A Comprehensive Survey on Learning from Rewards for Large Language Models: Reward Models and Learning Strategies

Xiaobao Wu
Nanyang Technological University
xiaobao002@e.ntu.edu.sg

Abstract

Recent developments in Large Language Models (LLMs) have shifted from pre-training scaling to post-training and test-time scaling. Across these developments, a key unified paradigm has arisen: *Learning from Rewards*, where reward signals act as the guiding stars to steer LLM behavior. It has underpinned a wide range of prevalent techniques, such as reinforcement learning (RLHF, RLAIF, DPO, and GRPO), reward-guided decoding, and post-hoc correction. Crucially, this paradigm enables the transition from passive learning from static data to active learning from dynamic feedback. This endows LLMs with aligned preferences and deep reasoning capabilities for diverse tasks. In this survey, we present a comprehensive overview of learning from rewards, from the perspective of reward models and learning strategies across training, inference, and post-inference stages. We further discuss the benchmarks for reward models and the primary applications. Finally we highlight the challenges and future directions. [1]

1 Introduction

Recent years have witnessed the rapid advancement of Large Language Models (LLMs), such as ChatGPT (OpenAI, 2023), Claude (Anthropic, 2025), and Llama (Meta, 2023, 2024). These models are initially empowered by *pre-training scaling* (Kaplan et al., 2020), which trains LLMs on massive corpora through next-token prediction. While this approach enables broad linguistic and knowledge representations, it suffers from several fundamental limitations: misalignment with human values (Bai et al., 2022b; Zhang et al., 2023b; Deshpande et al., 2023), difficulty in adapting to various task objectives (Lyu et al., 2023; Wang et al., 2023a), and deficiencies in deep reasoning (Mirzadeh et al., 2024; Wu et al., 2024b). As a result, these confine pre-trained models to surface-level tasks, falling short of the long-term goal of robust and general AI. To address these limitations, recent efforts have turned to *post-training* and *test-time scaling*, which seek to further refine LLMs after pre-training.

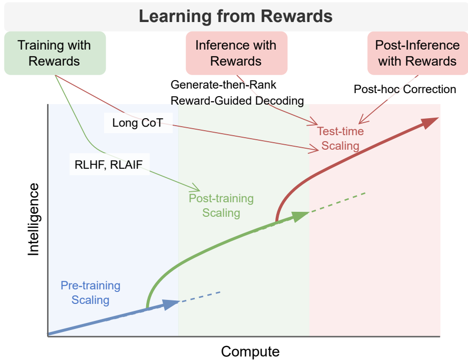

Figure 1: Illustration of the scaling phases of LLMs. The learning-from-rewards paradigm plays a pivotal role in the post-training and test-time scaling.

Across the post-training and test-time scaling, a critical unified paradigm has emerged as illustrated in Figure 1: ***Learning from Rewards***, which leverages reward signals to guide model behavior through diverse learning strategies. For post-training scaling, this paradigm has underpinned several key techniques, including preference alignment through Reinforcement Learning from Human Feedback (RLHF, Ouyang et al., 2022) or AI Feedback (RLAIF, Bai et al., 2022b) with scalar rewards and PPO (Schulman et al., 2017), and DPO (Rafailov et al., 2023) with implicit rewards. For test-time scaling, this paradigm supports eliciting long Chain-of-Thoughts reasoning via GRPO (Shao et al., 2024) with rule-based rewards, generate-then-rank (Cobbe et al., 2021; Lightman et al., 2023), reward-guided decoding (Deng and Raffel, 2023; Khanov et al., 2024), and post-hoc correction (Akyurek et al., 2023; Madaan et al., 2023). Through these techniques,

[1]We maintain a paper collection at https://github.com/bobxwu/learning-from-rewards-llm-papers.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 17847–17875
November 4-9, 2025 ©2025 Association for Computational Linguistics

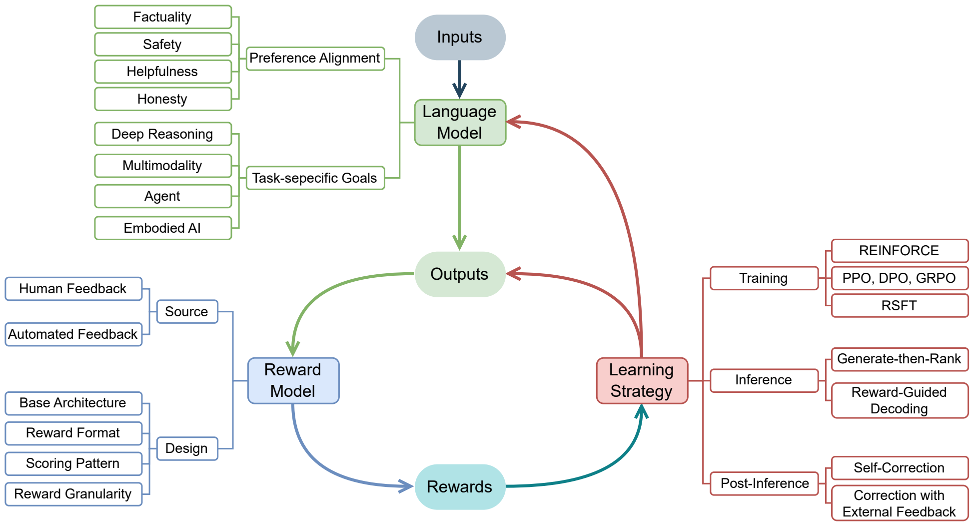

Figure 2: A unified framework of learning from rewards. The language model generates outputs; the reward model evaluates the outputs and provides reward signals; the learning strategy leverages the rewards to either fine-tune the language model or refine the outputs, occurring at the training, inference, or post-inference stages.

this paradigm enables LLMs to learn actively from dynamic feedback, in contrast to learning passively from static data. As such, this endows LLMs with aligned preferences and deep reasoning and planning abilities, leading to more intelligent agents. In consequence, this paradigm has inspired many applications, such as mathematical reasoning (DeepSeek-AI, 2025), code generation (Zhu et al., 2024), multimodality (Liu et al., 2025h), and agents (OpenAI, 2025).

Due to this growing prevalence, we comprehensively survey the *learning from rewards* for LLMs. We first introduce a taxonomy that categorizes existing works with a unified conceptual framework regarding reward model design and learning strategies (Sec. 2). Then we review representative techniques across three main stages: *training with rewards*, *inference with rewards*, and *post-inference with rewards* (Sec. 3 to 5). We additionally summarize primary applications, recent reward model benchmarks, and key challenges and promising directions for future research (Appendices A to C).

2 A Taxonomy of Learning from Rewards for LLMs

We first introduce a unified conceptual framework that captures the key components and interactions to understand learning from rewards systemically. Building upon this framework, we categorize the primary dimensions along which existing methods vary: (i) **Reward Source**; (ii) **Reward Model**; (iii) **Learning Stage**; (iv) **Learning Strategies**. Each dimension reflects a distinct aspect of how reward signals are acquired, represented, and utilized in language models.

2.1 A Unified Conceptual Framework

We present a unified conceptual framework for learning from rewards in Figure 2. It abstracts the key components and interactions involved in learning from rewards for language models. In this framework, the *language model* generates outputs conditioned on the inputs; the reward model then provides rewards to evaluate the output quality; the learning strategy leverages the reward signals to update the language model or adjusts the outputs.

Language Model. A language model $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ generates an output $\hat{y} \in \mathcal{Y}$ given an input $x \in \mathcal{X}$. This formulation covers a wide range of tasks, such as question answering, summarization, and image captioning.

Reward Model. A reward model evaluates the quality of an output $\hat{y}$ given an input x and produces a reward signal r that reflects desired properties, such as helpfulness, safety, or task-specific correctness. In different contexts, a reward model may be referred to as a verifier and an evaluator. We emphasize that here we adopt a broad definition of the reward model: it can be model-based or model-free. We will discuss these later.

Learning Strategy. A learning strategy uses reward signals to adjust the behavior of the language model. Here we consider both the training-

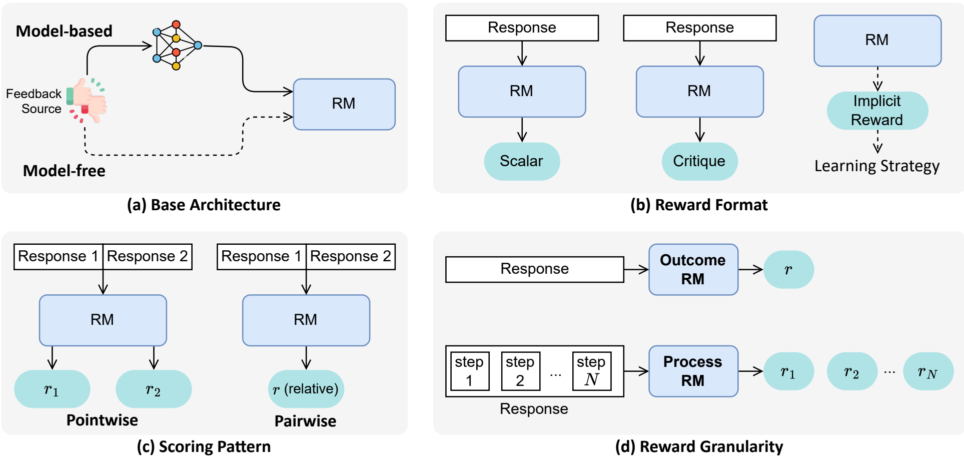

Figure 3: Reward Model (RM) design dimensions.

based (updating model parameters) and training-free strategies (directly refining model outputs).

2.2 Reward Source

Reward signals originate from two primary sources: **Human Feedback** and **Automated Feedback**. Each offers trade-offs in terms of reliability, scalability, and cost. We introduce them respectively as follows.

Human Feedback. Human feedback provides high-quality reward signals grounded in human judgment and intent. It typically collects human annotations through pairwise comparisons between alternative model outputs, *e.g.,* chosen and rejected responses. The collected preference data can be used to train explicit reward models like RLHF (Ouyang et al., 2022) or directly fine-tune the language model like DPO (Rafailov et al., 2023). While effective, this approach is resource-intensive and may not scale easily across domains or tasks.

Automated Feedback. To reduce the cost of human annotations and scale up the reward model training, automated feedback has been increasingly explored as an alternative. The automated feedback mainly includes (**i**) **Self-Rewarding**, where the language model critiques its own outputs (Yuan et al., 2024b; Wang et al., 2024d); (**ii**) **Trained Models**, such as powerful LLMs following the LLM-as-a-Judge design (Bai et al., 2022b; Lee et al., 2023); (**iii**) **Predefined Rules**, *i.e.,* verifiable rewards, such as accuracy and format rules used in DeepSeek-R1 (Shao et al., 2024; DeepSeek-AI et al., 2025). (**iv**) **Knowledge**, such as structured knowledge base or Wikipedia (Peng et al., 2023; Tian et al., 2023). (**v**) **Tools**, such as program compilers and interactive systems (Le et al., 2022; Liu et al., 2023). The automated feedback enables scalable reward generation but may introduce limitations in interpretability, generality, and alignment quality.

2.3 Reward Model

Reward models are the central foundation of learning from rewards. As shown in Figure 3, we organize the design space of reward model into four key dimensions: (i) **Base Architecture**; (ii) **Reward Format**; (iii) **Scoring Pattern**; (iv) **Reward Granularity**.

Base Architecture. As shown in Figure 3(a), this refers to the base architecture of a reward model. Here we consider a broad view of reward models, including both model-based and model-free architectures.

- **Model-based Architecture**. A dedicated reward model is trained to evaluate outputs. Common variants include
  (**a**) **Scalar Reward Models**. These models assign a scalar score to a candidate response, indicating its quality. Typically, they are built upon Transformer backbones (*e.g.,* GPT or BERT variants) with a value head that outputs scalars. They are trained with preference data via pairwise ranking losses such as the Bradley-Terry loss (Nakano et al., 2021; Ouyang et al., 2022; Liu et al., 2024a).
  (**b**) **Generative Reward Models**. These models generate natural language critiques as reward signals. They commonly follow LLM-as-a-Judge with general models (Zheng et al., 2023) or training specialized models (Li et al., 2023a; Cao et al., 2024; Ye et al., 2024; McAleese et al.,

2024). They have become more popular recently because they can leverage the deep reasoning capabilities of large reasoning models and provide finer-grained supervision (Huang et al., 2025a; Guo et al., 2025a).

(**c**) **Semi-scalar Reward Models**. These models combine scalars with critiques, offering both quantitative and qualitative assessment (Yu et al., 2024a; Zhang et al., 2025f). Their architectures usually involve two heads, one for scalar rewards and another for critique rewards.

- **Model-free Architecture**. Instead of an explicit reward model, model-free approaches derive reward signals directly from diverse feedback sources, such as preference data, tools, or knowledge. The resulting rewards can be scalar, critique, or implicit signals. For example, DPO (Rafailov et al., 2023) circumvents the need to train a reward model by directly aligning the language model with preference data through fine-tuning. Similarly, GRPO (DeepSeek-AI et al., 2025) adopts rule-based rewards from handcrafted constraints and task-specific heuristics.

Model-based and model-free approaches each present distinct trade-offs in reward specification and practical applicability. Model-based approaches provide flexible and generalizable reward evaluation. Once trained, reward models can be reused across tasks and enable iterative optimization. However, they require costly preference data, are prone to overfitting, and may introduce bias or reward hacking issues. Model-free methods avoid training a separate reward model, offering a simpler, sample-efficient, and usually more stable pipeline. However, they are typically task-specific, lack generalization, and offer limited flexibility for downstream reuse.

In order to align with previous literature, **we hereafter refer to the reward model as the model-based by default**.

Reward Format. As shown in Figure 3(b), this describes the specific format of reward signals:

- **Scalar Rewards**, numerical scores that quantify the quality of model outputs. They are the most commonly used format due to their simplicity and compatibility with learning strategies such as reinforcement learning. Their limitation lies in the sparsity and interpretability.
- **Critique Rewards**, natural language feedback that evaluates the quality of outputs (Saunders et al., 2022; Kwon et al., 2023), such as *"The score of this response is 3 out of 5"*. They are more expressive and interpretable than scalar rewards, enabling finer-grained guidance, but they may require additional processing to be used in certain learning algorithms.
- **Implicit Rewards** are signals implicitly embedded in the source without explicit supervision, such as preference data in DPO (Rafailov et al., 2023; Meng et al., 2024). This format simplifies the implementation but places more burden on the learning strategies to infer appropriate optimization signals.

Scoring Pattern. As shown in Figure 3(c), this dimension determines how responses are scored:

- **Pointwise Scoring** assigns a score to each response independently. It is the most widely used scoring pattern in reward models.
- **Pairwise Scoring** compares response pairs and selecting the preferred one. The pairwise scoring can be expressed as a scalar score indicating relative preference or a natural language critique such as *"Response 1 is better than Response 2"*.

Reward Granularity. As shown in Figure 3(d), we identify two kinds of reward granularity: reward granularity reflects the level of resolution at which feedback is provided:

- **Outcome Reward Models** evaluate the holistic quality of outputs, treating it as a single unit.
- **Process Reward Models** evaluate intermediate steps within the reasoning process of outputs, enabling fine-grained supervision during generation (Lightman et al., 2023; Wang et al., 2023b).

2.4 Learning Stage

Learning from rewards can occur at different stages of the language model lifecycle, including **Training**, **Inference**, and **Post-Inference**.

- **Training with Rewards.** At the training stage, reward signals can be transformed into optimization signals by training algorithms to fine-tune the language model, which is the most extensively explored in the literature. It can support post-training alignment with human preference (Ouyang et al., 2022; Bai et al., 2022b) and test-time scaling by eliciting the language models' deep reasoning capabilities through long Chain-of-Thoughts (CoT) (DeepSeek-AI et al., 2025).

- **Inference with Rewards.** During inference, reward signals can guide the decoding of model outputs without modifying model parameters. This enables test-time scaling by searching in a larger decoding space, such as *Best-of-N* and tree search (Cobbe et al., 2021; Snell et al., 2025).
- **Post-Inference with Rewards.** This stage uses rewards to refine model outputs after generation without modifying model parameters. Post-inference with rewards also supports test-time scaling by iteratively refining the outputs (Shinn et al., 2023).

2.5 Learning Strategy

Various learning strategies have been developed to incorporate reward signals to steer model behavior. These strategies are commonly divided into two types: **Training-based** and **Training-free**.

- **Training-based Strategies.** Training-based strategies optimize the language model by converting reward signals into gradient-based updates. The optimization mainly depends on Reinforcement Learning (RL) where language models act as policy models, or Supervised Fine-Tuning (SFT). Representative examples include Proximal Policy Optimization (PPO, Schulman et al., 2017; Ouyang et al., 2022), Direct Preference Optimization (DPO, Rafailov et al., 2023; Meng et al., 2024), Group Relative Policy Optimization (GRPO, Shao et al., 2024), and Rejection-Sampling Fine-Tuning (RSFT, Nakano et al., 2021; Yuan et al., 2023a; Dong et al., 2023)
- **Training-free Strategies.** Training-free strategies leverage reward signals to guide or refine model outputs without updating the language model parameters. They include generate-then-rank, such as *Best-of-N* (Cobbe et al., 2021; Lightman et al., 2023), reward-guided decoding (Deng and Raffel, 2023; Khanov et al., 2024), and post-inference correction (Shinn et al., 2023; Pan et al., 2023a). These methods provide a relatively lightweight mechanism for improving model outputs, and some are highly compatible with various model architectures. They are particularly useful when model fine-tuning is infeasible or computationally expensive.

The above presents a detailed taxonomy of learning from rewards for LLMs. We will review the representative studies across the three learning stages: training, inference, and post-inference with rewards in the following Sec. 3 to 5.

3 Training with Rewards

In this section, we introduce the methods for training LLMs with rewards. They contribute to post-training scaling for preference alignment and test-time scaling by eliciting long CoT abilities.

3.1 Training with Scalar Rewards

Training the language model with *scalar rewards* is the most extensively studied strategy in the literature. We classify these methods based on human and automated feedback as follows.

Scalar Rewards from Human Feedback. Human feedback is a key source for constructing scalar rewards. The most prominent example is RLHF (Ziegler et al., 2019; Ouyang et al., 2022; Bai et al., 2022a; Glaese et al., 2022). RLHF trains a scalar reward model on human preference data (pairwise comparisons with chosen and rejected responses). The reward models commonly adopt the Transformer architecture with a value head that outputs scalars, and their training objectives follow the Bradley-Terry loss (Bradley and Terry, 1952), which maximizes the reward differences between preferred and dispreferred outputs. The trained reward model assigns evaluative scalar scores to the model outputs, serving as a proxy for human judgment. With the reward model, RLHF fine-tunes the language model through PPO to align it with human preferences, such as harmlessness and helpfulness. Various variants have been explored, such as *Safe RLHF* (Dai et al., 2023) and *Fine-Grained RLHF* (Wu et al., 2023).

Scalar Rewards from Automated Feedback. A growing body of work explores automated feedback as a substitute to provide scalar rewards, which bypasses expensive human annotations. A prominent example is RLAIF (Bai et al., 2022b). RLAIF uses an LLM as a proxy judge to generate preference data following the idea of *LLM-as-a-Judge* (Zheng et al., 2023; Yu et al., 2025a). RLAIF also trains a scalar reward model on them and then uses it to fine-tune the language model. Automated feedback can also come from other models (Wang et al., 2024d; Dutta et al., 2024; Ahn et al., 2024) and various tools, such as code compilers (Liu et al., 2023; Dou et al., 2024; Gehring et al., 2024).

3.2 Training with Critique Rewards

Another line of work explores training with *critique rewards*. They commonly rely on generative reward models, and some could provide explanations and refinement suggestions through reasoning. For instance, *Auto-J* (Li et al., 2023a) generates critiques that support pointwise and pairwise evaluation. It adopts GPT-4 to produce evaluation judgments as the training data. *CompassJudger-1* (Cao et al., 2024) and *Con-J* (Ye et al., 2024) follow a similar design. *SFR-Judges* (Wang et al., 2024c) fine-tunes an LLM on the response deduction task to improve its judging ability.

3.3 Training with Implicit Rewards

Besides, many methods adopt *implicit rewards* for training. The reward signals are not provided directly but are implicitly embedded in the structure of the training data, such as preference pairs. Some use a scalar reward model to construct training data, but not for fine-tuning. Their reward signals for fine-tuning are encoded in the training data, so we treat them as training with implicit rewards.

Implicit Rewards from Human Feedback. A pioneering approach using implicit rewards from human feedback is DPO (Rafailov et al., 2023). DPO encodes implicit rewards via the log-likelihood difference between preferred and dispreferred responses. As such, DPO effectively reduces complicated RLHF into supervised fine-tuning. Several variants have been proposed based on DPO to further simplify the training or expand its applicability, such as SimPO (Meng et al., 2024) and KTO (Ethayarajh et al., 2024).

Apart from the DPO style, another line of work follows a Rejection-Sampling Fine-Tuning (RSFT) scheme. They typically select high-quality samples from a large number of candidate data for SFT. Representative work includes *RAFT* (Dong et al., 2023), *ReST* (Gulcehre et al., 2023), *RSO* (Liu et al., 2024b), and *RRHF* (Yuan et al., 2023b).

Implicit Rewards from Automated Feedback. Implicit rewards can originate from diverse automated feedback as well, such as AI feedback, feedback from external knowledge and external tools. **AI feedback** is a common source of implicit rewards, including self-rewarding and other trained models. *Self-Rewarding* (Yuan et al., 2024b) leverages the language model to evaluate its own outputs and construct preference data for fine-tuning with iterative DPO. *Meta-Rewarding* (Wu et al., 2024a) additionally evaluates its own judgments. Zhang et al. (2025c) extend self-rewarding to the process-level. Instead of direct self-assessment, some methods depend on self-consistency to model implicit rewards, like *SCPO* (Prasad et al., 2024) and *PFPO* (Jiao et al., 2024a). **External knowledge and tools** can provide feedback to model implicit rewards. Tian et al. (2023) and *FLAME* (Lin et al., 2024a) construct preference pairs by checking whether model outputs are supported by Wikipedia. *TRICE* (Qiao et al., 2023), *CodeLutra* (Tao et al., 2024), and Xiong et al. (2025) leverage tool execution results to construct preference data.

3.4 Training with Rule-based Rewards

Recently, training with *rule-based rewards* has gained prominence, since DeepSeek-R1 shows they can elicit long CoT abilities for LLMs (DeepSeek-AI et al., 2025). Rule-based rewards are derived by verifying outputs against specific rules, such as format constraints and evaluation metrics. Rule-based rewards are also referred to as *verifiable rewards/outcomes* due to their clean evaluation criteria. In detail, DeepSeek-R1 (DeepSeek-AI et al., 2025) defines two types of rule-based rewards: *accuracy rewards* and *format rewards*. With these rule-based rewards, it fine-tunes the language model through the RL algorithm GRPO (Shao et al., 2024). GRPO eliminates the dependence on the reward and value model in PPO and the preference data construction in DPO. Later, many following studies have been proposed. *DAPO* (Yu et al., 2025b) and *Open-R1* (Face, 2025) introduce open-source training frameworks, and some extended GRPO algorithms are introduced (Xu et al., 2025b; Zuo et al., 2025; Feng et al., 2025c; Zhang et al., 2025b).

3.5 Training with Process Rewards

An emerging line of work focuses on training with *process rewards*.Figure 3(d) shows these methods commonly employ a Process Reward Model (PRM) to assess the intermediate steps of model outputs. This provides more fine-grained supervision, which especially benefits complex reasoning tasks.

Process Rewards from Human Feedback. Early studies leverage human annotations to train PRMs. For instance, Uesato et al. (2022); Lightman et al. (2023) train PRMs using human annotations on intermediate mathematical reasoning steps. Uesato et al. (2022) then use the trained PRM to fine-tune the language model via reinforcement learning to improve its math reasoning.

Process Rewards from Automated Feedback. Recent efforts leverage automated feedback to supervise PRM's training at scale and avoid intensive step-level human annotations. One major direction leverages strong LLMs to generate step-level annotations, such as *WizardMath* (Luo et al., 2023) and *ActPRM* (Duan et al., 2025). Alternatively, other methods estimate process rewards without explicit annotations, including Monte Carlo estimation like *Math-Shepherd* (Wang et al., 2023b) and Jiao et al. (2024b), ranking estimation (Li and Li, 2024), and trajectory sampling like *OmegaPRM* (Luo et al., 2024) and *HRM* (Wang et al., 2025b). Others attempt to derive process rewards from outcome rewards, such as Yuan et al. (2024a), *PRIME* (Cui et al., 2025), and *OREAL* (Lyu et al., 2025). Others design generative PRMs with reasoning processes, such as *GenPRM* (Zhao et al., 2025b), *R-PRM* (She et al., 2025) and *ThinkPRM* (Khalifa et al., 2025).

4 Inference with Rewards

After the training stage, inference with rewards offers a flexible and lightweight mechanism to adapt and steer the model behavior without modifying model parameters. We identify two primary inference-with-rewards strategies: (i) *Generate-then-Rank* and (ii) *Reward-Guided Decoding*. These strategies play a critical role for achieving ***test-time scaling***: They allow the language model to search, reflect, and revise its outputs on the fly.

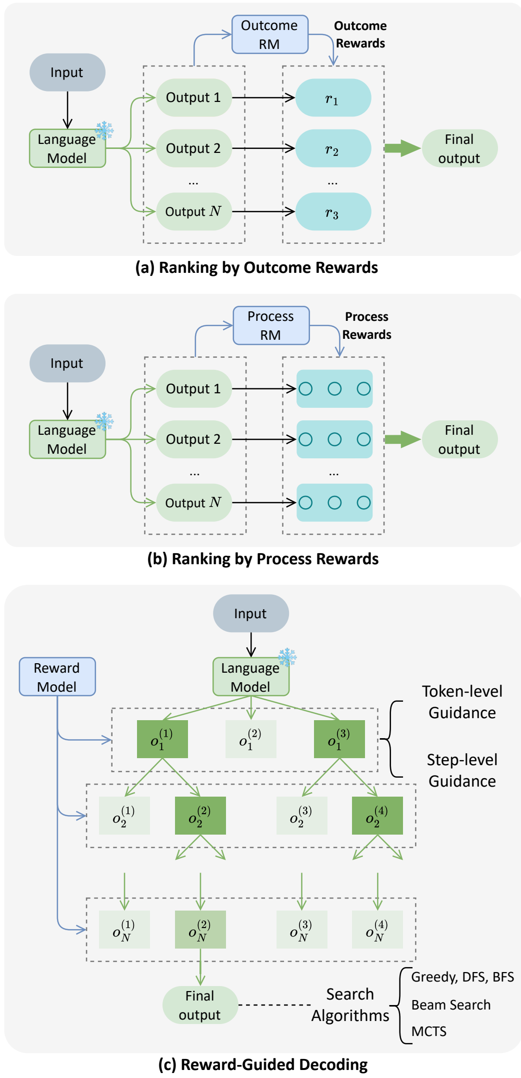

Figure 4: Illustrations of strategies for **Inference with Rewards**. (a,b): Generate-then-rank with outcome and process rewards. (c): Reward-guided decoding at the token and step level with search algorithms.

4.1 Generate-then-Rank

The generate-then-rank approach, usually referred to as *Best-of-N*, easily scales test-time compute to improve model outputs. It samples a number of candidate responses from the language model, scores them with a reward model, and selects the best one as the final output by ranking or voting (Wang et al., 2022). Based on the reward granularity, we distinguish two kinds of methods: (i) *ranking by outcome rewards* and (ii) *ranking by process rewards* as shown in Figure 4(a,b).

Ranking by Outcome Rewards. As shown in Figure 4(a), this method adopts an outcome reward model (ORM) to assess the holistic quality of candidate responses. Early work by Cobbe et al. (2021) trains a binary ORM to evaluate the correctness of candidate math solutions and selects the top-ranked one as the final output. Uesato et al. (2022) adopt the same idea and conduct comprehensive experiments on ranking outputs by ORMs. *LEVER* (Ni et al., 2023) trains a binary classifier as the ORM with code execution results as supervision. *V-STaR* (Hosseini et al., 2024) trains a verifier as the ORM on preference pairs through DPO to rank candidate math/code solutions during inference. *GenRM* (Zhang et al., 2024c) follows a generative way using the token generation probability. *Fast Best-of-N* (Sun et al., 2024a) accelerates this following a speculative rejection scheme.

Ranking by Process Rewards. As aforementioned, outcome reward models may struggle to discern the nuance among candidate responses. Thus many methods adopt process reward models (PRMs) for the generate-then-rank strategy. These methods score intermediate steps of candidate responses through a PRM and aggregate these step-

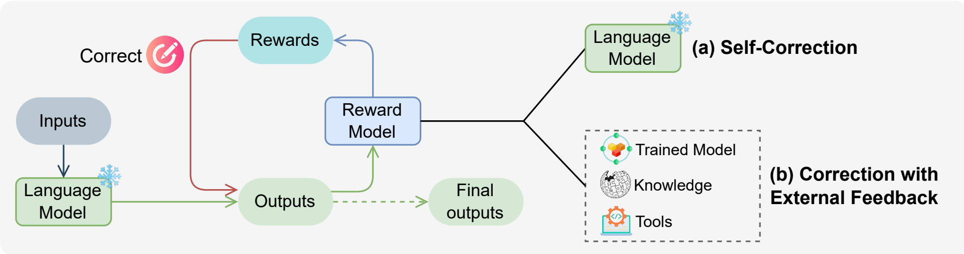

Figure 5: Illustration of **Post-Inference with Rewards**. (a): Self-Correction, using the language model itself. (b): Correction with External Feedback, such as trained model, external knowledge, and external tools.

level scores through multiplication or minimum to compute an overall score for ranking or voting (Zhang et al., 2025g). Early work by Lightman et al. (2023) trains a PRM to rank candidate math solutions by the product of their step-level scores. More extensions are also proposed, including *DIVERSE* (Li et al., 2023b), *Math-Shepherd* (Wang et al., 2023b), and *VisualPRM* (Wang et al., 2025c).

4.2 Reward-Guided Decoding

The above generate-then-rank decouples generation from evaluation; In contrast, *reward-guided decoding* tightly incorporates reward signals to guide the generation of language models. Figure 5(c) shows that it guides the language model's *token-level* or *step-level* decoding based on the reward signals through a search algorithm, such as greedy search, beam search, or MCTS. This enables fine-grained control over output quality and can foster reasoning and planning abilities.

Token-level Guidance. Token-level guidance steers language model generation by incorporating reward signals into the token decoding. This strategy commonly combines the token likelihoods with the reward signals from a reward model to select the next token, such as *RAD* (Deng and Raffel, 2023), *ARGS* (Khanov et al., 2024), *PG-TD* (Zhang et al., 2023c), *ARM* (Troshin et al., 2024), and *FaRMA* (Rashid et al., 2025).

Step-level Guidance. Beyond token-level guidance, step-level guidance operates on intermediate steps of generation. Figure 4(d) shows the generation is decomposed into multiple intermediate steps. At each step, a search algorithm, such as beam search and MCTS, explores the output space and selects appropriate steps guided by reward signals. This mechanism enables the model to recover from earlier errors and enhance reasoning. Representative work includes *GRACE* (Khalifa et al., 2023), Xie et al. (2023), Snell et al. (2025), *ORPS* (Yu et al., 2024b) and *RSD* (Liao et al., 2025). Some studies guide the decoding based on the step-level value, *i.e.,* cumulative future rewards, such as *Tree-of-Thoughts* (Yao et al., 2023) and *OVM* (Yu et al., 2023a). Other methods use reward signals to guide MCTS, including *RAP* (Hao et al., 2023), *STILL-1* (Jiang et al., 2024), and *rStar* (Qi et al., 2024). Several extensions leverage process reward models to precisely guide MCTS, such as *ReST-MCTS** (Zhang et al., 2024a), *LE-MCTS* (Park et al., 2024), and *rStar-Math* (Guan et al., 2025).

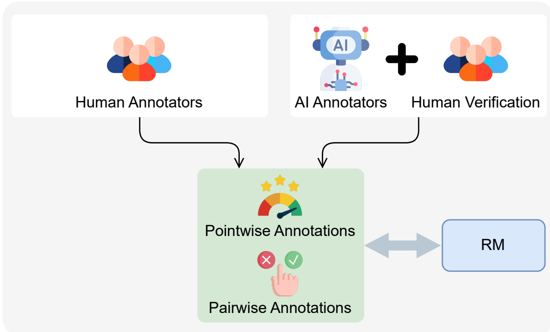

Figure 6: Illustration of Benchmarking Reward Models. Pointwise or pairwise annotations originate from human annotators or AI annotators with human verification.

5 Post-Inference with Rewards

Post-inference with rewards aims to correct and refine the model outputs after they have been generated. This approach enables iterative enhancement without updating model parameters, offering a lightweight and compatible mechanism for *test-time scaling*. It commonly incorporates critique rewards as augmented contexts to revise outputs, which provide fine-grained signals for correction, such as error locations and revision suggestions. We categorize these methods into two kinds: *Self-Correction* and *Correction with external rewards*.

5.1 Self-Correction

Figure 5(a) depicts that self-correction leverages the language model itself as a generative reward model to evaluate and revise its own outputs, similar to the aforementioned self-rewarding strategy. Early work *Self-Refine* (Madaan et al., 2023) follows this design. Similarly, *Reflexion* (Shinn et al., 2023) generates reflection feedback through the language model itself. It additionally maintains a memory bank to store previous feedback, outputs, and scalar feedback from evaluation metrics. *CoVe* (Dhuliawala et al., 2023) prompts the language model to generate and answer verification questions to identify factual errors in its own outputs. Others train the language model to improve its self-correction capability, such as *SCoRE* (Kumar et al., 2024) and *RISE* (Qu et al., 2024).

5.2 Correction with External Feedback

Prior studies argue that general language models struggle to identify and correct their errors without external feedback (Huang et al., 2023; Kamoi et al., 2024; Madaan et al., 2023; Pan et al., 2023b). Owing to this, increasing attention has been devoted to incorporating external feedback as reward signals as shown in Figure 5(b). We classify these works based on the feedback source: *trained models*, *external knowledge*, and *external tools*.

Trained Model. Many methods rely on more capable trained models (commonly referred to as critic models) to provide feedback as reward signals. The early work *CodeRL* (Le et al., 2022) uses a trained critic model to predict the functional correctness of the generated code. Following this, various studies are proposed, for instance, Welleck et al. (2022) for toxicity control; *RL4F* (Akyurek et al., 2023) for summarization; *Shepherd* (Wang et al., 2023c) and *A2R* (Lee et al., 2024) for factuality; *CTRL* (Xie et al., 2025b) and *CriticGPT* McAleese et al. (2024) for code generation. Moreover, some studies focus on step-level feedback for correction, such as *REFINER* (Paul et al., 2023) and *AutoMathCritique* (Xi et al., 2024). Others follow the multi-agent debate design, where critiques from peer agents support reflection and improvement, such as *MAD* (Liang et al., 2023), Cohen et al. (2023), and Du et al. (2023).

External Knowledge and Tools. External knowledge mainly provides factual critiques along with retrieved evidence to improve factuality and reduce hallucinations. Several methods follow this idea, such as *RARR* (Gao et al., 2022), *ReFeed* (Yu et al., 2023c), *LLM-Augmenter* (Peng et al., 2023), Varshney et al. (2023), and *FACTOOL* (Chern et al., 2023). External tools can execute and verify the model outputs, and their feedback can work as reward signals for correction. A primary tool is code compilers. They provide execution feedback to guide the refinement, such as *Self-Edit* (Zhang et al., 2023a) and *Self-Evolve* (Jiang et al., 2023). *Self-Debug* (Chen et al., 2023) and *CYCLE* (Ding et al., 2024) extend them with more feedback, for instance, unit test results and program explanations. Other tools can provide feedback, such as logic reasoner (Pan et al., 2023a), symbolic interpreter (Qiu et al., 2023), proof checker (First et al., 2023), search engines (Gou et al., 2023; Kim et al., 2023).

6 Benchmarking Reward Models

Rigorous and diverse benchmarks are essential for evaluating the performance of reward models. As illustrated in Figure 6, recent benchmarks primarily rely on human annotators or AI annotators followed by human verification. The resulting annotations are mainly pointwise (*e.g.,* scalar scoring) or pairwise (*e.g.,* selecting the preferred response given two options). *RewardBench* (Lambert et al., 2024) is the first comprehensive benchmarks for reward models. It aggregates preference data from existing datasets to evaluate reward model performance in chatting, reasoning, and safety. *RM-Bench* (Liu et al., 2024d) and *RMB* (Zhou et al., 2024a) extend it to more scenarios. Some focus on PRMs, like *ProcessBench* (Zheng et al., 2024), *MR-Ben* (Zeng et al., 2024), and *PRMBench* (Song et al., 2025b).

Due to the page limitation, we discuss more about benchmarks, applications, challenges, and future directions in Appendices A to C.

7 Conclusion

We comprehensively survey the emerging paradigm of *learning from rewards*. We introduce its landscape from three key stages: training, inference, and post-inference, each reflecting a distinct way to integrate reward signals into steering LLMs' behavior. In addition, we summarize recent progress in benchmarking reward models and applications. Finally we identify core challenges and outline promising future directions. We hope this survey provides a structured understanding of the field and inspires future research.

Limitations

This paper comprehensively surveys the emerging paradigm of learning from rewards in the post-training and test-time scaling of LLMs, but we believe there are some limitations:

- Due to the page constraints, we cannot cover the full technical details of all methods. Based on our comprehensive survey, we encourage interested readers to refer to the original papers for in-depth explanations and implementation specifics.
- We primarily focus on the representative methods and recent trends associated with learning from rewards. As a result, we may omit some earlier approaches and domain-specific techniques.

References

David Abel, André Barreto, Benjamin Van Roy, Doina Precup, Hado P van Hasselt, and Satinder Singh. 2023. A definition of continual reinforcement learning. *Advances in Neural Information Processing Systems*, 36:50377–50407.

Daechul Ahn, Yura Choi, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. 2024. Tuning large multimodal models for videos using reinforcement learning from ai feedback. *arXiv preprint arXiv:2402.03746*.

Afra Feyza Akyurek, Ekin Akyurek, Ashwin Kalyan, Peter Clark, Derry Tanti Wijaya, and Niket Tandon. 2023. RL4F: Generating natural language feedback with reinforcement learning for repairing model outputs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7716–7733, Toronto, Canada. Association for Computational Linguistics.

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

Anthropic. 2025. Introducing deep research.

Alisson Azzolini, Hannah Brandon, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, Francesco Ferroni, Rama Govindaraju, et al. 2025. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

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

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Michael Bowling and Esraa Elelimy. 2025. Rethinking the foundations for continual reinforcement learning. *arXiv preprint arXiv:2504.08161*.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Maosong Cao, Alexander Lam, Haodong Duan, Hongwei Liu, Songyang Zhang, and Kai Chen. 2024. Compassjudger-1: All-in-one judge model helps model evaluation and evolution. *arXiv preprint arXiv:2410.16256*.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. 2025a. R1-v: Reinforcing super generalization ability in vision-language models with less than $3.

Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. 2025b. Research: Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*.

Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, et al. 2024b. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? *arXiv preprint arXiv:2407.04842*.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

Sanjiban Choudhury. 2025. Process reward models for llm agents: Practical framework and directions. *arXiv preprint arXiv:2502.10325*.

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

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. Lm vs lm: Detecting factual errors via cross examination. *arXiv preprint arXiv:2305.13281*.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, et al. 2025. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*.

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

DeepSeek-AI. 2025. Deepseek-prover-v2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition. *arXiv preprint arXiv:2504.21801*.

DeepSeek-AI et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Haikang Deng and Colin Raffel. 2023. Reward-augmented decoding: Efficient controlled text generation with a unidirectional reward model. *arXiv preprint arXiv:2310.09520*.

Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, Ryan Soklaski, Alex Tamkin, Jared Kaplan, et al. 2024. Sycophancy to subterfuge: Investigating reward-tampering in large language models. *arXiv preprint arXiv:2406.10162*.

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

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Yangruibo Ding, Marcus J Min, Gail Kaiser, and Baishakhi Ray. 2024. Cycle: Learning to self-refine the code generation. *Proceedings of the ACM on Programming Languages*, 8(OOPSLA1):392–418.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*.

Shihan Dou, Yan Liu, Haoxiang Jia, Limao Xiong, Enyu Zhou, Wei Shen, Junjie Shan, Caishuang Huang, Xiao Wang, Xiaoran Fan, et al. 2024. Stepcoder: Improve code generation with reinforcement learning from compiler feedback. *arXiv preprint arXiv:2402.01391*.

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

Keyu Duan, Zichen Liu, Xin Mao, Tianyu Pang, Changyu Chen, Qiguang Chen, Michael Qizhe Shieh, and Longxu Dou. 2025. Efficient process reward model training via active learning. *arXiv preprint arXiv:2504.10559*.

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.

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

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. 2021. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198(Suppl 27):6435–6467.

Hugging Face. 2025. Open r1: A fully open reproduction of deepseek-r1.

Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. 2025a. Retool: Reinforcement learning for strategic tool use in llms. *arXiv preprint arXiv:2504.11536*.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. 2025b. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*.

Zihao Feng, Xiaoxue Wang, Ziwei Bai, Donghang Su, Bowen Wu, Qun Yu, and Baoxun Wang. 2025c. Improving generalization in intent detection: Grpo with reward-based curriculum sampling. *arXiv preprint arXiv:2504.13592*.

Emily First, Markus N Rabe, Talia Ringer, and Yuriy Brun. 2023. Baldur: Whole-proof generation and repair with large language models. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 1229–1241.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios N Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024. How to evaluate reward models for rlhf. *arXiv preprint arXiv:2410.14872*.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, et al. 2024. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2022. Rarr: Researching and revising what language models say, using language models. *arXiv preprint arXiv:2210.08726*.

Minghe Gao, Xuqi Liu, Zhongqi Yue, Yang Wu, Shuang Chen, Juncheng Li, Siliang Tang, Fei Wu, Tat-Seng Chua, and Yueting Zhuang. 2025. Benchmarking multimodal cot reward model stepwise by visual program. *arXiv preprint arXiv:2504.06606*.

Jonas Gehring, Kunhao Zheng, Jade Copet, Vegard Mella, Taco Cohen, and Gabriel Synnaeve. 2024. RLEF: grounding code llms in execution feedback with reinforcement learning. *arXiv preprint arXiv:2410.02089*.

Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin J. Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Sona Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Anna Goldie, Azalia Mirhoseini, Hao Zhou, Irene Cai, and Christopher D Manning. 2025. Synthetic data generation & multi-step rl for reasoning & tool use. *arXiv preprint arXiv:2504.04736*.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alexa Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, A. Doucet, Orhan Firat, and Nando de Freitas. 2023. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*.

Jiaxin Guo, Zewen Chi, Li Dong, Qingxiu Dong, Xun Wu, Shaohan Huang, and Furu Wei. 2025a. Reward reasoning model. *arXiv preprint arXiv:2505.14674*.

Yanjiang Guo, Jianke Zhang, Xiaoyu Chen, Xiang Ji, Yen-Jen Wang, Yucheng Hu, and Jianyu Chen. 2025b. Improving vision-language-action model with online reinforcement learning. *arXiv preprint arXiv:2501.16664*.

Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. 2025c. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*.

Srishti Gureja, Lester James V. Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. 2024. M-rewardbench: Evaluating reward models in multilingual settings. *arXiv preprint arXiv:2410.15522*.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. 2024. V-star: Training verifiers for self-taught reasoners. *arXiv preprint arXiv:2402.06457*.

Hui Huang, Yancheng He, Hongli Zhou, Rui Zhang, Wei Liu, Weixun Wang, Wenbo Su, Bo Zheng, and Jiaheng Liu. 2025a. Think-j: Learning to think for generative llm-as-a-judge. *arXiv preprint arXiv:2505.14268*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Yao Hu, and Shaohui Lin. 2025b. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Erik Jenner and Adam Gleave. 2022. Preprocessing reward functions for interpretability. *arXiv preprint arXiv:2203.13553*.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704.

Jinhao Jiang, Zhipeng Chen, Yingqian Min, Jie Chen, Xiaoxue Cheng, Jiapeng Wang, Yiru Tang, Haoxiang Sun, Jia Deng, Wayne Xin Zhao, et al. 2024. Enhancing llm reasoning with reward-guided tree search. *arXiv preprint arXiv:2411.11694*.

Pengcheng Jiang, Jiacheng Lin, Lang Cao, Runchu Tian, SeongKu Kang, Zifeng Wang, Jimeng Sun, and Jiawei Han. 2025. Deepretrieval: Hacking real search engines and retrievers with large language models via reinforcement learning. *arXiv preprint arXiv:2503.00223*.

Shuyang Jiang, Yuhao Wang, and Yu Wang. 2023. Self-evolve: A code evolution framework via large language models. *arXiv preprint arXiv:2306.02907*.

Fangkai Jiao, Geyang Guo, Xingxing Zhang, Nancy F Chen, Shafiq Joty, and Furu Wei. 2024a. Preference optimization for reasoning with pseudo feedback. *arXiv preprint arXiv:2411.16345*.

Fangkai Jiao, Chengwei Qin, Zhengyuan Liu, Nancy F Chen, and Shafiq Joty. 2024b. Learning planning-based reasoning by trajectories collection and process reward synthesizing. *arXiv preprint arXiv:2402.00658*.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Zhuoran Jin, Hongbang Yuan, Tianyi Men, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Rag-rewardbench: Benchmarking reward models in retrieval augmented generation for preference alignment. *arXiv preprint arXiv:2412.13746*.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

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. *arXiv preprint arXiv:2001.08361*.

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

Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Grace: Discriminator-guided chain-of-thought reasoning. *arXiv preprint arXiv:2305.14934*.

Maxim Khanov, Jirayu Burapacheep, and Yixuan Li. 2024. Args: Alignment as reward-guided search. In *The Twelfth International Conference on Learning Representations*.

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. Language models can solve computer tasks. *Advances in Neural Information Processing Systems*, 36:39648–39677.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. 2024. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*.

Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. 2023. Reward design with language models. *arXiv preprint arXiv:2303.00001*.

Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. 2024. Step-DPO: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofeng Yang. 2025. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*.

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

Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. 2022. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:21314–21328.

Dongyub Lee, Eunhwan Park, Hodong Lee, and Heui-Seok Lim. 2024. Ask, assess, and refine: Rectifying factual consistency and hallucination in llms with metric-guided feedback learning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2422–2433.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2023. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. *arXiv preprint arXiv:2309.00267*.

Bolian Li, Yifan Wang, Ananth Grama, and Ruqi Zhang. 2024a. Cascade reward sampling for efficient decoding-time alignment. *arXiv preprint arXiv:2406.16306*.

Jiazheng Li, Yuxiang Zhou, Junru Lu, Gladys Tyen, Lin Gui, Cesare Aloisi, and Yulan He. 2025a. Two heads are better than one: Dual-model verbal reflection at inference-time. *arXiv preprint arXiv:2502.19230*.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023a. Generative judge for evaluating alignment. *arXiv preprint arXiv:2310.05470*.

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, et al. 2024b. Vlrewardbench: A challenging benchmark for vision-language generative reward models. *arXiv preprint arXiv:2411.17451*.

Lihe Li, Ruotong Chen, Ziqian Zhang, Zhichao Wu, Yi-Chen Li, Cong Guan, Yang Yu, and Lei Yuan. 2024c. Continual multi-objective reinforcement learning via reward model rehearsal. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 4434–4442.

Lin Li, Wei Chen, Jiahui Li, and Long Chen. 2025b. Relation-r1: Cognitive chain-of-thought guided reinforcement learning for unified relational comprehension. *arXiv preprint arXiv:2504.14642*.

Ming Li, Shitian Zhao, Jike Zhong, Yuxiang Lai, and Kaipeng Zhang. 2025c. Cls-rl: Image classification with rule-based reinforcement learning. *arXiv preprint arXiv:2503.16188*.

Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. 2025d. Q-insight: Understanding image quality via visual reinforcement learning. *arXiv preprint arXiv:2503.22679*.

Wendi Li and Yixuan Li. 2024. Process reward model with q-value rankings. *arXiv preprint arXiv:2410.11287*.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. 2025e. Webthinker: Empowering large reasoning models with deep research capability.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. 2025f. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. 2025g. Torl: Scaling tool-integrated rl. *arXiv preprint arXiv:2503.23383*.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023b. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. 2024. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19401–19411.

Baohao Liao, Yuhui Xu, Hanze Dong, Junnan Li, Christof Monz, Silvio Savarese, Doyen Sahoo, and Caiming Xiong. 2025. Reward-guided speculative decoding for efficient llm reasoning. *arXiv preprint arXiv:2501.19324*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Jiacheng Lin, Tian Wang, and Kun Qian. 2025. Rec-r1: Bridging generative large language models and user-centric recommendation systems via reinforcement learning. *arXiv preprint arXiv:2503.24289*.

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Scott Yih, and Xilun Chen. 2024a. Flame: Factuality-aware alignment for large language models. *Advances in Neural Information Processing Systems*, 37:115588–115614.

Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. 2024b. Criticbench: Benchmarking llms for critique-correct reasoning. *arXiv preprint arXiv:2402.14809*.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024a. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*.

Fangfu Liu, Hanyang Wang, Yimo Cai, Kaiyan Zhang, Xiaohang Zhan, and Yueqi Duan. 2025a. Video-t1: Test-time scaling for video generation. *arXiv preprint arXiv:2503.18942*.

Jiate Liu, Yiqin Zhu, Kaiwen Xiao, Qiang Fu, Xiao Han, Wei Yang, and Deheng Ye. 2023. RLTF: reinforcement learning from unit test feedback. *Trans. Mach. Learn. Res.*, 2023.

Tianqi Liu, Wei Xiong, Jie Ren, Lichang Chen, Junru Wu, Rishabh Joshi, Yang Gao, Jiaming Shen, Zhen Qin, Tianhe Yu, Daniel Sohn, Anastasiia Makarova, Jeremiah Liu, Yuan Liu, Bilal Piot, Abe Ittycheriah, Aviral Kumar, and Mohammad Saleh. 2025b. Rrm: Robust reward model training mitigates reward hacking. *arXiv preprint arXiv:2409.13156*.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. 2024b. Statistical rejection sampling improves preference optimization. In *The Twelfth International Conference on Learning Representations*.

Wei Liu, Junlong Li, Xiwen Zhang, Fan Zhou, Yu Cheng, and Junxian He. 2024c. Diving into self-evolving training for multimodal reasoning. *arXiv preprint arXiv:2412.17451*.

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

Yuhang Liu, Pengxiang Li, Congkai Xie, Xavier Hu, Xiaotian Han, Shengyu Zhang, Hongxia Yang, and Fei Wu. 2025c. Infigui-r1: Advancing multimodal gui agents from reactive actors to deliberative reasoners. *arXiv preprint arXiv:2504.14239*.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. 2025d. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*.

Zhaowei Liu, Xin Guo, Fangqi Lou, Lingfeng Zeng, Jinyi Niu, Zixuan Wang, Jiajie Xu, Weige Cai, Ziwei Yang, Xueqian Zhao, et al. 2025e. Fin-r1: A large language model for financial reasoning through reinforcement learning. *arXiv preprint arXiv:2503.16252*.

Zhiyuan Liu, Yuting Zhang, Feng Liu, Changwang Zhang, Ying Sun, and Jun Wang. 2025f. Othink-mr1: Stimulating multimodal generalized reasoning capabilities through dynamic reinforcement learning. *arXiv preprint arXiv:2503.16081*.

Zihan Liu, Yang Chen, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024e. Acemath: Advancing frontier math reasoning with post-training and reward modeling. *arXiv preprint arXiv:2412.15084*.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025g. Inference-time scaling for generalist reward modeling. *arXiv preprint arXiv:2504.02495*.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025h. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.

Zhengxi Lu, Yuxiang Chai, Yaxuan Guo, Xi Yin, Liang Liu, Hao Wang, Guanjing Xiong, and Hongsheng Li. 2025. Ui-r1: Enhancing action prediction of gui agents by reinforcement learning. *arXiv preprint arXiv:2503.21620*.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.

Haoran Luo, Yikai Guo, Qika Lin, Xiaobao Wu, Xinyu Mu, Wenhao Liu, Meina Song, Yifan Zhu, Luu Anh Tuan, et al. 2025. Kbqa-o1: Agentic knowledge base question answering with monte carlo tree search. *arXiv preprint arXiv:2501.18922*.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, et al. 2024. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2.

Chengqi Lyu, Songyang Gao, Yuzhe Gu, Wenwei Zhang, Jianfei Gao, Kuikun Liu, Ziyi Wang, Shuaibin Li, Qian Zhao, Haian Huang, et al. 2025. Exploring the limit of outcome reward for learning mathematical reasoning. *arXiv preprint arXiv:2502.06781*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*.

Peixian Ma, Xialie Zhuang, Chengjin Xu, Xuhui Jiang, Ran Chen, and Jian Guo. 2025. Sql-r1: Training natural language to sql reasoning model by reinforcement learning. *arXiv preprint arXiv:2504.08600*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

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

Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. 2024. Llm critics help catch llm bugs. *arXiv preprint arXiv:2407.00215*.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. 2025. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.

Meta. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2307.09288*.

Meta. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Ansong Ni, Srini Iyer, Dragomir Radev, Veselin Stoyanov, Wen-tau Yih, Sida Wang, and Xi Victoria Lin. 2023. Lever: Learning to verify language-to-code generation with execution. In *International Conference on Machine Learning*, pages 26106–26128. PMLR.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2025. Introducing deep research. openai.com.

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.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. 2022. The effects of reward misspecification: Mapping and mitigating misaligned models. *arXiv preprint arXiv:2201.03544*.

Alexander Pan, Erik Jones, Meena Jagadeesan, and Jacob Steinhardt. 2024a. Feedback loops with language models drive in-context reward hacking. *arXiv preprint arXiv:2402.06627*.

Jane Pan, He He, Samuel R Bowman, and Shi Feng. 2024b. Spontaneous reward hacking in iterative self-refinement. *arXiv preprint arXiv:2407.04549*.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. 2023a. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *arXiv preprint arXiv:2305.12295*.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023b. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*.

Sungjin Park, Xiao Liu, Yeyun Gong, and Edward Choi. 2024. Ensembling large language models with process reward-guided tree search for better complex reasoning. *arXiv preprint arXiv:2412.15797*.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2023. Refiner: Reasoning feedback on intermediate representations. *arXiv preprint arXiv:2304.01904*.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Hao Peng, Yunjia Qi, Xiaozhi Wang, Zijun Yao, Bin Xu, Lei Hou, and Juanzi Li. 2025. Agentic reward modeling: Integrating human preferences with verifiable correctness signals for reliable reward systems. *arXiv preprint arXiv:2502.19328*.

Archiki Prasad, Weizhe Yuan, Richard Yuanzhe Pang, Jing Xu, Maryam Fazel-Zarandi, Mohit Bansal, Sainbayar Sukhbaatar, Jason Weston, and Jane Yu. 2024. Self-consistency preference optimization.

Zhenting Qi, Mingyuan Ma, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. 2024. Mutual reasoning makes smaller llms stronger problem-solvers. *arXiv preprint arXiv:2408.06195*.

Cheng Qian, Emre Can Acikgoz, Qi He, Hongru Wang, Xiusi Chen, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. 2025. Toolrl: Reward is all tool learning needs. *arXiv preprint arXiv:2504.13958*.

Shuofei Qiao, Honghao Gui, Chengfei Lv, Qianghuai Jia, Huajun Chen, and Ningyu Zhang. 2023. Making language models better tool learners with execution feedback. *arXiv preprint arXiv:2305.13068*.

Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, et al. 2023. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. *arXiv preprint arXiv:2310.08559*.

Yuxiao Qu, Tianjun Zhang, Naman Garg, and Aviral Kumar. 2024. Recursive introspection: Teaching language model agents how to self-improve. *Advances in Neural Information Processing Systems*, 37:55249–55285.

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

Ahmad Rashid, Ruotian Wu, Rongqi Fan, Hongliang Li, Agustinus Kristiadi, and Pascal Poupart. 2025. Towards cost-effective reward guided text generation. *arXiv preprint arXiv:2502.04517*.

Noam Razin, Zixuan Wang, Hubert Strauss, Stanley Wei, Jason D Lee, and Sanjeev Arora. 2025. What makes a reward model a good teacher? an optimization perspective. *arXiv preprint arXiv:2503.15477*.

Manon Revel, Matteo Cargnelutti, Tyna Eloundou, and Greg Leppert. 2025. Seal: Systematic error analysis for value alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27599–27607.

Jiacheng Ruan, Wenzhen Yuan, Xian Gao, Ye Guo, Daoxin Zhang, Zhe Xu, Yao Hu, Ting Liu, and Yuzhuo Fu. 2025. VLRMBench: A comprehensive and challenging benchmark for vision-language reward models. *arXiv preprint arXiv:2503.07478*.

Jacob Russell and Eugene Santos. 2019. Explaining reward functions in markov decision processes. In *Proceedings of the Thirty-Second International Florida Artificial Intelligence Research Society Conference, Sarasota, Florida, USA, May 19-22 2019*, pages 56–61. AAAI Press.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

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

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. 2025a. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*.

Wei Shen, Guanlin Liu, Zheng Wu, Ruofei Zhu, Qingping Yang, Chao Xin, Yu Yue, and Lin Yan. 2025b. Exploring data scaling trends and effects in reinforcement learning from human feedback. *arXiv preprint arXiv:2503.22230*.

Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *arxiv preprint arXiv:2303.11366*.

David Silver and Richard S Sutton. 2025. Welcome to the era of experience. *Google AI*.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, volume 2, page 7.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025a. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. 2025b. Prmbench: A fine-grained and challenging benchmark for process-level reward models. *arXiv preprint arXiv:2501.03124*.

Hanshi Sun, Momin Haider, Ruiqi Zhang, Huitao Yang, Jiahao Qiu, Ming Yin, Mengdi Wang, Peter Bartlett, and Andrea Zanette. 2024a. Fast best-of-n decoding via speculative rejection. *arXiv preprint arXiv:2410.20290*.

Shichao Sun, Junlong Li, Weizhe Yuan, Ruifeng Yuan, Wenjie Li, and Pengfei Liu. 2024b. The critique of critique. *arXiv preprint arXiv:2401.04518*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. 2025. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*.

Leitian Tao, Xiang Chen, Tong Yu, Tung Mai, Ryan A. Rossi, Yixuan Li, and Saayan Mitra. 2024. Codelutra: Boosting LLM code generation via preference-guided refinement. *arXiv preprint arXiv:2411.05199*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2023. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations*.

Sergey Troshin, Vlad Niculae, and Antske Fokkens. 2024. Efficient controlled language generation with low-rank autoregressive reward models. *arXiv preprint arXiv:2407.04615*.

Haoqin Tu, Weitao Feng, Hardy Chen, Hui Liu, Xianfeng Tang, and Cihang Xie. 2025. Vilbench: A suite for vision-language process reward modeling. *arXiv preprint arXiv:2503.20271*.

Gladys Tyen, Hassan Mansoor, Victor Cărbune, Peter Chen, and Tony Mak. 2023. Llms cannot find reasoning errors, but can correct them given the error location. *arXiv preprint arXiv:2311.08516*.

Jonathan Uesato, Ramana Kumar, Victoria Krakovna, Tom Everitt, Richard Ngo, and Shane Legg. 2020. Avoiding tampering incentives in deep rl via decoupled approval. *arXiv preprint arXiv:2011.08827*.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023a. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS*.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024a. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*.

Hongru Wang, Cheng Qian, Wanjun Zhong, Xiusi Chen, Jiahao Qiu, Shijue Huang, Bowen Jin, Mengdi Wang, Kam-Fai Wong, and Heng Ji. 2025a. Otc: Optimal tool calls via reinforcement learning. *arXiv preprint arXiv:2504.14870*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024b. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Peifeng Wang, Austin Xu, Yilun Zhou, Caiming Xiong, and Shafiq Joty. 2024c. Direct judgement preference optimization. *arXiv preprint arXiv:2409.14664*.

Peiyi Wang, Lei Li, Zhihong Shao, RX Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2023b. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*.

Teng Wang, Zhangyi Jiang, Zhenqi He, Wenhan Yang, Yanan Zheng, Zeyu Li, Zifan He, Shenyang Tong, and Hailei Gong. 2025b. Towards hierarchical multi-step reward models for enhanced reasoning in large language models. *arXiv preprint arXiv:2503.13551*.

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

Tianlu Wang, Ping Yu, Xiaoqing Ellen Tan, Sean O'Brien, Ramakanth Pasunuru, Jane Dwivedi-Yu, Olga Golovneva, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023c. Shepherd: A critic for language model generation. *arXiv preprint arXiv:2308.04592*.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, et al. 2025c. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yibin Wang, Zhiyu Tan, Junyan Wang, Xiaomeng Yang, Cheng Jin, and Hao Li. 2024e. Lift: Leveraging human feedback for text-to-video model alignment. *arXiv preprint arXiv:2412.04814*.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. 2025d. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*.

Zhiqiang Wang, Pengbin Feng, Yanbin Lin, Shuzhang Cai, Zongao Bian, Jinghua Yan, and Xingquan Zhu. 2025e. Crowdvlm-r1: Expanding r1 ability to vision language model for crowd counting using fuzzy group relative policy reward. *arXiv preprint arXiv:2504.03724*.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, Monica Lam, Yiping Lu, Kyunghyun Cho, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, and Manling Li. 2025f. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning.

Yuxiang Wei, Olivier Duchenne, Jade Copet, Quentin Carbonneaux, Lingming Zhang, Daniel Fried, Gabriel Synnaeve, Rishabh Singh, and Sida I Wang. 2025. Swe-rl: Advancing llm reasoning via reinforcement learning on open software evolution. *arXiv preprint arXiv:2502.18449*.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2022. Generating sequences by learning to self-correct. *arXiv preprint arXiv:2211.00053*.

Xueru Wen, Xinyu Lu, Xinyan Guan, Yaojie Lu, Hongyu Lin, Ben He, Xianpei Han, and Le Sun. 2024. On-policy fine-grained knowledge feedback for hallucination mitigation. *arXiv preprint arXiv:2406.12221*.

Lilian Weng. 2024. Reward hacking in reinforcement learning. *lilianweng.github.io*.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024a. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. *arXiv preprint arXiv:2407.19594*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36:59008–59033.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024b. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862.

Zhiheng Xi, Dingwen Yang, Jixuan Huang, Jiafu Tang, Guanyu Li, Yiwen Ding, Wei He, Boyang Hong, Shihan Do, Wenyu Zhan, et al. 2024. Enhancing llm reasoning via critique models with test-time and training-time supervision. *arXiv preprint arXiv:2411.16579*.

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. 2024. Evaluating mathematical reasoning beyond accuracy. *arXiv preprint arXiv:2404.05692*.

Yu Xia, Jingru Fan, Weize Chen, Siyu Yan, Xin Cong, Zhong Zhang, Yaxi Lu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2025. Agentrm: Enhancing agent generalization with reward modeling. *arXiv preprint arXiv:2502.18407*.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025a. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, James Xu Zhao, Min-Yen Kan, Junxian He, and Michael Xie. 2023. Self-evaluation guided beam search for reasoning. *Advances in Neural Information Processing Systems*, 36:41618–41650.

Zhihui Xie, Liyu Chen, Weichao Mao, Jingjing Xu, Lingpeng Kong, et al. 2025b. Teaching language models to critique via reinforcement learning. *arXiv preprint arXiv:2502.03492*.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2024. Llava-critic: Learning to evaluate multimodal models. *arXiv preprint arXiv:2410.02712*.

Wei Xiong, Hanning Zhang, Chenlu Ye, Lichang Chen, Nan Jiang, and Tong Zhang. 2025. Self-rewarding correction for mathematical reasoning. *arXiv preprint arXiv:2502.19613*.

Huimin Xu, Xin Mao, Feng-Lin Li, Xiaobao Wu, Wang Chen, Wei Zhang, and Anh Tuan Luu. 2025a. Full-step-dpo: Self-supervised preference optimization with step-wise rewards for mathematical reasoning. *arXiv preprint arXiv:2502.14356*.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935.

Yixuan Even Xu, Yash Savani, Fei Fang, and Zico Kolter. 2025b. Not all rollouts are useful: Down-sampling rollouts in llm reinforcement learning. *arXiv preprint arXiv:2504.13818*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Michihiro Yasunaga, Luke Zettlemoyer, and Marjan Ghazvininejad. 2025. Multimodal rewardbench: Holistic evaluation of reward models for vision language models. *arXiv preprint arXiv:2502.14191*.

Ziyi Ye, Xiangsheng Li, Qiuchi Li, Qingyao Ai, Yujia Zhou, Wei Shen, Dong Yan, and Yiqun Liu. 2024. Beyond scalar reward model: Learning generative judge from preference data. *arXiv preprint arXiv:2410.03742*.

Fei Yu, Anningzhe Gao, and Benyou Wang. 2023a. Ovm, outcome-supervised value models for planning in mathematical reasoning. *arXiv preprint arXiv:2311.09724*.

Jiachen Yu, Shaoning Sun, Xiaohui Hu, Jiaxu Yan, Kaidong Yu, and Xuelong Li. 2025a. Improve llm-as-a-judge ability as a general ability. *arXiv preprint arXiv:2502.11689*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. 2025b. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2023b. RLHF-V: towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*.

Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. 2023c. Improving language models via plug-and-play retrieval feedback. *arXiv preprint arXiv:2305.14002*.

Yue Yu, Zhengxing Chen, Aston Zhang, Liang Tan, Chenguang Zhu, Richard Yuanzhe Pang, Yundi Qian, Xuewei Wang, Suchin Gururangan, Chao Zhang, Melanie Kambadur, Dhruv Mahajan, and Rui Hou. 2024a. Self-generated critiques boost reward modeling for language models. *arXiv preprint arXiv:2411.16646*.

Zhuohao Yu, Weizheng Gu, Yidong Wang, Zhengran Zeng, Jindong Wang, Wei Ye, and Shikun Zhang. 2024b. Outcome-refining process supervision for code generation. *arXiv preprint arXiv:2412.15118*.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. 2024a. Free process rewards without process labels. *arXiv preprint arXiv:2412.01981*.

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

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023a. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*.

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

Zhongshen Zeng, Pengguang Chen, Shu Liu, Haiyun Jiang, and Jiaya Jia. 2023. Mr-gsm8k: A meta-reasoning benchmark for large language model evaluation. *arXiv preprint arXiv:2312.17080*.

Zhongshen Zeng, Yinhong Liu, Yingjia Wan, Jingyao Li, Pengguang Chen, Jianbo Dai, Yuxuan Yao, Rongwu Xu, Zehan Qi, Wanru Zhao, et al. 2024. Mr-ben: A meta-reasoning benchmark for evaluating system-2 thinking in llms. *arXiv preprint arXiv:2406.13975*.

Yufei Zhan, Yousong Zhu, Shurong Zheng, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. 2025. Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. *arXiv preprint arXiv:2503.18013*.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024a. Rest-mcts*: Llm self-training via process reward guided tree search. *Advances in Neural Information Processing Systems*, 37:64735–64772.

Han Zhang, Yu Lei, Lin Gui, Min Yang, Yulan He, Hui Wang, and Ruifeng Xu. 2024b. Cppo: Continual learning for reinforcement learning with human feedback. In *The Twelfth International Conference on Learning Representations*.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025a. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*.

Kechi Zhang, Zhuo Li, Jia Li, Ge Li, and Zhi Jin. 2023a. Self-edit: Fault-aware code editor for code generation. *arXiv preprint arXiv:2305.04087*.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2024c. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023b. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. 2025b. Right question is already half the answer: Fully unsupervised llm reasoning incentivization. *arXiv preprint arXiv:2504.05812*.

Shimao Zhang, Xiao Liu, Xin Zhang, Junxiao Liu, Zheheng Luo, Shujian Huang, and Yeyun Gong. 2025c. Process-based self-rewarding language models. *arXiv preprint arXiv:2503.03746*.

Shun Zhang, Zhenfang Chen, Yikang Shen, Mingyu Ding, Joshua B Tenenbaum, and Chuang Gan. 2023c. Planning with large language models for code generation. *arXiv preprint arXiv:2303.05510*.

Wenqi Zhang, Mengna Wang, Gangao Liu, Xu Huixin, Yiwei Jiang, Yongliang Shen, Guiyang Hou, Zhe Zheng, Hang Zhang, Xin Li, et al. 2025d. Embodied-reasoner: Synergizing visual search, reasoning, and action for embodied interactive tasks. *arXiv preprint arXiv:2503.21696*.

Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. 2025e. Tinyllava-video-r1: Towards smaller lmms for video reasoning. *arXiv preprint arXiv:2504.09641*.

Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, et al. 2025f. Mm-rlhf: The next step forward in multimodal llm alignment. *arXiv preprint arXiv:2502.10391*.

Yudi Zhang, Yali Du, Biwei Huang, Ziyan Wang, Jun Wang, Meng Fang, and Mykola Pechenizkiy. 2023d. Interpretable reward redistribution in reinforcement learning: A causal approach. *Advances in Neural Information Processing Systems*, 36:20208–20229.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025g. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*.

Baining Zhao, Ziyou Wang, Jianjie Fang, Chen Gao, Fanhang Man, Jinqiang Cui, Xin Wang, Xinlei Chen, Yong Li, and Wenwu Zhu. 2025a. Embodied-r: Collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning. *arXiv preprint arXiv:2504.12680*.

Jian Zhao, Runze Liu, Kaiyan Zhang, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, et al. 2025b. Genprm: Scaling test-time compute of process reward models via generative reasoning. *arXiv preprint arXiv:2504.00891*.

Shuai Zhao, Linchao Zhu, and Yi Yang. 2025c. Learning from reference answers: Versatile language model alignment without binary human preference data. *arXiv preprint arXiv:2504.09895*.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2024. Processbench: Identifying process errors in mathematical reasoning. *arXiv preprint arXiv:2412.06559*.

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

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. *arXiv preprint arXiv:2504.03160*.

Changzhi Zhou, Xinyu Zhang, Dandan Song, Xiancai Chen, Wanli Gu, Huipeng Ma, Yuhang Tian, Mengdi Zhang, and Linmei Hu. 2025a. Refinecoder: Iterative improving of large language models via adaptive critique refinement for code generation. *arXiv preprint arXiv:2502.09183*.

Enyu Zhou, Guodong Zheng, Binghai Wang, Zhiheng Xi, Shihan Dou, Rong Bao, Wei Shen, Limao Xiong, Jessica Fan, Yurong Mou, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024a. Rmb: Comprehensively benchmarking reward models in llm alignment. *arXiv preprint arXiv:2410.09893*.

Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. 2025b. R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*.

Zihao Zhou, Shudong Liu, Maizhen Ning, Wei Liu, Jindong Wang, Derek F Wong, Xiaowei Huang, Qiufeng Wang, and Kaizhu Huang. 2024b. Is your model really a good math reasoner? evaluating mathematical reasoning with checklist. *arXiv preprint arXiv:2407.08733*.

Jie Zhu, Qian Chen, Huaixia Dou, Junhui Li, Lifan Guo, Feng Chen, and Chi Zhang. 2025. Dianjin-r1: Evaluating and enhancing financial reasoning in large language models.

Qihao Zhu, Daya Guo, Zhihong Shao, Dejian Yang, Peiyi Wang, Runxin Xu, Y Wu, Yukun Li, Huazuo Gao, Shirong Ma, et al. 2024. Deepseek-coder-v2: Breaking the barrier of closed-source models in code intelligence. *arXiv preprint arXiv:2406.11931*.

Xinyu Zhu, Junjie Wang, Lin Zhang, Yuxiang Zhang, Ruyi Gan, Jiaxing Zhang, and Yujiu Yang. 2022. Solving math word problems via cooperative reasoning induced language models. *arXiv preprint arXiv:2210.16257*.

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

Yuxin Zuo, Kaiyan Zhang, Shang Qu, Li Sheng, Xuekai Zhu, Biqing Qi, Youbang Sun, Ganqu Cui, Ning Ding, and Bowen Zhou. 2025. Ttrl: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*.

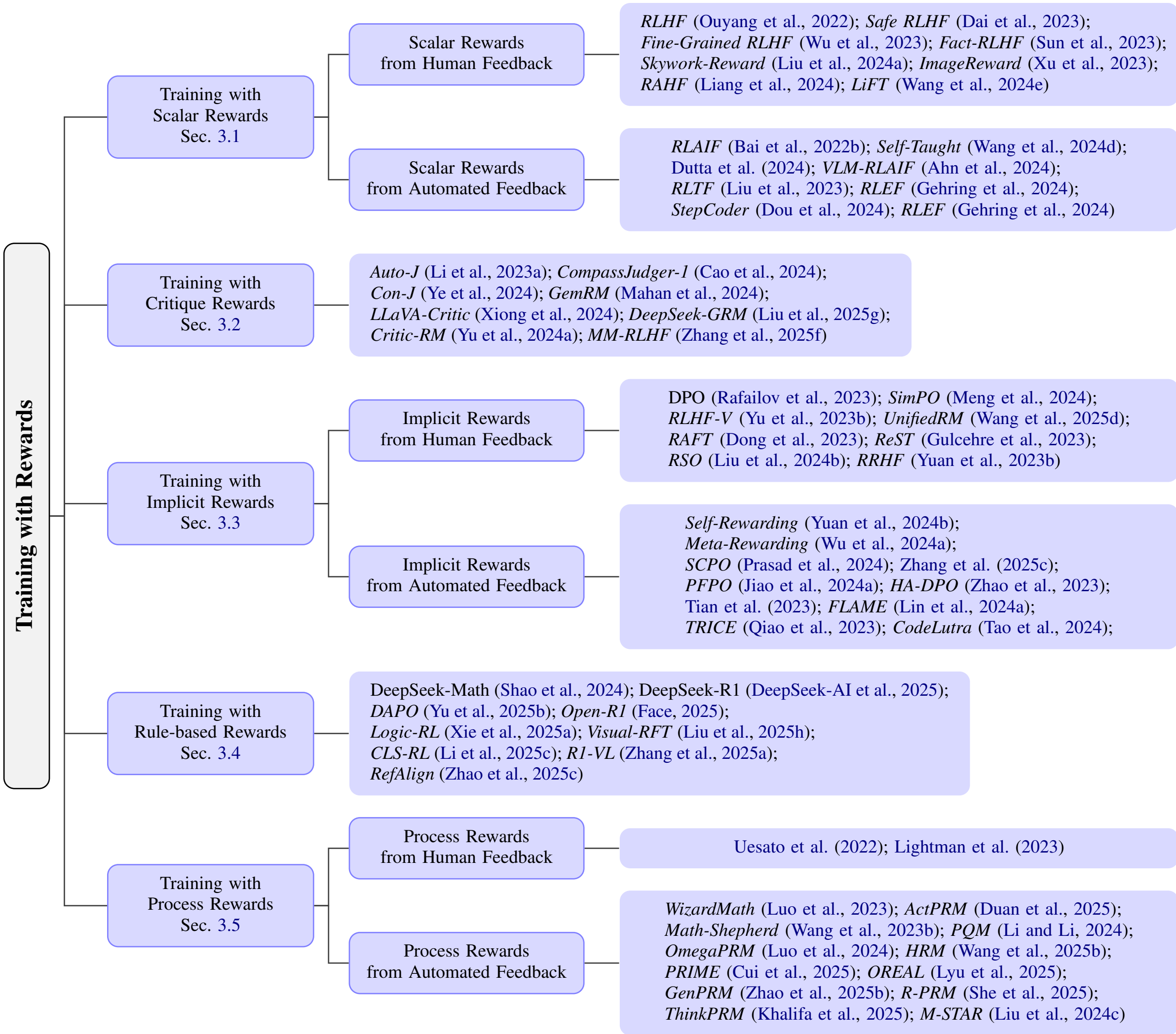

Figure 7: Overview of **Training with Rewards**.

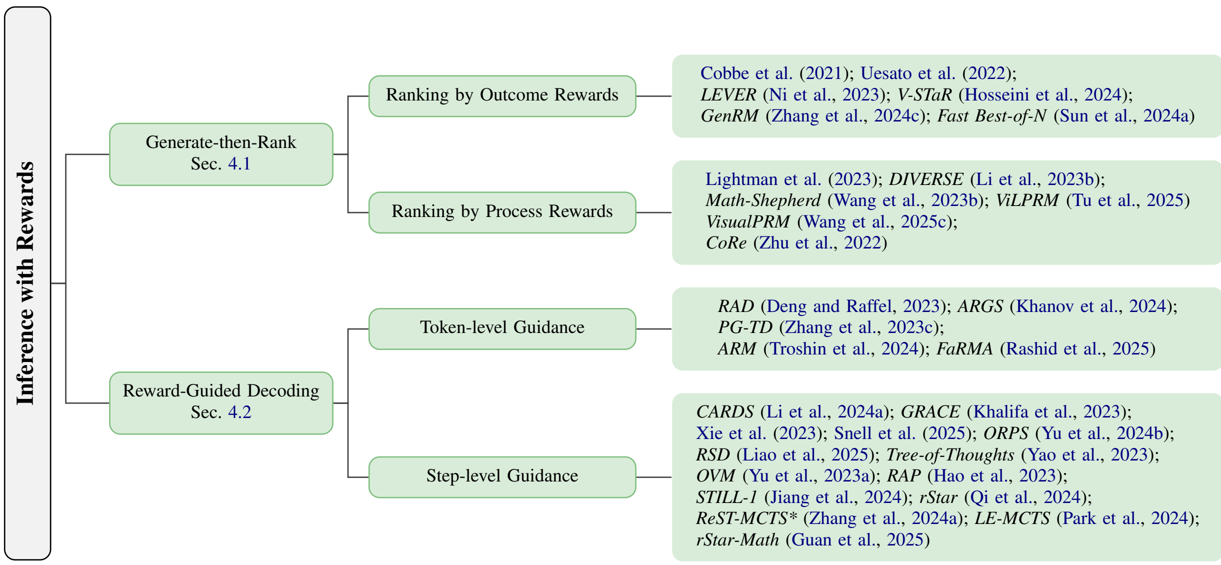

Figure 8: Overview of **Inference with Rewards**.

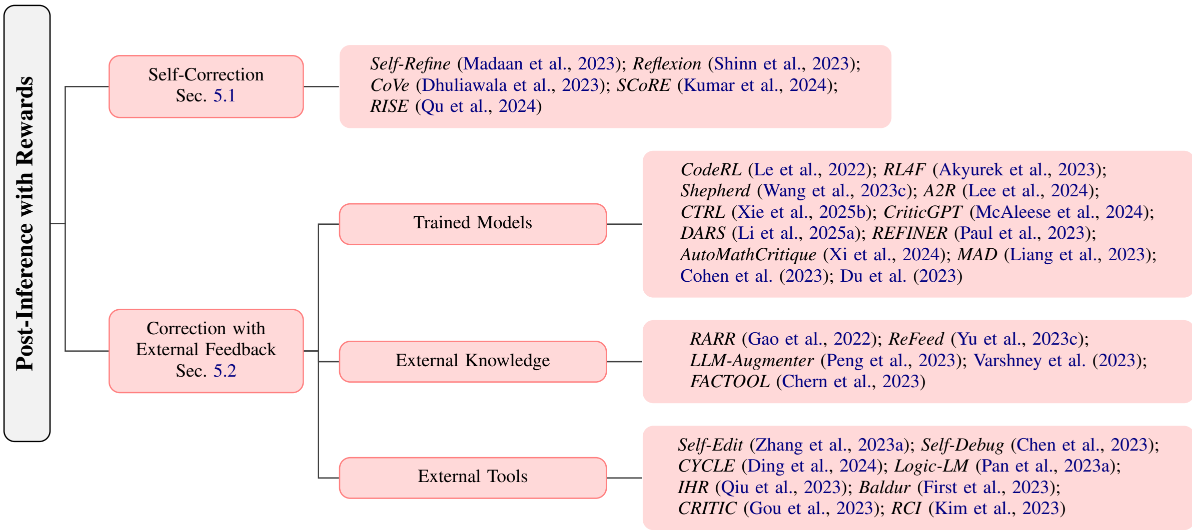

Figure 9: Overview of **Post-Inference with Rewards**.

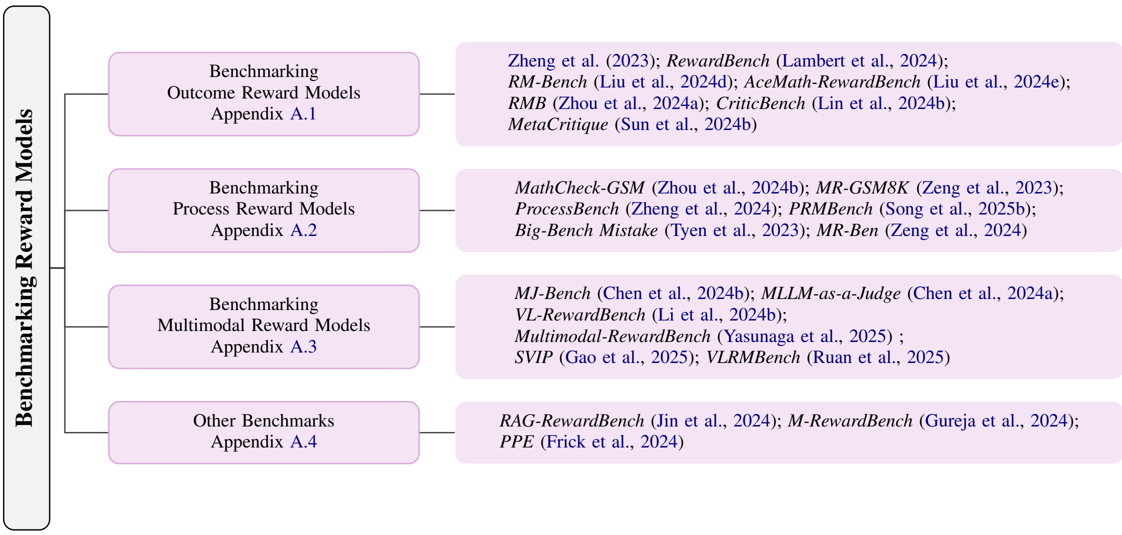

Figure 10: Overviews of **Benchmarking Reward Models**.

A Benchmarking Reward Models ((Extended))

A.1 Benchmarking Outcome Reward Models

A dominant line of benchmarking studies centers on outcome reward models that evaluate the overall quality of generated outputs. Zheng et al. (2023) is an early work that evaluates LLMs' judging ability by directly prompting them. As LLMs can naturally function as generative reward models, this study also represents one of the earliest benchmarks for reward models. *RewardBench* (Lambert et al., 2024) is the first comprehensive benchmarks for reward models. It aggregates preference data from existing datasets, such as AlpacaEval and MTBench, to evaluate reward model performance in chatting, reasoning, and safety. *RM-Bench* (Liu et al., 2024d) introduces evaluation for reward models on sensitivity to subtle content changes and robustness to style biases. It constructs preference pairs across chat, code, math, and safety domains using GPT-4o. *AceMath-RewardBench* (Liu et al., 2024e) focuses on math-specific evaluations. It tests whether reward models can identify correct solutions from candidates across various mathematical tasks and difficulty levels. *RMB* (Zhou et al., 2024a) furthermore broadens the evaluation scope to 49 real-world scenarios.

Apart from evaluating with preference data, some benchmarks focus on the critique ability of reward models. *CriticBench* (Lin et al., 2024b) assess whether reward models can generate critiques that accurately identify the correctness of a response and effectively guide the correction. Similarly, *MetaCritique* (Sun et al., 2024b) benchmarks LLM-generated critiques by decomposing them into atomic information units and assessing their correctness.

A.2 Benchmarking Process Reward Models

Recently more benchmarks focus on process reward models due to their increasing significance. In detail, several benchmarks focus on math reasoning, such as *MathCheck-GSM* (Zhou et al., 2024b), *MR-GSM8K* (Zeng et al., 2023), and *MR-MATH* (Xia et al., 2024). They require reward models to locate the first error step in a math reasoning solution. Their testing samples are adapted from existing math datasets, including GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021). Furthermore, *ProcessBench* (Zheng et al., 2024) features diversity and higher difficulty levels by scaling this up to Olympiad- and competition-level math problems (He et al., 2024; Gao et al., 2024). Beyond step correctness, *PRMBench* (Song et al., 2025b) offers a more fine-grained benchmark. It annotates each step in the reasoning path with specific error types grouped into three dimensions: simplicity, soundness, and sensitivity. The annotations come from LLM-generated perturbations and are subsequently verified by human annotators.

Besides mathematical reasoning, *Big-Bench Mistake* (Tyen et al., 2023) targets logical reasoning. It annotates chain-of-thought trajectories from BIG-Bench (bench authors, 2023), each labeled with the first logical error. Furthermore, *MR-Ben* (Zeng et al., 2024) expands this to the reasoning process of seven domains: math, logic, physics, chemistry, medicine, biology and code.

A.3 Benchmarking Multimodal Reward Models

Due to the prevalence of multimodal language models, another vital line of benchmarks focuses on multimodal reward models with diverse evaluation protocols.

MJ-Bench (Chen et al., 2024b) depends on text-to-image generation tasks for evaluation. It builds preference data across four dimensions: text-image alignment, safety, image quality, and social bias. *MLLM-as-a-Judge* (Chen et al., 2024a) uses image understanding tasks for benchmarking and includes pointwise and pairwise scoring. *VL-RewardBench* (Li et al., 2024b) includes three tasks: general multimodal instructions, hallucination detection, and multimodal reasoning. *Multimodal-RewardBench* (Yasunaga et al., 2025) spans six key capabilities of multimodal reward models: general correctness, human preference, factual knowledge, reasoning, safety, and VQA.

Beyond the outcome level, current benchmarks also assess multimodal process reward models. *SVIP* (Gao et al., 2025) targets process-level evaluation on relevance, logic, and attribute correctness of diverse multimodal tasks. It transforms reasoning paths into executable visual programs and automatically annotates each step. *VLRMBench* (Ruan et al., 2025) further integrates evaluation on three dimensions: reasoning steps, whole outcomes, and critiques on error analysis. It collects testing data of multimodal understanding through AI annotations and human verification.

A.4 Other Benchmarks

In addition to general-purpose evaluations, several benchmarks aim to address domain-specific or emerging challenges in reward modeling. *RAG-RewardBench* (Jin et al., 2024) targets reward model evaluation in RAG. It constructs preference data for RAG-specific scenarios, including multi-hop reasoning, fine-grained citation, appropriate abstention, and conflict robustness. *M-RewardBench* (Gureja et al., 2024) extends the evaluation to multilingual contexts. Instead of direct evaluation, *PPE* (Frick et al., 2024) indirectly evaluates reward models through RLHF pipelines. It measures the performance of trained LLMs with a reward model, offering a practical perspective.

B Applications

The strategies described above for learning from rewards have been widely adopted across diverse applications. Early applications focus on preference alignment, such as RLHF (Ouyang et al., 2022) and RLAIF (Bai et al., 2022b). In particular, the recent DeepSeek-R1 (DeepSeek-AI et al., 2025) has demonstrated the effectiveness of reinforcement learning to develop *large reasoning models*, which has inspired a wave of R-1 style applications for diverse areas. In this section, we review the primary applications following these strategies.

B.1 Preference Alignment

Learning-from-rewards strategies have become the cornerstone for aligning LLMs with human preferences. These strategies design diverse reward signals to encourage desirable attributes, such as factuality, harmlessness, and helpfulness, while penalizing undesired behaviors like toxicity, bias, and hallucination. We summarize three major objectives of preference alignment as follows.

- **Factuality and Reducing hallucination.** Hallucination refers to generating fluent but factually incorrect or fabricated content (Tian et al., 2023). It is a pervasive issue for language models, especially in knowledge-intensive tasks such as healthcare and scientific research. The methods for this alignment span the training, inference, and post-inference stages (Sun et al., 2023; Lin et al., 2024a; Zhao et al., 2023; Peng et al., 2023; Wang et al., 2023c). The rewards mainly stem from human preferences about factuality as well as external knowledge sources. For instance, *Fact-RLHF* (Sun et al., 2023) trains a factuality-aware reward model on human preferences and additional supervision from image captions and multiple-choice answers The reward model is then used to fine-tune the multimodal language model via PPO, guiding the model to reduce hallucinations. *RLFH* (Wen et al., 2024) decomposes the model responses into atomic statements, verifies their truthfulness against external knowledge, and converts them into dense token-level scalar rewards. To reduce hallucination, it directly uses these reward signals to fine-tune the model via PPO.
- **Safety and Harmlessness.** Safety and harmlessness constitute another critical axis of alignment, particularly in adversarial or socially sensitive contexts (Bai et al., 2022b; Ji et al., 2023). Language models must be discouraged from producing toxic, offensive, or biased content before being deployed in real-world systems. To this end, the methods primarily focus on the training (Ouyang et al., 2022; Bai et al., 2022a) and inference stages (Deng and Raffel, 2023; Khanov et al., 2024). For instance, *RAD* (Deng and Raffel, 2023) depends on reward signals to produce non-toxicity content during decoding.
- **Helpfulness.** Meanwhile, helpfulness emphasizes that language models are expected to provide relevant, informative, and context-aware responses to fulfill user intent (Taori et al., 2023). This alignment is imperative in areas like instruction-following and dialogue systems. Reward signals are generally sourced from human preferences and task-specific quality metrics (Bai et al., 2022a).

B.2 Mathematical Reasoning

Mathematical reasoning is vital to measure the language model's ability to solve complex reasoning problems. Some methods build reward models and fine-tune the language model for math reasoning (Shao et al., 2024; DeepSeek-AI, 2025), particularly using process reward models (Uesato et al., 2022; Luo et al., 2023) like *Math-Shepherd* (Wang et al., 2023b). They can provide step-level reward signals for a math reasoning solution. Moreover, some approaches construct preference data for math reasoning, *i.e.,* correct and incorrect solutions, and then fine-tune the language model through DPO (Lai et al., 2024; Xu et al., 2025a). Others include inference-time scaling strategies, such as generate-then-rank (Cobbe et al., 2021;

Lightman et al., 2023), and reward-guided decoding with search algorithms like MCTS (Hao et al., 2023; Guan et al., 2025).

B.3 Code Generation

The code generation task has made significant strides due to the development of LLMs, which improves software engineering productivity by a large margin. To improve the code language model through fine-tuning, the reward signals can come from various sources, including (Zhu et al., 2024), and code compiler feedback, unit test results, and code analysis (Liu et al., 2023; Dou et al., 2024; Tao et al., 2024; Zhou et al., 2025a). For example, DeepSeek-Coder-V2 (Zhu et al., 2024) trains a reward model for code generation and fine-tunes the language model via the reinforcement learning algorithm GRPO (Shao et al., 2024). Additionally, some approaches guide the inference of language models during code generation with reward models, including the generate-then-rank (Ni et al., 2023; Hosseini et al., 2024) and reward-guided decoding (Yu et al., 2024b). Another popular direction refines the generated code to correct errors and bugs through the language model itself (Shinn et al., 2023; Zhang et al., 2023a; Chen et al., 2023) or external feedback (Xie et al., 2025b).

B.4 Multimodal Tasks

Learning-from-rewards strategies have been widely applied to multimodal tasks, including multimodal understanding and generation. Most studies adopt reinforcement learning and reward-guided decoding methods. For instance, *Q-Insight* (Li et al., 2025d) focuses on improving comprehensive image quality understanding with reinforcement learning. *VLM-R1* (Shen et al., 2025a) applies reinforcement learning to fine-tune vision-language models and focuses on two tasks: referring expression compression and object detection. *Vision-R1* (Huang et al., 2025b) enhances multimodal reasoning of vision-language models for mathematical VQA. Zhan et al. (2025) proposes another *Vision-R1*, but it mainly facilitates object localization tasks with vision-language models.

Video-R1 (Feng et al., 2025b), *VideoChat-R1* (Li et al., 2025f), and *TinyLLaVA-Video-R1* (Zhang et al., 2025e) apply GRPO into video reasoning. *R1-V* (Chen et al., 2025a) and *CrowdVLM-R1* (Wang et al., 2025e) focus on visual counting. More example applications include multimodal reasoning (Zhou et al., 2025b; Meng et al., 2025; Tan et al., 2025; Li et al., 2025b; Liu et al., 2025f), object detection (Liu et al., 2025h), segmentation (Liu et al., 2025d), and image/video generation (Guo et al., 2025c; Liu et al., 2025a).

B.5 Agents

LLM Agent is an autonomous system that automatically performs complex tasks through task decomposition and action execution in dynamic environments (Wang et al., 2024b). Various learning-from-rewards strategies have been applied to training or guiding the agents. *AgentRM* (Xia et al., 2025) targets general-purpose decision-making agents across domains such as web navigation, embodied planning, text games, and tool use. During inference, a reward model guides the agents to choose candidate actions or trajectories. *AgentPRM* (Choudhury, 2025) trains LLM agents with a process reward model. *KBQA-o1* (Luo et al., 2025) guides MCTS with a reward model for the knowledge base question answering task with agents. *DeepResearch* (OpenAI, 2025) and *DeepResearcher* (Zheng et al., 2025) design agents for research tasks. They both use reinforcement learning to fine-tune the agents. *UI-R1* (Lu et al., 2025) introduces a rule-based reinforcement learning framework for GUI action prediction with multimodal agents. *InfiGUI-R1* (Liu et al., 2025c) is a similar work with GUI agents. *RAGEN* (Wang et al., 2025f) propose training the agents via multi-turn reinforcement learning with a new algorithm based on GRPO.

B.6 Other Applications

Many other applications have been developed following the learning-from-rewards strategies.

Embodied AI is essential for the development of artificial general intelligence. AI systems, such as embodied robots, must interact with the physical world and complete complex tasks through high-level planning and low-level control. They generally aim to enhance the embodied reasoning abilities with reinforcement learning, such as *Cosmos-reason1* (Azzolini et al., 2025), *iRe-VLA* (Guo et al., 2025b), *Embodied-Reasoner* (Zhang et al., 2025d), and *Embodied-R* (Zhao et al., 2025a).

Several approaches apply reinforcement learning to reason with information retrieval from knowledge databases or the real-world web. These approaches include *R1-Searcher* (Song et al., 2025a), *Search-R1* (Jin et al., 2025), *DeepRetrieval* (Jiang et al., 2025), *ReSearch* (Chen et al., 2025b), and

WebThinker (Li et al., 2025e). They adopt different reward designs to improve search performance.

Applications for other applications also emerge. *ToRL* (Li et al., 2025g), *ReTool* (Feng et al., 2025a), *SWi-RL* (Goldie et al., 2025), *ToolRL* (Qian et al., 2025) and *OTC* (Wang et al., 2025a) are proposed to improve LLMs' reasoning ability to call various tools through reinforcement learning. *Rec-R1* (Lin et al., 2025) applies reinforcement learning for recommendation system. *SWE-RL* (Wei et al., 2025) aims at software engineering with reinforcement learning. *SQL-R1* (Ma et al., 2025) focuses on natural language to SQL reasoning. It uses a composite reward function with format correctness, execution success, result accuracy, and reasoning completeness.

Some applications are designed for specific areas. *Med-R1* Lai et al. (2025) and *MedVLM-R1* (Pan et al., 2025) are proposed for medical field. They target medical VQA across various imaging modalities (*e.g.,* CT, MRT, and X-ray) and several clinical tasks, like diagnosis, and anatomy identification. *Fin-R1* (Liu et al., 2025e) develops LLMs for the financial field, targeting financial QA and decision-making. It leverages accuracy and format rule-based rewards to train a language model on domain-specific data. *DianJin-R1* (Zhu et al., 2025) is another LLM for the financial field with reinforcement learning.

C Challenges and Future Directions

In this section, we discuss the current challenges and future directions of learning from rewards. Figure 11 summarizes the key challenges and future directions from the perspective of reward model design and learning strategies. Ultimately, we envision the development of interpretable, robust, and continually evolving agent systems capable of interacting with and adapting to the complexities of the real world.

C.1 Interpretability of Reward Models

Interpretability of reward models remains an open challenge for the learning-from-rewards strategies (Russell and Santos, 2019; Zhang et al., 2023d; Jenner and Gleave, 2022). Most reward models are typically treated as black boxes that produce scalars or critiques without exposing human-interpretable explanations. Such opacity hinders human trust and oversight and may lead to misaligned optimization. In consequence, enhancing reward model interpretability is essential for reliable alignment, enabling humans to inspect and verify the internal decision process and steer models toward desired behavior. Recent efforts have attempted to address this issue. For instance, *ArmoRM* (Wang et al., 2024a) improves the interpretability with multi-objective reward modeling, where each objective corresponds to a human-interpretable dimension, such as helpfulness, correctness, coherence, complexity, and verbosity. While this approach is effective, its interpretability is limited to these predefined objectives. In addition, emerging generative reward models can disclose their rationales of reward scoring (Zhao et al., 2025b; Khalifa et al., 2025). While promising, their interpretability remains limited and demands further investigation into consistency, reliability, and faithfulness.

C.2 Generalist Reward Models

A promising future direction is the development of *generalist reward models*. Most existing reward models are designed for narrow domains; thus they often suffer from weak generalization across tasks. Moreover, their reward outputs are typically static and lack support for inference-time scalability, hindering their application in diverse and open-ended scenarios (Liu et al., 2024a; Zhang et al., 2024c; Snell et al., 2025).

In contrast, a generalist reward model seeks to overcome these limitations. They demand flexibility for input types, including single, paired, or multiple responses, and also require accurate reward generation in various domains, such as question answering, math reasoning, and code generation. Besides, they are expected to generate higher-quality reward signals with increased inference-time computing. Such models offer a unified interface for reward modeling across domains and enable scalable, interpretable reward generation. For example, *DeepSeek-GRM* (Liu et al., 2025g), a recent attempt in this direction, proposes a pointwise generative reward model. Rather than only scalars, it can generate evaluative natural language principles and critiques, enabling effective inference-time scaling through multi-sample voting and meta-reward filtering.

C.3 Reward Hacking

Reward hacking is a fundamental challenge in learning from rewards (Everitt et al., 2021; Amodei et al., 2016; Weng, 2024; Liu et al., 2025b). It occurs when models exploit unintended shortcuts in

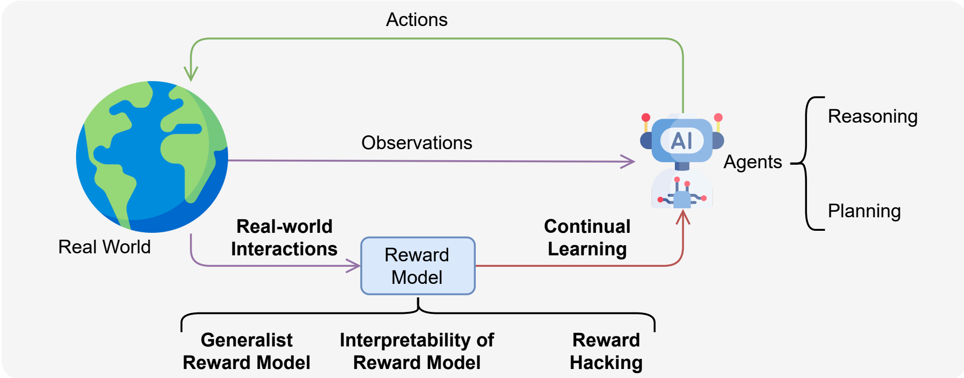

Figure 11: Illustration of challenges and future directions.

the reward function to obtain high rewards without truly learning the desired behaviors or completing the task as designed. This phenomenon has been observed across domains. For instance, LLMs may fabricate plausible yet incorrect answers, and code LLMs subtly modify unit tests to pass evaluations (Denison et al., 2024). Reward backing can also happen during inference, called *in-context reward hacking* (Pan et al., 2024b,a). It arises in self-refinement loops where the same model acts as both the generator and the judge. In such cases, the model may learn to produce outputs that exploit its own evaluation heuristics, leading to inflated internal scores while deviating from true objectives.

Reward hacking fundamentally arises from the difficulty of specifying a reward function that perfectly captures the true objectives. As articulated by Goodhart's Laws—*When a measure becomes a target, it ceases to be a good measure*—any proxy metric used as a reward will eventually be exploited once applying optimization pressure. To mitigate reward hacking, the following directions are worth exploring: (i) Designing more robust and tamper-resistant reward functions (Razin et al., 2025; Shen et al., 2025b; Peng et al., 2025); (ii) Detecting misalignment via behavioral or distributional anomaly detection (Pan et al., 2022); (iii) Decoupling feedback mechanisms to prevent contamination (Uesato et al., 2020); (iv) Auditing the dataset for training reward models to reduce reward hacking risks (Revel et al., 2025).

C.4 Grounded Rewards from Real-World Interactions

Despite recent advances in learning from rewards for LLMs, most methods fundamentally rely on human preferences or well-curated automated feedback. The LLMs are typically optimized to maximize the rewards derived from these feedback. In consequence, this inherently limits the agent's ability to surpass existing human knowledge and adapt to complex environments.

Due to these limitations, moving beyond chat-driven rewards toward grounded real-world rewards is another promising direction. This movement requires LLMs to be integrated into agentic frameworks, and agents should increasingly interact directly with their environment and derive reward signals from observed outcomes. For example, a health assistant could optimize behavior based on physiological signals rather than user ratings, and a scientific agent could refine hypotheses based on experimental data rather than expert approval (Silver and Sutton, 2025). This shift would enable agents to close the feedback loop with the real world, allowing for autonomous discovery, adaptation, and pursuit of goals beyond human understanding. The transition to real-world interactions raises substantial technical challenges. Agents must handle noisy, delayed, or partial feedback from complex environments, requiring advances in credit assignment, robust exploration, and uncertainty modeling.

C.5 Continual Learning from Rewards

Current learning-from-rewards strategies often assume a fixed dataset, a predefined reward model, and short episodic interactions. Once trained, models typically exhibit limited abilities to adapt to new tasks or evolving environments (Zhang et al., 2024b; Silver and Sutton, 2025). This episodic and offline paradigm sharply contrasts with real-world intelligence's dynamic, ongoing nature, where agents must continually learn from experience and recalibrate based on new feedback.

As such, a vital direction is continual learning from rewards. It is a crucial foundation for building lifelong competent and aligned agents. By abandoning the traditional assumption of fixed objectives, models can remain responsive to changing reward signals, avoid performance degradation under distributional shifts, and better reflect long-term user intent. Notably, it is a broader idea of continual reinforcement learning (Abel et al., 2023; Li et al., 2024c; Bowling and Elelimy, 2025). Achieving continual learning from rewards presents significant challenges. It requires addressing catastrophic forgetting, maintaining stability while enabling plasticity, and designing dynamic reward modeling mechanisms.