic data. However, in superalignment, how to synthesize the high-quality data that effectively challenge the current model remains challenging. Additionally, how can we provide reliable oversight of such synthetic data largely requires human-AI collaboration. Due to the scalability issue, in most cases we have to rely on AI feedbacks, but when human experts will intervene and how they will be evolved in the pipeline is a complex problem. There are still many research problems worth doing in the future.

Aligning large language models from multiple dimensions. Aligning large language models to human values is an extremely complex problem which requires considerations of diverse aspects including cultures, regions, and countries. Existing studies mainly focus on a single perspective; however, integrating multiple perspectives, such as values, safety, and social norms, within a unified paradigm ethics paradigm is yet to be considered. Therefore, it is indispensable to design multi-objective optimization alignment algorithms to simultaneously model these factors [65].

Acknowledgements This work was supported by National Natural Science Foundation of China Distinguished Young Scholar Project (Grant No. 62125604). We would like to thank students Bosi WEN and Jiale CHENG for minor edits to the manuscript and thank the students for earlier discussions of this superalignment framework, including Chujie ZHENG, Haozhe JI, Yuxian GU, Zhexin ZHANG, and many others.

References

1 OpenAI. Introducing ChatGPT, 2022. https://openai.com/index/chatgpt/
2 Zhao W X, Zhou K, Li J, et al. A survey of large language models. 2023. ArXiv:2303.18223
3 OpenAI. OpenAI Charter. 2018. https://openai.com/charter/
4 Marcus G. Definition to AGI. 2022. https://twitter.com/GaryMarcus/status/1529457162811936768
5 LeCun Y. Human Intelligence. 2024. https://twitter.com/ylecun/status/1846034609521246234
6 LeCun Y. AMI (advanced machine intelligence). 2024. https://twitter.com/ylecun/status/1794249923329720415
7 Shevlane T, Farquhar S, Garfinkel B, et al. Model evaluation for extreme risks. 2023. ArXiv:2305.15324
8 Bengio Y, Hinton G, Yao A, et al. Managing extreme AI risks amid rapid progress. Science, 2024, 384: 842–845
9 Shen T, Jin R, Huang Y, et al. Large language model alignment: a survey. 2023. ArXiv:2309.15025
10 Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. 2017. ArXiv:1707.06347
11 Rafailov R, Sharma A, Mitchell E, et al. Direct preference optimization: your language model is secretly a reward model. In: Proceedings of Advances in Neural Information Processing Systems, 2024
12 Ji H, Lu C, Niu Y, et al. Towards efficient exact optimization of language model alignment. In: Proceedings of the 41st International Conference on Machine Learning, 2024
13 OpenAI. Introducing superalignment. 2023. https://openai.com/index/introducing-superalignment/
14 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd International Conference on Learning Representations, 2015
15 Yuan W, Pang R Y, Cho K, et al. Self-rewarding language models. In: Proceedings of the 41st International Conference on Machine Learning, 2024
16 Liu T, Zhao Y, Joshi R, et al. Statistical rejection sampling improves preference optimization. 2023. ArXiv:2309.06657
17 Song F, Yu B, Li M, et al. Preference ranking optimization for human alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2024. 18990–18998
18 Ethayarajh K, Xu W, Muennighoff N, et al. KTO: model alignment as prospect theoretic optimization. 2024. ArXiv:2402.01306
19 Azar M G, Guo Z D, Piot B, et al. A general theoretical paradigm to understand learning from human preferences. In: Proceedings of International Conference on Artificial Intelligence and Statistics, 2024. 4447–4455
20 Burns C, Izmailov P, Kirchner J H, et al. Weak-to-strong generalization: eliciting strong capabilities with weak supervision. In: Proceedings of the 41st International Conference on Machine Learning, 2024
21 Bowman S R, Hyun J, Perez E, et al. Measuring progress on scalable oversight for large language models. 2022. ArXiv:2211.03540
22 Christiano P, Shlegeris B, Amodei D. Supervising strong learners by amplifying weak experts. 2018. ArXiv:1810.08575
23 Wen J, Zhong R, Ke P, et al. Learning task decomposition to assist humans in competitive programming. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, 2024. 11700–11723
24 Leike J, Krueger D, Everitt T, et al. Scalable agent alignment via reward modeling: a research direction. 2018. ArXiv:1811.07871
25 Sun Z, Shen Y, Zhou Q, et al. Principle-driven self-alignment of language models from scratch with minimal human supervision. In: Proceedings of Advances in Neural Information Processing Systems, 2024
26 Bai Y, Kadavath S, Kundu S, et al. Constitutional AI: harmlessness from AI feedback. 2022. ArXiv:2212.08073

27 Du Y, Li S, Torralba A, et al. Improving factuality and reasoning in language models through multiagent debate. In: Proceedings of the 41st International Conference on Machine Learning, 2024
28 Irving G, Christiano P, Amodei D. AI safety via debate. 2018. ArXiv:1805.00899
29 McAleese N, Pokorny R M, Uribe J F C, et al. LLM critics help catch LLM bugs. 2024. ArXiv:2407.00215
30 Cobbe K, Kosaraju V, Bavarian M, et al. Training verifiers to solve math word problems. 2021. ArXiv:2110.14168
31 Hendrycks D, Burns C, Basart S, et al. Measuring massive multitask language understanding. In: Proceedings of International Conference on Learning Representations, 2021
32 Chen M, Tworek J, Jun H, et al. Evaluating large language models trained on code. 2021. ArXiv:2107.03374
33 Zheng L, Chiang W L, Sheng Y, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. In: Proceedings of Advances in Neural Information Processing Systems, 2023. 46595–46623
34 Ke P, Wen B, Feng A, et al. CritiqueLLM: towards an informative critique generation model for evaluation of large language model generation. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, 2024. 13034–13054
35 Zhang L, Hosseini A, Bansal H, et al. Generative verifiers: reward modeling as next-token prediction. 2024. ArXiv:2408.15240
36 Cheng J, Lu Y, Gu X, et al. AutoDetect: towards a unified framework for automated weakness detection in large language models. In: Proceedings of Findings of the Association for Computational Linguistics, Miami, 2024. 6786–6803
37 Skalse J, Howe N, Krasheninnikov D, et al. Defining and characterizing reward gaming. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 9460–9471
38 Wen J, Zhong R, Khan A, et al. Language models learn to mislead humans via RLHF. 2024. ArXiv:2409.12822
39 Cheng J, Liu X, Wang C, et al. SPaR: self-play with tree-search refinement. In: Proceedings of the International Conference on Learning Representations, 2025
40 Zou A, Wang Z, Kolter J Z, et al. Universal and transferable adversarial attacks on aligned language models. 2023. ArXiv:2307.15043
41 Liu C, Dong Y, Xiang W, et al. A comprehensive study on robustness of image classification models: benchmarking and rethinking. Int J Comput Vis, 2025, 133: 567–589
42 Zhang C, Hu M, Li W, et al. Adversarial attacks and defenses on text-to-image diffusion models: a survey. Inf Fusion, 2025, 114: 102701
43 Wen J, Ke P, Sun H, et al. Unveiling the implicit toxicity in large language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 2023. 1322–1338
44 Ye Z, Greenlee-Scott F, Bartolo M, et al. Improving reward models with synthetic critiques. 2024. ArXiv:2405.20850
45 Cao Y, Sheng Q Z, McAuley J, et al. Reinforcement learning for generative AI: a survey. 2023. ArXiv:2308.14328
46 Wadhwa M, Zhao X, Li J J, et al. Learning to refine with fine-grained natural language feedback. In: Proceedings of Findings of the Association for Computational Linguistics, Miami, 2024. 12281–12308
47 Xie Y, Zhou W, Prakash P, et al. Improving model factuality with fine-grained critique-based evaluator. 2024. ArXiv:2410.18359
48 Panickssery A, Bowman S R, Feng S. LLM evaluators recognize and favor their own generations. In: Proceedings of the 38th Annual Conference on Neural Information Processing Systems, 2024
49 Gao Y, Xu G, Wang Z, et al. Bayesian calibration of win rate estimation with LLM evaluators. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Miami, 2024. 4757–4769
50 Jung J, Brahman F, Choi Y. Trust or escalate: LLM judges with provable guarantees for human agreement. 2024. ArXiv:2407.18370
51 Li J, Sun S, Yuan W, et al. Generative judge for evaluating alignment. In: Proceedings of the 12th International Conference on Learning Representations, 2023
52 Zhu L, Wang X, Wang X. JudgeLM: fine-tuned large language models are scalable judges. 2023. ArXiv:2310.17631
53 Saunders W, Yeh C, Wu J, et al. Self-critiquing models for assisting human evaluators. 2022. ArXiv:2206.05802
54 Zelikman E, Wu Y, Mu J, et al. Star: bootstrapping reasoning with reasoning. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 15476–15488
55 Chen Z, Deng Y, Yuan H, et al. Self-play fine-tuning converts weak language models to strong language models. In: Proceedings of the 41st International Conference on Machine Learning, 2024
56 Madaan A, Tandon N, Gupta P, et al. Self-refine: iterative refinement with self-feedback. In: Proceedings of Advances in Neural Information Processing Systems, 2024
57 Bai Y, Ying J, Cao Y, et al. Benchmarking foundation models with language-model-as-an-examiner. In: Proceedings of Advances in Neural Information Processing Systems, 2024
58 Cohen R, Hamri M, Geva M, et al. LM vs LM: detecting factual errors via cross examination. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023. 12621–12640
59 Shumailov I, Shumaylov Z, Zhao Y, et al. The curse of recursion: training on generated data makes models forget. 2024. ArXiv:2305.17493
60 Alemohammad S, Casco-Rodriguez J, Luzi L, et al. Self-consuming generative models go MAD. 2023. ArXiv:2307.01850
61 Guo Y, Shang G, Vazirgiannis M, et al. The curious decline of linguistic diversity: training language models on synthetic text. 2024. ArXiv:2311.09807
62 Briesch M, Sobania D, Rothlauf F. Large language models suffer from their own output: an analysis of the self-consuming training loop. 2024. ArXiv:2311.16822
63 Wu T, Li X, Liu P. Progress or regress? Self-improvement reversal in post-training. 2024. ArXiv:2407.05013
64 Song Y, Zhang H, Eisenach C, et al. Mind the gap: examining the self-improvement capabilities of large language models. 2024. ArXiv:2412.02674
65 Zhong Y, Ma C, Zhang X, et al. Panacea: Pareto alignment via preference adaptation for LLMs. 2024. ArXiv:2402.02030