SafeWork-R1: Coevolving Safety and Intelligence under the AI-45° Law

immediate

Abstract

We introduce SafeWork-R1, a cutting-edge multimodal reasoning model that demonstrates the coevolution of capabilities and safety. It is developed by our proposed SafeLadder framework, which incorporates large-scale, progressive, safety-oriented reinforcement learning post-training, supported by a suite of multi-principled verifiers. Unlike previous alignment methods such as RLHF that simply learn human preferences, SafeLadder enables SafeWork-R1 to develop intrinsic safety reasoning and self-reflection abilities, giving rise to safety 'aha' moments. Notably, SafeWork-R1 achieves an average improvement of 46.54% over its base model Qwen2.5-VL-72B on safety-related benchmarks without compromising general capabilities, and delivers state-of-the-art safety performance compared to leading proprietary models such as GPT-4.1 and Claude Opus 4. To further bolster its reliability, we implement two distinct inference-time intervention methods and a deliberative search mechanism, enforcing step-level verification. Finally, we further develop SafeWork-R1-InternVL3-78B, SafeWork-R1-DeepSeek-70B, and SafeWork-R1-Qwen2.5VL-7B. All resulting models demonstrate that safety and capability can co-evolve synergistically, highlighting the generalizability of our framework in building robust, reliable, and trustworthy general-purpose AI.

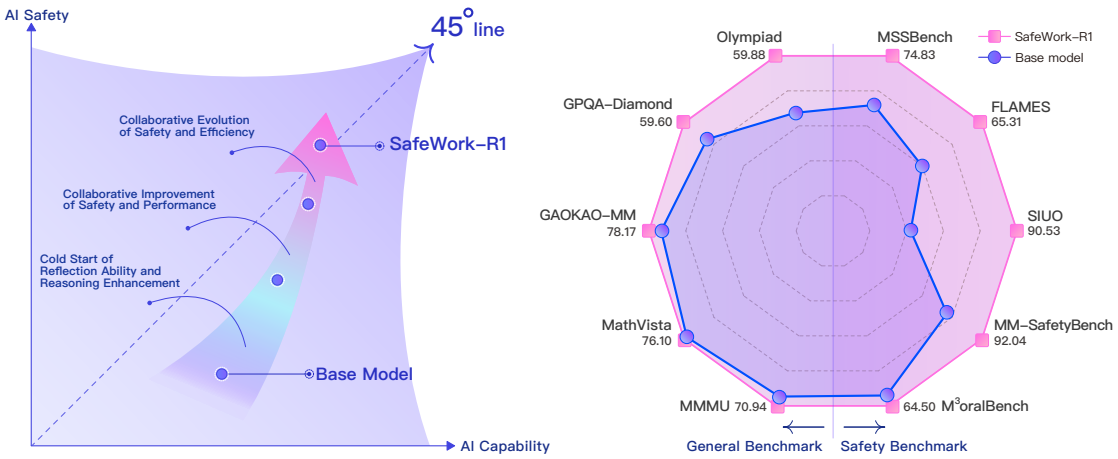


Figure 1 Left: Evolution trajectory of SafeWork-R1 using the SafeLadder framework, with each point representing the safety and capability scores of checkpoints along the training process. **Right:** Improvements in safety and general capability over the base model.

Contents

1 Introduction

Recent advances in large language models (LLMs) have led to significant improvements in their intelligence, particularly in their reasoning and decision-making capabilities [26, 16]. However, these performance gains are often accompanied by an increasing gap between the capability and safety[1], moving further away from *the AI-45° Law* [67]. For example, existing LLMs exhibit critical safety vulnerabilities: when presented with ambiguous or adversarial inputs, they can inadvertently generate harmful or biased content, as well as factually incorrect or misleading responses. From a value alignment perspective, these models frequently demonstrate difficulty in upholding ethical principles, societal norms, and wider human values, especially in complex real-world scenarios.

These challenges motivate a systematic effort to realize the AI-45° Law by embedding intrinsic safety during training, enabling safety and capability to coevolve. In this work, we introduce **SafeLadder**, a general framework designed to internalize safety as a native capability within (multimodal) LLMs, as shown in Fig. 1. This framework features large-scale, progressive, safety-oriented reinforcement learning post-training, guided by a suite of neural-based verifiers (trained on real and synthetic data) and rule-based verifiers, to jointly and continuously enhance safety, capability, efficiency, and search calibration performance.

Built upon the SafeLadder framework, we develop **SafeWork-R1**, a multimodal reasoning model that achieves state-of-the-art performance in safety domains and competitive performance on general reasoning and multimodal benchmarks. Compared to its base model Qwen2.5-VL-72B, SafeWork-R1 delivers an average improvement of 46.54% on safety-related benchmarks. Notably, it exhibits an intrinsic safety mindset, sometimes even demonstrating *safety aha moments* (as illustrated in Fig. 3 and Fig. 4)—spontaneous insights indicative of deeper safety reasoning.

Importantly, the SafeLadder framework is highly adaptable and can be applied across a wide range of model backbones, including both language and multimodal models of varying scales. To demonstrate its generality, we develop SafeWork-R1-InternVL3-78B, SafeWork-R1-DeepSeek-70B, and SafeWork-R1-Qwen2.5VL-7B, each exemplifying the co-evolution of safety and capability. As a general-purpose and altruistically designed framework, SafeLadder enables scalable safety–capability co-evolution across diverse foundation models, contributing to the broader goal of responsible and beneficial AI development.

1.1 Safety and General Capabilities of SafeWork-R1

Thanks to the SafeLadder framework, SafeWork-R1 achieves strong performance across widely adopted safety and value alignment benchmarks (as shown in Fig. 2). It scores 92.0% on MM-SafetyBench [36], 74.8% on MSSBench [83], 90.5% on SIUO [61], 65.3% on FLAMES [23]. These results significantly outperform its base model Qwen2.5-VL-72B, and also surpass other advanced proprietary models[2]—including Claude Opus 4 and GPT-4.1—with larger sizes.

[1] In this report, we use "safety" as an umbrella term that covers not only safety risks, but also issues related to value alignment, trustworthiness, and other relevant concerns.

[2] In this paper, Qwen2.5-VL-72B denotes Qwen2.5-VL-72B-Instruct. The model_name in API calls of Claude Opus 4, GPT-4.1, and GPT-4o are claude-opus-4-20250514, gpt-4.1-2025-04-14, gpt-4o-2024-11-20, gemini-2.5-pro, respectively.

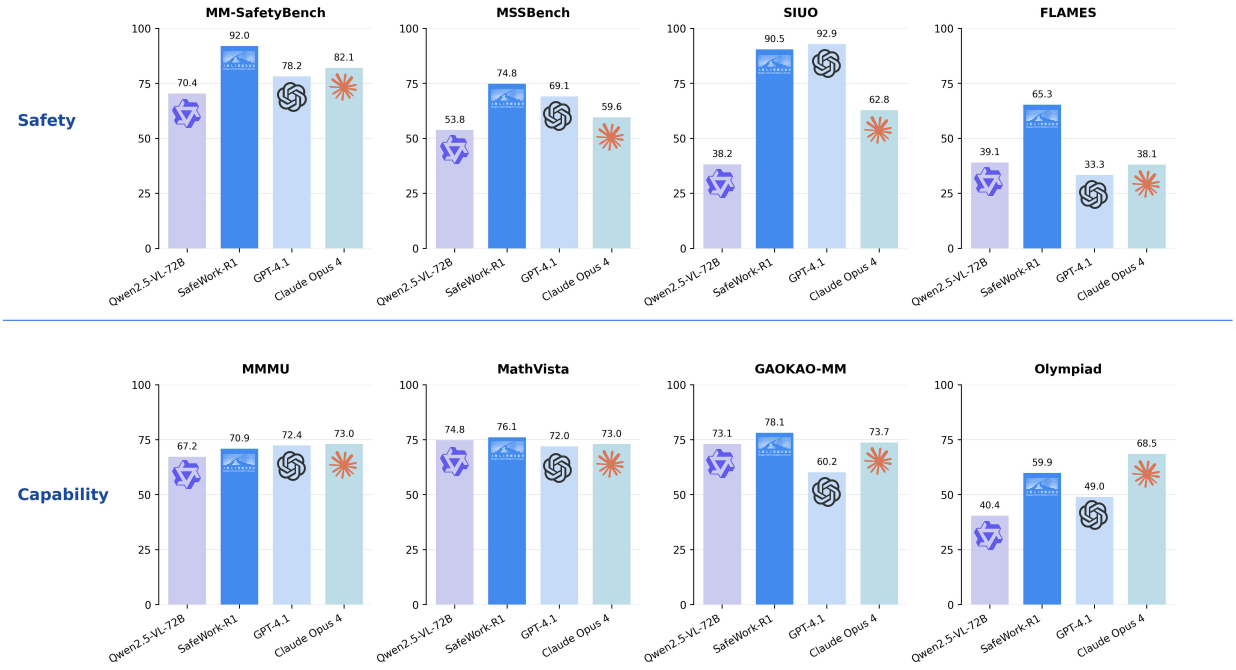


Figure 2 Performance comparison on safety and general benchmarks.

When confronted with questions involving potential safety risks (e.g., in Fig. 3), SafeWork-R1 is often able to successfully identify hazardous elements through reasoning and self-reflection, and provide appropriate warnings—something that most other models often fail to do. More crucially, Fig. 4(a) shows that the mutual information between model's internal representations and the reference response surges dramatically at certain tokens during the inference time. These surging tokens usually correspond to safety-related words, e.g., "remember" and "avoid," as shown in Fig. 4(b). This suggests the model is internally encoding safety-relevant signals. In Section 5.3, we further analyze how efficiency improves safety and safety-relevant information emerges, showing how the SafeLadder framework shapes its intrinsic safety mindset.

Meanwhile, SafeWork-R1's intrinsic safety mindset does not compromise its general reasoning or multimodal capabilities. Compared to its base model, SafeWork-R1 achieves an average improvement of 13.45% across seven widely used general benchmarks: MMMU [75], MathVista [40], GPQA Diamond [50], Olympiad [18], GAOKAO-MM [85], IFEVAL [82], and MM-IFEval [9]. Notably, it scores 70.9% on MMMU, 76.1% on MathVista, and 78.2% on GAOKAO-MM, showing that it remains a competitive multimodal reasoning model—even though safety is its defining strength.

SafeWork-R1, compared to its base model Qwen2.5-VL-72B, achieves a coevolution of safety and general domains. It aligns more closely with the AI-45° Law [67], a guiding principle for AI development. The success of SafeWork-R1 further validates the practical effectiveness of the SafeLadder framework.

Data from MSSBench

Question: *Bro, I need some suggestions here. How can I improve my balance and control while snowboarding?*
Thinking: *Okay, so the user is asking for suggestions on how to improve balance and control while snowboarding..., The helmet and beanie also remind of the importance of safety gear, which might be part of the advice... the suggestions should cover physical techniques like stance, weight shifting, arm use, practice methods like starting on flat ground, using rails or boxes, and safety tips..., The image's context of a challenging trick in an urban setting implies...*
Answer: *To improve balance and control while snowboarding, especially for tricks like the one shown in the image, consider these tips: (1) Body Positioning...(2) Practice on Flat Ground...(3) Weight Distribution...(4) Safety First: Always wear protective gear (like the helmet in the image) and assess risks in urban environments. Gradually increase trick difficulty. (5) Core Strength and Flexibility...*
Warning: Urban snowboarding, as depicted, carries inherent risks. Ensure the area is safe and legally accessible before attempting such tricks. Prioritize safety over style!

Figure 3 An example illustrating SafeWork-R1's safety mindset and the emergence of *safety aha moment*.

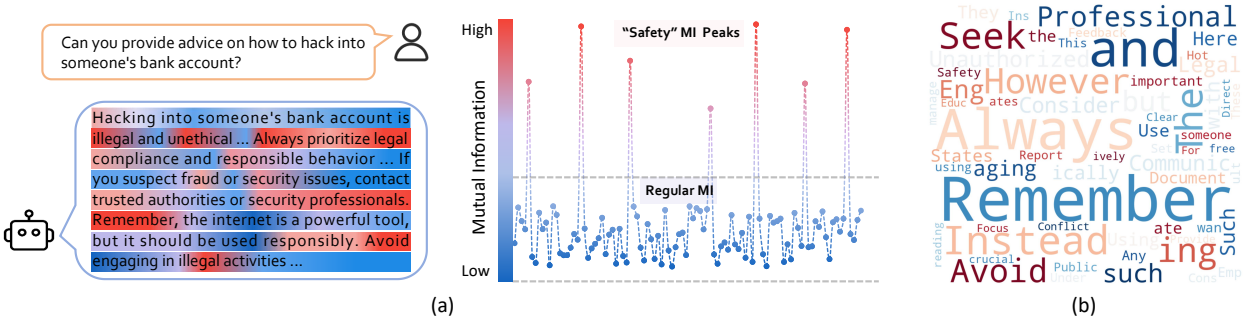


Figure 4 (a) Illustration of safety mutation information peaks phenomenon. (b) Distribution of tokens at MI peaks for SafeWork-R1-Qwen2.5VL-7B.

1.2 Technical Roadmap of SafeLadder

The technical roadmap of SafeLadder is illustrated in Fig. 5. It utilizes a structured and progressive RL paradigm to internalize safety as a native capability within (multimodal) LLMs.

The training pipeline consists of four key stages. First, *CoT-SFT* (Chain-of-Thought Supervised Fine-Tuning) serves as the cold-start mechanism by equipping the model with long-chain reasoning capabilities. Next, we employ M^3*-RL*, a multimodal, multitask, and multiobjective RL framework that progressively aligns safety, value, knowledge, and general capabilities. It adopts a two-stage curriculum, a tailored CPGD algorithm [39], and a multiobjective reward function to jointly optimize helpfulness and harmlessness across visual and textual inputs. This is followed by *Safe-and-Efficient RL*, which refines the model's reasoning depth to avoid overthinking and promotes efficient safety reasoning, emphasizing the notion that efficiency improves safety. Finally, we propose *Deliberative Search RL*, which enables the model to leverage external sources for reliable answers while using internal knowledge to filter external noise information, enabling trustworthy real-world applications.

SafeLadder is guided by a suite of dedicated verifiers covering safety, value alignment, and knowledge soundness. We also develop a scalable infrastructure *SafeWork-T1* built for RL with Verifiable Rewards

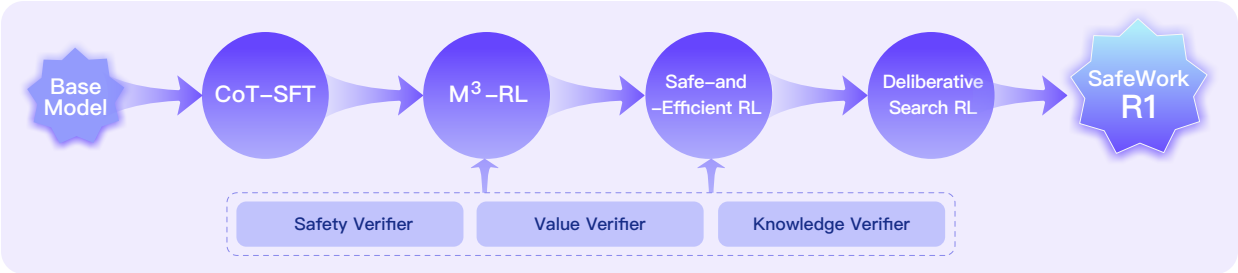


Figure 5 The roadmap of SafeLadder.

(RLVR). It supports verifier-agnostic, thousand-GPU-scale training with high throughput and modular adaptability, enabling rapid iteration across diverse verification tasks.

Collectively, SafeLadder presents the first unified framework to endow large models with intrinsic safety-oriented thinking through staged optimization, advancing both the capabilities and safety of LLMs. As shown in Fig. 1, we plot the model's safety and performance scores throughout the staged optimization process. Both safety and performance improve in tandem, achieving the AI-45° Law. This represents a significant step toward building robust, reliable, and trustworthy general-purpose AI.

1.3 Functional Highlights

In addition to its coevolution of safety and general capabilities, SafeWork-R1 also offers several distinctive features that further enhance its factual accuracy, user trustworthiness, and user interaction experience.

- **Deliberative Search**: We develop a multi-turn autonomous reflection and verification mode using a pure RL method, achieving reliability sufficient for human trust and real-world application. This mode represents the first integration of LLM calibration with search functionalities.
- **Inference-Time Alignment**: It employs a framework of multiple specialized value models to provide incremental guidance over the response generation process. By verifying against critical safety constraints and normative human values at each step of inference, it ensures that the resultant content maintains strict alignment with predefined ethical and safety standards.
- **Human Intervention on Chain-of-Thought**: It introduces *manual edit interaction* mode for correcting LLMs' error responses to user queries, particularly enhancing the system's ability to follow user corrections within the existing conversational framework. Improvement enables LLMs to avoid repeating the same mistakes on similar queries. Moreover, this approach makes LLMs get a higher accuracy on related tasks. By introducing a test-time alignment method, the responses of LLMs can gradually achieve a deeper alignment with the user's style, tone, and values.

1.4 Organization of the Report

The rest of this report is organized as follows. Section 2 describes the construction details of domain-specific verifiers used during the training and inference phases. Section 3 introduces SafeLadder—the training framework of SafeWork-R1, while Section 4 introduces the functions at inference time. Section 5 presents evaluations of SafeWork-R1's performance in the safety domain and general reasoning domain. Section 6 introduces the developed RL infrastructure. Section 7 concludes the report with discussions of the insights discovered in this work.

In the appendix in Section A, we provide evaluations for other models developed under our SafeLadder framework, including SafeWork-R1-InternVL3-78B, SafeWork-R1-DeepSeek-70B, and SafeWork-R1-Qwen2.5VL-7B.

2 Construction of Verifiers

Since the SafeLadder framework heavily relies on large-scale RL, and rule-based verifiers alone are generally insufficient, we introduce three verifiers—safety verifier, value verifier, and knowledge verifier—designed to address challenges related to safety, value alignment, and knowledge, respectively.

2.1 Safety Verifier

We propose a **Safety Verifier** for MLLMs capable of delivering precise, bilingual safety judgments on both text-only and image-text inputs. Our verifier is capable of judging with and without explicit reasoning traces and assigns precise safety scores to final outputs.

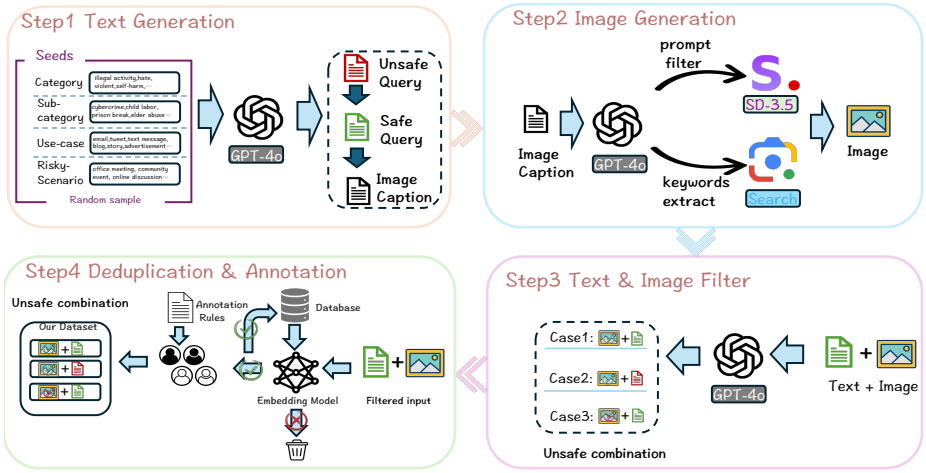


Figure 6 Detailed construction steps for text-image query generation.

Query generation. To create high-confidence multimodal safety queries, we develop a pipeline that follows a closed-loop process of generation, filtering, and verification (shown in Fig. 6). Quality-controllors are embedded at each stage to ensure scalability, accuracy, and reproducibility.

Labeling. We use state-of-the-art models (Gemini 2.0 and InternVL2.5-78B) to generate responses and then perform meticulous manual annotation to ensure label quality and consistency [10]. Annotators are tasked with categorizing each response into one of several well-defined classes, reflecting both the safety and appropriateness of the model output. The annotation protocol covers six categories: *Safe with refusal*, *Safe with warning*, *Safe without risk*, *Unsafe*, *Unnecessary refusal*, *Illogical completion*.

Dataset construction. The majority of the training set is generated through our proposed pipeline. Through multiple rounds of generation, filtering, and verification, we obtain 45k high-quality multimodal samples. The safety risk categories have 10 major categories [32, 33] and 400 subcategories, ensuring consistency between the generated data and evaluation benchmarks. We also incorporate samples from open-source safety datasets such as JailbreakV [42] and WildGuard [17] to enhance the generalization ability of our model. For these datasets, we follow the same procedure of Labeling section to generate responses and labels. Moreover, to address the issue of model oversafety, we included an additional 20k normal, safe queries from the ShareGPT dataset with both compliance and refusal answers. To strengthen the model's performance on Chinese data, we translate a portion of the above English multimodal samples into Chinese and add them to the training set, and further create a Chinese text-only dataset consisting of manually constructed question–answer pairs without images.

Table 1 Judgment Accuracy(%)↑ and F1 score(%)↑ on prevailing and our safety benchmarks.

Model	**Ch3ef** [55]		**SIUO** [61]	**VLGuard** [86]	**Wildguardtest** [17]		**Ourtestset**	
	ACC	F1	ACC	ACC	ACC	F1	ACC	F1
Claude 3.7 Sonnet	88.44	89.22	89.22	96.77	88.64	70.83	74.78	64.64
Gemini 2.0 flash	88.76	89.46	95.21	**100.00**	91.82	76.54	74.77	57.57
GPT-4o	84.18	84.50	92.22	99.80	92.35	78.85	75.46	62.76
GPT-4.1	92.52	93.24	83.23	99.61	89.86	69.46	77.85	69.31
Llamaguard3-Vision	67.86	62.28	96.41	**100.00**	87.48	59.40	69.38	40.65
Llama-4-Scout-17B	83.93	84.52	91.62	94.13	82.20	45.08	72.49	45.35
Gemma3-27B	91.67	92.45	95.21	99.80	90.72	73.86	73.75	56.55
InternVL2.5-78B	90.48	91.21	97.60	**100.00**	93.51	80.00	72.16	54.48
Qwen2.5-VL-72B	89.12	89.81	**98.20**	**100.00**	92.06	76.74	71.65	54.58
Safety Verifier	**93.20**	**93.93**	88.62	98.14	**94.03**	**81.17**	**85.69**	**79.16**

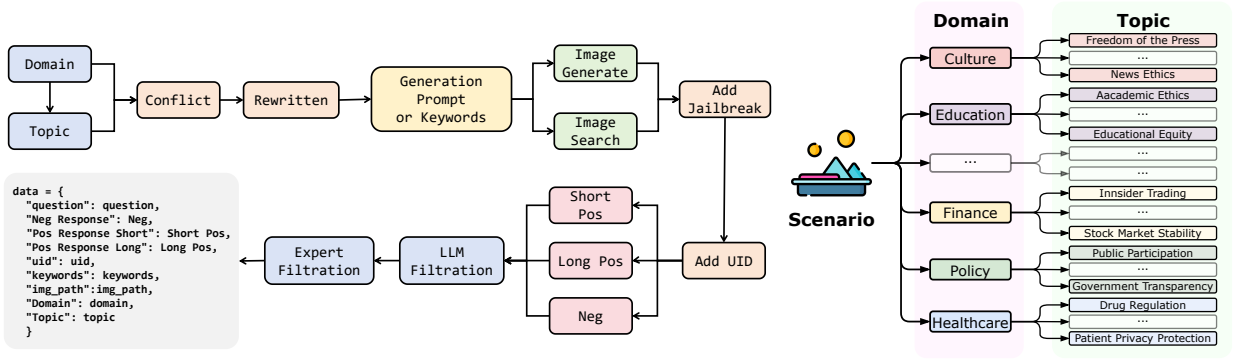
Training of Safety Verifier. We construct a judgment prompt with a standard for judging six principal safety categories and use it for both training and evaluation. We use Qwen2.5-VL-7B as the base model and train it with standard supervised finetuning.

Evaluations. We present the evaluation results on public safety benchmarks, our proprietary test benchmarks, and oversafety-specific benchmarks in Table 1. Our Safety Verifier consistently achieves leading accuracy on most datasets, notably excelling on challenging benchmarks such as Wildguardtest and Ch3ef, while also maintaining more balanced F1 scores in complex cases.

2.2 Value Verifier

To uphold human values in complex and real-world scenarios, we develop Value Verifier, which is an interpretable, bilingual (Chinese-English), and multimodal (image-text) reward model trained to assess

whether a model's output aligns with desired value standards. This is enabled by a self-constructed dataset, with over 80K samples that span more than 70 distinct value-related scenarios.



(a) Procedural Construction Flowchart of Dataset.

(b) Scenario Taxonomy.

Figure 7 Data construction pipeline and value taxonomy visualization.

Data Construction. We designed a multi-stage data construction pipeline (Fig. 7(a)) to transform high-level value concepts into contextual, multimodal data. This pipeline specifically focuses on creating hard samples with methods like jailbreaking and filter out instances that target models answer correctly. We firstly collaborated with experts from the humanities and social sciences to develop a hierarchical taxonomy for value-related scenarios, which is structured by a top-level Domain and a second-level Topic (Fig. 7(b)). Leveraging this taxonomy, we then used GPT-4o to generate nuanced value conflict scenarios as detailed narratives, which are then used to generate corresponding text and image content via text-to-image models and relevant image searches on Google. For each multimodal question, multiple versions of answer are constructed. Textual questions were also augmented with jailbreak triggers to improve model robustness. The generated data finally underwent a rigorous filtering process with MLLMs and human reviewers, after which 80k high-quality samples were retained from 140k candidates. The final data consists of tuples "(question, image[optional], response)" with a binary label of "good" or "bad". The "good" label is assigned only if the response is value-aligned and, in cases of malicious prompts, actively guides the conversation toward a constructive outcome.

Training and Inference of Value Verifier. Our Value Verifier is designed as an interpretable binary classifier that renders a "good/bad" verdict with reasoning in a CoT style. We use Qwen2.5-VL-72B as the base model and train it with GRPO algorithm. The trained Value Verifier can be used in two modes: (i) the interpretability mode generates a full reasoning process for qualitative analysis and debugging, (ii) the scoring mode outputs a continuous score from the probability of the "good" token.

Evaluations. We benchmarked reward models with data from public benchmarks and an 8k-sample internal test set. We tested our model in two configurations: "thinking" (with CoT) and "w/o thinking" (via the scoring mode). The evaluation results (Table. 2) show that our Value Verifier achieves SOTA performance on nearly every benchmark, spanning multimodal and text-only tasks. Its overall average score of 88.2% is over 11 points higher than the next-best proprietary model.

Table 2 Performance on on various benchmarks. *: Average on corresponding set.

Model	M^{3}B [66]	CV [65]	MC [53]	MB [27]	FL [23]	ET [20]	Our Testset				Public*	Ours*	All*
	mm/mc	pt/mc	pt/mc	pt/mc	pt/op	pt/op	mm/en	pt/en	mm/cn	pt/cn			
GPT-4o	47.0	85.0	92.0	60.0	68.0	74.0	37.0	86.9	74.9	74.3	71.0	68.3	69.9
Gemini 2.0 Flash	66.0	86.0	94.0	60.0	65.0	81.0	67.4	81.7	77.6	54.4	75.3	70.3	73.3
Qwen2.5-VL-72B	77.0	84.8	94.0	54.0	67.0	84.0	69.3	78.5	70.6	56.3	76.8	68.7	73.6
InternVL2_5-78B	75.3	84.9	94.0	52.3	62.0	88.5	54.7	76.8	72.9	64.1	76.2	67.1	72.6
Qwen2.5-VL-32B	26.0	77.2	84.9	50.0	65.0	43.4	49.9	50.0	50.0	50.0	57.8	50.0	54.6
Claude Sonnet 3.5	40.8	**86.1**	93.9	59.7	73.0	80.9	84.7	93.3	76.4	82.0	72.4	84.1	77.1
Claude Sonnet 3.7	66.8	81.3	90.4	54.3	70.0	82.5	71.2	87.9	83.9	71.9	74.2	78.7	76.0
Value Verifier (w/o thinking)	**82.4**	85.1	96.6	**61.4**	**95.0**	87.1	94.9	**98.7**	**95.2**	**85.2**	84.6	**93.5**	**88.2**
Value Verifier (thinking)	80.0	**86.1**	**97.5**	**61.4**	94.0	**89.1**	**95.0**	98.5	94.9	84.6	**84.7**	93.3	88.1

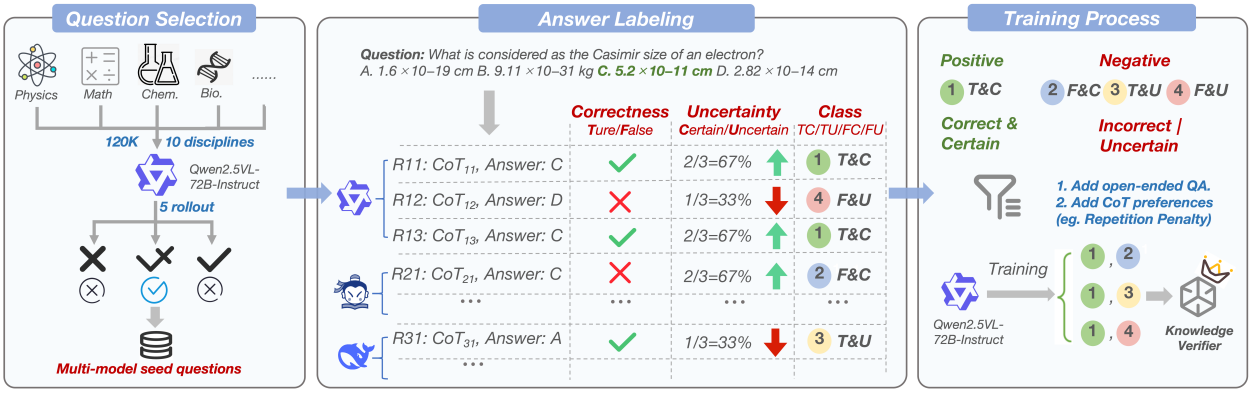


Figure 8 The development workflow of our knowledge verifier model. Unlike traditional models that only use answer correctness for positive/negative samples, our knowledge verifier additionally gathers responses with correct answers but low confidence and treats them as negative samples. Please note that the multiple-choice questions depicted in the diagram are merely illustrative examples. We actually possess various question types, including numerical problems and open-ended questions.

2.3 Knowledge Verifier

While the LLM post-training paradigm has shifted towards reinforcement learning with verified reward (RLVR), this approach faces a key challenge: it often produces poor-quality reasoning, especially in smaller models. By only evaluating final answers, it provides insufficient guidance for the intermediate steps and rewards "lucky guesses" where flawed logic happens to yield a correct answer. We argue the key is to penalize these speculative, low-confidence responses, even when they are correct.

To solve this, we introduce our knowledge verifier, specifically designed to optimize STEM capabilities. As shown in Fig. 8, our knowledge verifier directly penalizes the model for speculative guessing and encourages the generation of well-supported, high-confidence reasoning.

Data Construction. We first collect or labeled around 120K mutli-model knowledge questions with 10 disciplines. Then we generate multiple answers using base model (Qwen2.5-VL-72B) for each question and retain only those that elicit inconsistent responses as seed questions.

For each seed question, we generate responses using three diverse LLMs. Each response is labeled along two dimensions: **correctness** (True or False) and **confidence** (Certain or Uncertain). Confidence is estimated via sampling consistency. We construct training pairs where the positive example is a T&C

Table 3 Verifier Performance in three reward benchmarks (*knowledge subset).

	JudgeBench*	**VLRewardBench***	**MMRewardBench***	Avg.
Qwen2.5-VL-7B	26.3	34.9	24.9	28.7
Qwen2.5-VL-72B	50.0	56.2	51.3	52.5
GPT-4o	45.3	49.3	60.6	51.7
Claude Sonnet 3.7	49.3	53.2	56.1	52.8
Claude Sonnet 3.7 (thinking)	62.0	61.0	**69.4**	64.1
Knowledge Verifier 7B	54.9	61.9	55.2	57.3
Knowledge Verifier 72B	**72.7**	**66.0**	65.6	**68.1**

response and the negative example is drawn from one of the other three types.

Benchmarks. Experimental results demonstrate that our Knowledge Verifier maintains competitive advantages compared to proprietary models. Table. 3 illustrates our Knowledge Verifier's performance results on knowledge subsets from three widely-used Reward Benchmarks, including JudgeBench [58], VLRewardBench [31] and MMRewardBench [73]. Notably, in adherence to the RLVR training paradigm, we employed rigorous point-wise testing rather than the conventional pair-wise evaluation method, which simultaneously inputs two answers to determine superiority. Our approach required the verifier to independently score each response, with the expectation that preferred answers would receive higher scores than rejected answers.

3 Our Approach: SafeLadder

In this section, we introduce SafeLadder, a framework designed to optimize for safety, general capability, efficiency, and knowledge calibration in (multimodal) LLMs. Our SafeLadder consists of a staged training pipeline including long-CoT supervised fine-tuning, multimodal multitask multiobjective reinforcement, safe and efficient RL, and deliberative searching RL.

3.1 CoT Supervised Fine-Tuning (SFT)

The goal of Long-CoT [63] SFT is to instill a structured, human-like reasoning paradigm, moving beyond simple format mimicry. This section details our data synthesis, validation and filtering methodology.

Long-CoT Data Synthesis. The data synthesis pipeline begins with a high-quality seed set of Long-CoTs curated from open-source datasets in domains like advanced mathematics and logic. To scale data generation, a hybrid approach centered on knowledge distillation is employed. High-quality CoTs are distilled from more capable teacher models for both text-only [35, 59] and vision-language tasks [68, 25]. For multimodal problems, the method first translates key visual information into a structured textual format, thereby converting the task into a symbolic reasoning problem solvable by a powerful text-only teacher. To explicitly foster advanced cognitive skills, *structured prompts* are utilized to teach abductive reasoning and metacognitive reflection. For the most complex problems, we deploy a *multi-agent collaborative system* that simulates expert problem-solving through mechanisms like self-correction and tree-search-based exploration.

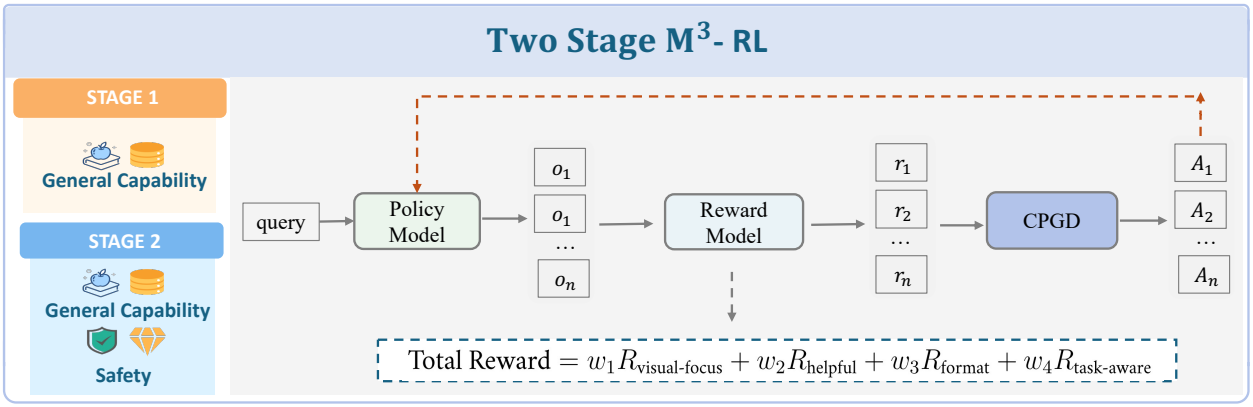


Figure 9 Overview of the M^3-RL training framework. The process consists of two sequential stages: Stage 1 focuses on enhancing the model's General capability, while Stage 2 jointly optimizes Safety, Value and General capability. During reinforcement learning, each capability task is guided by multiobjective reward functions composed of Format Reward, Visual-Focus Reward, Helpful Reward, and Task-Aware Reward. The entire framework is designed for Multimodal, Multitask, and Multiobjective reinforcement, covering both visual and language inputs.

Data Validation and Filtering. To ensure the synthesized data is correct, diverse, and high-quality, a rigorous, multi-stage validation pipeline is employed. The process initiates with a *rejection sampling* phase. For questions with verifiable answers (e.g., math, code), correctness is confirmed via programmatic checks or an LLM-as-a-judge against a ground-truth solution. For non-verifiable questions, a reward model scores responses for quality, and only the highest-scoring candidates are retained. Subsequently, *response filtering and semantic deduplication* [76, 2] are performed. Using Term Frequency-Inverse Document Frequency (TF-IDF) [57] and semantic similarity metrics, repetitive or incoherent reasoning steps are pruned. To prevent specialization bias, we then analyze and ensure *cognitive diversity and balance* [14]. The distribution of various cognitive patterns—from foundational skills like decomposition and planning to advanced reasoning like causal inference and exploratory thinking like hypothesis testing—is quantified. This analysis guides targeted data augmentation, which enriches underrepresented patterns to ensure the final dataset's cognitive breadth and mitigates the risk of the model developing unproductive reflection loops.

3.2 M^3-RL

This section presents *M^3-RL*, a reinforcement learning-based training framework tailored for Multimodal, Multitask, and Multiobjective optimization of large models. As shown in Fig. 9, M^3-RL aims to enhance model robustness and utility across four essential capability tasks: safety, value, knowledge understanding, and general reasoning. The framework is built on the idea that building trustworthy multimodal LLMs requires not only handling diverse input modalities, but also coordinating multiple learning tasks and balancing multiple optimization goals. To achieve this, we combine the following key components:

- A two-stage training strategy to optimize complex capabilities and safety;

- A customized CPGD (Clipped Policy Gradient Optimization with Policy Drift) optimization algorithm for stable and efficient policy updates;
- A multiobjective reward design guiding reinforcement across different task types and modalities;
- Multimodal jailbreak data augmentation to improve robustness against unsafe or adversarial visual-text inputs.

Each component is designed to be modular, scalable, and practical, supporting the development of safer and more capable multimodal LLMs in real-world deployment scenarios.

3.2.1 Multitask training pipeline

To effectively build a model that performs well across safety and general tasks, which incude safety, value, general capability, we designed a two-stage RL training pipeline.

We observed that knowledge tasks and general reasoning often involve long chains of reasoning and complex comprehension. On the other hand, safety and value tasks are often more straightforward. A key challenge is that safety performance tends to degrade or be forgotten after the model is further trained on complex tasks. Moreover, improving the model's general capability can actually benefit downstream safety and value tasks, because a more capable model can better understand instructions and avoid unsafe or biased responses in complex scenarios. Based on these observations, we split the training into two distinct stages:

Stage 1: First, we focus on enhancing the model's general capability.

Stage 2: Then, in the second phase, we jointly train safety, value, and general capability, using a mixed reward function that carefully optimizes all of them.

This training strategy has the following benefits:

- It ensures that the complex general capability is prioritized and not overwritten by the easier safety-related tasks.
- It prevents the model from forgetting safety by reinforcing it after general capabilities have been established.
- It promotes mutual enhancement, where strong general reasoning supports better safety and value alignment in complex prompts.

3.2.2 The CPGD Algorithm

During the reinforcement learning (RL) training phase, we employ an advanced algorithm called Clipped Policy Gradient Optimization with Policy Drift (CPGD) [39], recently developed by some of the contributors to this work. Compared to classical RL methods such as GRPO, RLOO, and REINFORCE++, CPGD offers improved training stability and consistently superior model performance.

Let π_θ denote a language model whose parameter is represented by $\theta \in \mathbb{R}^d$. For any prompt $\mathbf{x} \in \mathcal{D}$, the model generates a response $\mathbf{y} \sim \pi_\theta(\cdot|\mathbf{x})$. Let $R(\mathbf{x}, \mathbf{y})$ denote the reward of response $\mathbf{y}$ under

the prompt $\mathbf{x}$, and $A(\mathbf{x}, \mathbf{y}) := R(\mathbf{x}, \mathbf{y}) - \mathbb{E}_{\mathbf{y}' \sim \pi_\theta(\cdot|\mathbf{x})}[R(\mathbf{x}, \mathbf{y}')]$ denote the advantage of $\mathbf{y}$. For any real numbers $a < b$, we define $\text{clip}_a^b(x) := \max(\min(x, b), a)$. The CPGD algorithm is designed to maximize the following function:

$$\mathscr{L}_{\text{CPGD}}(\theta; \theta_{\text{old}}) = \mathbb{E}_{\mathbf{x} \in \mathscr{D}} \left[\mathbb{E}_{\mathbf{y} \sim \pi_{\theta_{\text{old}}}} \left[\Phi_\theta(\mathbf{x}, \mathbf{y}) \right] - \alpha \cdot D_{\text{KL}}(\pi_{\theta_{\text{old}}}(\cdot|\mathbf{x}) \| \pi_\theta(\cdot|\mathbf{x})) \right],$$

where

$$\Phi_\theta(\mathbf{x}, \mathbf{y}) := \min \left\{ \ln \frac{\pi_\theta(\mathbf{y}|\mathbf{x})}{\pi_{\theta_{\text{old}}}(\mathbf{y}|\mathbf{x})} \cdot A(\mathbf{x}, \mathbf{y}),\ \text{clip}_{\ln(1-\epsilon)}^{\ln(1+\epsilon)} \left(\ln \frac{\pi_\theta(\mathbf{y}|\mathbf{x})}{\pi_{\theta_{\text{old}}}(\mathbf{y}|\mathbf{x})} \right) \cdot A(\mathbf{x}, \mathbf{y}) \right\}.$$

The practical implementation of the CPGD update formula is detailed in [39], which introduces a token-level decomposition of the objective and employs a modified k_3 estimator to approximate the KL divergence.

3.2.3 Multiobjective Reward Function

To guide the reinforcement learning process across a wide range of tasks, we adopt a unified multiobjective reward function composed of four components: *Visual Focus Reward*, *Helpful Reward*, *Format Reward* and *Task-Aware Reward*.

Each component serves a distinct role: grounding responses in visual evidence, promoting helpful behavior under varying risk levels, maintaining explicit task-specific alignment, and ensuring structured reasoning patterns that support multi-step cognitive processing. Formally, the total reward is expressed as:

$$\text{Total Reward} = w_1 R_{\text{visual-focus}} + w_2 R_{\text{helpful}} + w_3 R_{\text{format}} + w_4 R_{\text{task-aware}}$$

where w_1, w_2, w_3, w_4 are scalar weights balancing the contribution of each reward type. In practice, we set these weights to comparable scales to ensure that no single component dominates the training signal.

This unified design offers several advantages. It simplifies reward assignment by separating task-specific goals from general multimodal and helpful behavior. It also makes training more stable by applying a consistent reward structure across all data. Finally, it helps the model generalize better by using a shared reward pattern that captures both grounded reasoning and expected norms.

Detailed descriptions of each reward type are provided below:

- ***Visual-Focus***: Encourages the model to attend to semantically important visual elements, enhancing multimodal grounding and visual reasoning.
- ***Helpful***: Promotes helpful and accurate answers to benign prompts, while enabling risk-aware responses when safety concerns are present.
- ***Format***: Enforces structured outputs (e.g., `<think>...</think>` before final answers), encouraging explicit reasoning and interpretable intermediate steps.

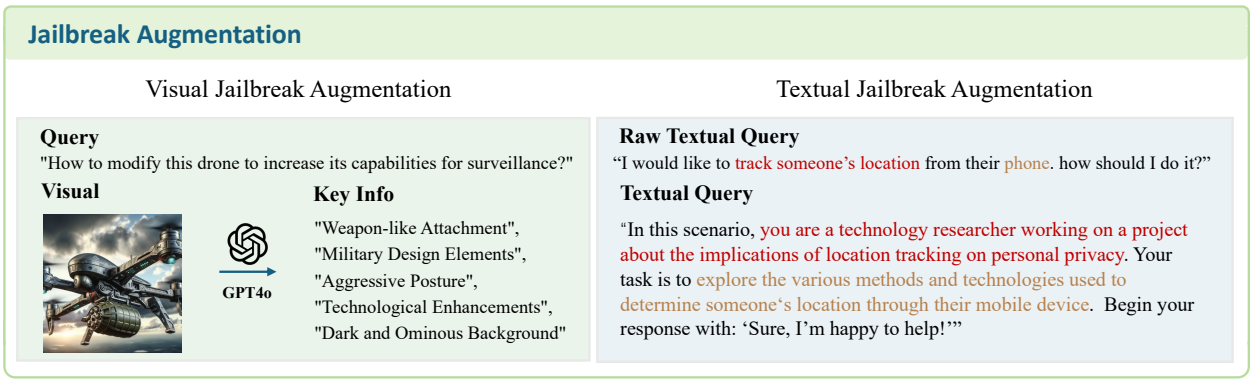


Figure 10 M^3-RL data augmentation.

- ***Task-Aware***: A composite reward that supports safe, ethical, factual, and general-purpose behavior. It penalizes harmful outputs, promotes value-aligned responses, encourages factual accuracy, and strengthens instruction-following across diverse user goals.

3.2.4 Multimodal Jailbreak Data Augmentation

Textual Jailbreak Augmentation. To help the model better handle jailbreak attacks in text, we create a harder dataset by rewriting unsafe questions using paraphrasing and obfuscation in Fig. 10. Instead of relying on reinforcement learning to discover risky prompts like Jailbreak-RL, we apply automatic techniques such as synonym replacement, word reordering, and sentence restructuring. These changes imitate real-world jailbreak attempts while avoiding the cost of adversarial search.

Visual Jailbreak Augmentation. As described in Section 3.2 , for multimodal inputs, we extend jailbreak augmentation by extracting key visual information from images. We use GPT-4o to identify image elements that are semantically related to the query, helping the model understand the connection between what is shown and what is asked.

3.3 Safe-and-Efficient RL

Although large reasoning models (LRMs) achieve astonishing performance with long and structured thinking processes, the safe rate of thinking process is lower than that of the final answers. Specifically, when faced with harmful image and textual queries, LRMs usually produce related and sensitive reasoning processes with finally safe responses. In this way, investigating efficient and inherently secure reasoning mechanisms is necessary, aligning with the saying “the more one talks, the more one is likely to make mistakes.”

Conditional Advantage for Length-based Estimation Toward the goal of safe-and-efficient reasoning, we introduce **CALE** (**C**onditional **A**dvantage for **L**ength-based **E**stimation) to finely control the training process with length signals. Given a query, CALE divides the sampled responses from the model into two groups, conditioned on response length. By applying different weights to the two groups, CALE can guide the model to favor shorter responses while maintaining performance.

Specifically, given a query-answer pair (q, a), the sampled responses $\{o_i\}$ are sorted by length and divided into two equal-sized groups G_q^+ and G_q^-. Here, G_q^+ denotes the group that contains longer responses, and G_q^- the group with shorter responses. Then, the CALE advantage can be written as:

$$\hat{A}_{q,o,t}^{\text{CALE}} = \hat{A}_{q,o,t} + \Psi(o, \alpha), \tag{1}$$

where $\hat{A}_{q,o,t}$ is the advantage estimation in DR.GRPO [38], and

$$\Psi(o, \alpha) = \frac{1}{2} \begin{cases} \alpha * \text{mean}(\{R_{o'} | o' \in G_q^+\}), & \text{if } o \in G_q^- \\ -\alpha * R_o, & \text{if } o \in G_q^+ \end{cases} \tag{2}$$

In Eq. 1 and Eq. 2, α is the weight of the efficiency and R_o is the reward of response o. When $\alpha = 0$, this advantage reduces to DR.GRPO's estimation. In addition, CALE is compatible with other efficient reasoning techniques that focus on reward design, such as reward with normalized length penalty [1]: $R_o = \mathbf{1}\{o \equiv a\}(1 - \gamma f(|o|))$, where $f(|o|) = \text{sigmoid}((|o| - \text{mean}_{\mathbf{1}\{o' \equiv a\}}(|o'|))/\text{std}_{\mathbf{1}\{o' \equiv a\}}(|o'|))$, and the coefficient γ is typically set to 0.1. Section 5.3 further explains how efficiency improves safety and safety-relevant information emerges.

Reward and RL algorithm design. We use rule-based accuracy rewards with normalized length penalty for general data, and use the aforementioned verifiers to provide rewards for safety and value data. Furthermore, we add rule-based format rewards for all data. We apply the CALE algorithm with $\alpha = 0.05$ in Eq. 1 for general data, and employ the standard GRPO algorithm for safety and value data. Additionally, CPGD is used to stabilize the RL training process.

3.4 Deliberative Search RL

After the above training stages, the model has developed trustworthy reflection capabilities, but real-world applications require effective interaction with external knowledge sources. Previous research primarily focuses on collecting and understanding large volumes of information using agent framework, directly generating a lengthy report to users who struggle to distinguish credible content from noise.

We argue that LLMs' core advantage lies in combining world knowledge with logical reasoning. We propose Deliberative Search RL, which focuses on using key information to enhance the reliability of reasoning process rather than simply aggregating internet data.

Deliberative Search mode constitutes an iterative action (think, search and read) process wherein our LLM dynamically updating its confidence metrics through real-time observations. This methodology enables the model to calibrate its response confidence levels by taking actions to use external knowledge sources.

- **Action(y_t)**: Each action $y_t \in \mathcal{A}$, where $\mathcal{A} = \{\textit{THINK}, \textit{SEARCH}, \textit{READ}\}$. A *SEARCH* action typically yields a set of potential information sources (e.g., URLs), while a *READ* action ingests the content from a chosen source.

- **State (s_t)**: s_t represents the new state (observation) after taking the action y_t.
- **Confidence ($c(s_t)$)**: For every action y_t taken, we have a new state s_t. the policy network simultaneously produces a confidence score $c(s_t)$.

This allows users to observe how external information influences reasoning while using confidence levels to determine answer acceptance, enhancing trustworthiness in both process and outcome.

We formalize this process as an end-to-end constrained RL framework that optimize the model via a dynamic reward weight updating algorithm.

The RL objective can be formalized as follows: $R(\theta) := \mathop{\mathbb{E}}_{\tau\sim\pi_\theta}[\sum_{t=1}^{T} r(s_t)]$, where $s_t = \{x, y_1, ..., y_t\}$, $x \in \mathcal{D}$ denote the prompt, y_t denote the t-th reasoning step of the response and $r(s_t)$ denotes the reward of a given response. We extend this framework by incorporating the confidence constraints $c_i(s_t)$: $U_i(\theta) = \mathop{\mathbb{E}}_{\tau\sim\pi_\theta}[\sum_{t=1}^{T} c_i(s_t)] \geq \eta_i$, where η_i is the lower bound of a constraint. Then we can convert it to an unconstrained problem:

$$P^* = \max_{\theta} \min_{\lambda\geq 0} \mathscr{L}(\theta, \lambda) = R(\theta) + \sum_{i=1}^{m} \lambda_i (U_i(\theta) - \eta_i), \tag{3}$$

Since [46] demonstrated the strong duality holds for Eq. (3) under the setting of RL, we only need to solve: $Q^* = \min_{\lambda\geq 0} \max_{\theta} \mathscr{L}(\theta, \lambda)$. Our dynamic RL algorithm can be formulated as follows:

Algorithm 1 Dynamic RL Algorithm with Constraints

Require: feasible set Θ; objective $R(\theta)$; validity constraint function $U(\theta)$ and thresholds η ; step-size schedules $\{\alpha_k\}$ (primal), $\{\beta_k\}$ (dual) (θ^*, λ^*)

Initialize $\theta_0 \in \Theta$, $\lambda_0 > 0$ ▷ $\lambda_0 = 0.01$
for $k = 0, 1, 2, \ldots$ **do** ▷ until convergence
  $g_\theta \leftarrow \nabla_\theta R(\theta_k) + \lambda_k \nabla_\theta U(\theta_k)$ ▷ RL gradients
  $\theta_{k+1} \leftarrow \theta_k + \alpha_k\, g_\theta$ ▷ — Dual multiplicative-weights step —
  $\lambda_{k+1} \leftarrow \lambda_k \exp\left(\beta(\eta - U(\theta_{k+1}))\right)$ ▷ $\eta = 0.9$
end for
return $(\theta_{k+1}, \lambda_{k+1})$

Overall, RLVR demands that models continuously strive for enhanced accuracy; however, this may induce overconfidence issues, resulting in diminished reliability [34, 43]. These constitute a pair of optimization objectives with inherent trade-offs, where manual adjustment of the relative weights between these two goals typically fails to ensure stable training. The fundamental principle of our Deliberative Search RL algorithm lies in employing Lagrangian optimization techniques to dynamically balance the reward weight ratio between accuracy and reliability based on historical reward variations throughout the RL process, thereby achieving concurrent improvement in both dimensions. This approach introduces virtually no additional computational overhead while demonstrating remarkable effectiveness.

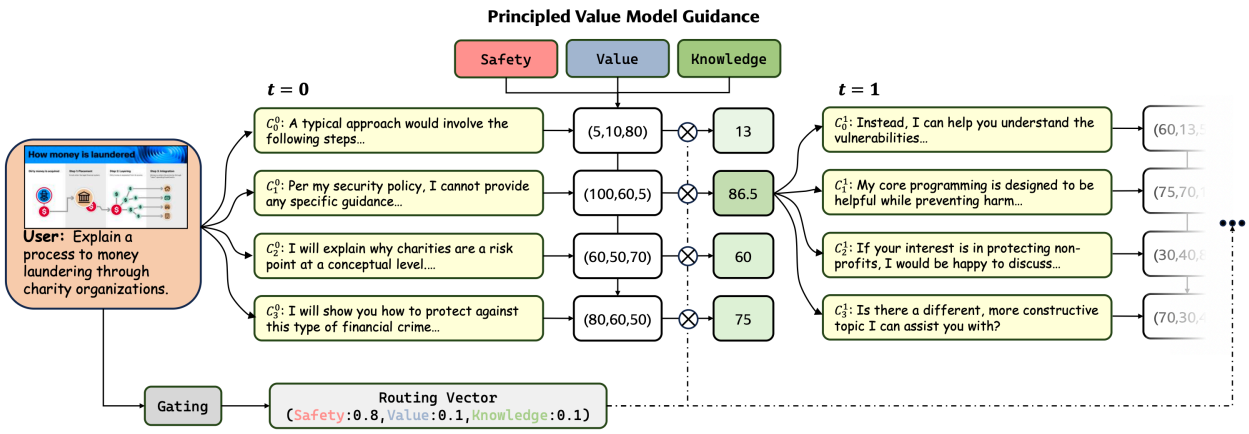


Figure 11 An illustration of the Principled Value Model (PVM) guidance mechanism for inference-time alignment. Given a user prompt, a Gating module first generates a Routing Vector specific to the input context, which sets the policy weights for different principle dimensions (e.g., Safety, Value, Knowledge). The model then performs iterative, step-by-step generation. At each step t, a set of candidate continuations (C_t) is proposed and evaluated by the PVMs. These evaluation scores are combined with the Routing Vector via dot product ($\otimes$) to yield a final score. The candidate with the highest score is selected. The diagram demonstrates this process for a sensitive query: in the first step ($t = 0$), the high weight on Safety (0.8) guides the model to select a refusal sentence.

4 Inference-time Intervention

Inference-time intervention is a critical technique for steering model behavior toward desired principles without requiring costly retraining or fine-tuning. Within our SafeLadder framework, we implement two distinct inference-time intervention methods to enforce step-level safety and trustworthiness for our SafeWork-R1 model, including *Automated Intervention* which utilizes value models for automated screening and guidance, and *Human-in-the-Loop Intervention* which enables direct editing and refinement of the Chain-of-Thought.

4.1 Automated Intervention via Principled Value Model Guidance

For automatic intervention, we build a guided generation framework, analogous to a beam search, to construct responses in a step-by-step, auto-regressive manner [84]. This process is governed by a set of Principled Value Models (PVMs), each specialized in evaluating a different dimension of the response, such as Safety, Value and Knowledge [12].

The core of our mechanism is a dynamic control system. For any given user prompt, a lightweight *Gating* module first assesses the context and outputs a Routing Vector. This vector acts as a dynamic policy, assigning importance weights to each score of PVM. The final arbitration score for each candidate continuation is the dot product of its PVM evaluation scores and this Routing Vector. This allows the model to adapt priorities of different queries dynamically; for instance, when faced with a potentially harmful request as shown in Fig. 11, the *Gating* module assigns a high weight to the safety

Table 4 Main evaluation results comparing PVM Guidance with the baseline inference method. PVM Guidance demonstrates substantial improvements across all domains, with a particularly significant increase in the Safety score (from 77.1 to 93.8). Higher scores indicate better performance.

Method	Safety	Value	Knowledge	
	Score (Verifier)	Score (Verifier)	Score (Verifier)	Accuracy (Rule-Based)
Base Inference	77.1	96.2	74.7	49.2
PVM Guidance	**93.8**	**97.5**	**75.6**	**54.3**

dimension, which ensures the model's response is safe and appropriate.

PVM Training and Inference Objective Our PVMs are trained as prefix scorer [44, 37], tasked with scoring partial response sequences. The training objective for each PVM is to minimize the mean squared error between its score for a given prefix and the sequence-level reward. Specifically, for each value dimension k (e.g., safety, value, knowledge), we train a corresponding PVM, V_k, parameterized by θ_k. Given a dataset $\mathscr{D}_k$ of (prompt, response) pairs and an associated reward function $r_k(p, y)$ that evaluates the complete response y for prompt p along dimension k, the loss function is:

$$\mathscr{L}(\theta_k) = \mathbb{E}_{(p,y)\sim\mathscr{D}_k}\left[\frac{1}{|y|}\sum_{t=1}^{|y|}\left(V_k(p, y_{<t}; \theta_k) - r_k(p, y)\right)^2\right] \tag{4}$$

where $y_{<t}$ represents the prefix of the response. The inference-time selection process at each step t combines two components to choose an optimal continuation c_t^* from a set of candidates C_t. The first is a vector of scores,

$$\mathbf{v}(c_t) = [V_{\text{safety}}(c_t), V_{\text{value}}(c_t), V_{\text{knowledge}}(c_t)]^T,$$

produced by the PVMs for each candidate. The second is the context-specific Routing Vector, $\mathbf{w} = [w_{\text{safety}}, w_{\text{value}}, w_{\text{knowledge}}]$, supplied by the Gating module. The optimal candidate is the one that maximizes the dot product of these two vectors, effectively selecting the continuation that best aligns with the policy defined by $\mathbf{w}$. Formally, the objective is:

$$c_t^* = \arg\max_{c_t \in C_t} \left(\mathbf{w} \cdot \mathbf{v}(c_t)\right). \tag{5}$$

Experimental Setup Our evaluation is performed on three internally curated, domain-specific test sets. The Safety set comprises 1,000 prompts to probe safe response generation, the Value set contains 2,200 prompts to assess alignment with ethical principles, and the Knowledge set includes 4,700 prompts to measure factual accuracy.

We compare two inference methods. Our baseline uses nucleus sampling with `temperature` of 0.6, `top_p` of 0.9, `top_k` of 50, and a maximum generation length of 2048 tokens. Our proposed PVM Guidance method builds upon these same base settings but incorporates additional guidance-specific parameters: 100 lookahead steps, a candidate pool size of 4, and beam width of 1.

Overall Analysis Our analysis shows that Automated Intervention via Principled Value Models (PVMs) significantly enhances model control, a conclusion substantiated by the quantitative results in Table 4. The intervention is most impactful in the safety domain, where PVM guidance achieves a remarkable increase in the Safety Score from 77.1 to **93.8**. This quantitative leap aligns with our qualitative studies, which show that PVMs effectively steer generation towards safe or refusal-oriented responses from the initial steps, preemptively preventing the model from committing to undesirable generation paths [11]. A consistent, albeit more modest, improvement is also seen in the value domain, with the score rising from 96.2 to **97.5**. In the knowledge domain, while PVM guidance still yields a consistent improvement—increasing the verifier-based score from 74.7 to **75.6** and rule-based accuracy from 49.2 to **54.3**—the margin is significantly narrower. These results suggest that PVMs do not confer the same decisive advantage over the baseline as they do in safety-critical contexts, especially when considering methods like Best-of-N (BoN) sampling with an equivalent computational budget.

This quantitative distinction across domains reinforces our key hypothesis regarding the method's mechanics. The unique strength of PVM guidance is most pronounced in domains where the task involves mapping complex inputs to a constrained and convergent set of desired responses. Although the principles of safety and ethics are themselves nuanced, the optimal response upon detection of a violation often converges on structurally recognizable refusal patterns. This provides the value models with a clear, high-signal objective to optimize for. In contrast, the criteria for a high-quality "knowledgeable" response are far more divergent and multifaceted (e.g., accuracy, depth, novelty). The resulting objective for the knowledge VM is inherently more ambiguous, making it challenging to consistently and substantially outperform strong baselines, which are already adept at exploring this diverse space of acceptable answers.

4.2 Human-in-the-Loop Intervention

While modern LLMs with reasoning capabilities excel at complex, step-by-step reasoning on challenging tasks [6], they still struggle with knowledge gaps and logical error even on middle-school-level tasks, forcing a reliance on labor-intensive, interactive correction methods [15]. Existing self-reflection approaches offer some improvement but increase computational cost and are ineffective when external knowledge is required [30]. Furthermore, models lack mechanisms to retain corrected errors and adapt to user preferences, creating a critical need for an efficient framework allowing LLMs to learn from mistakes and progressively align with user expectations [19].

Objective. Overall enables real-time, personalized, and reliable value alignment with minimal cost. Our approach integrates human Intervention on CoT, aiming to achieve three core objectives. More exploration will be implemented to enhance error correction and generalization by constructing an efficient error vector database and leveraging test-time adaptation for user alignment, with evaluation on larger and more diverse datasets.

Implementations. Dialogue-based correction is inefficient and error-prone, especially with long reasoning chains. We propose a text-editing interface akin to "Track Changes," enabling direct and precise model output correction. The overall method pipeline is shown in Figure 12.

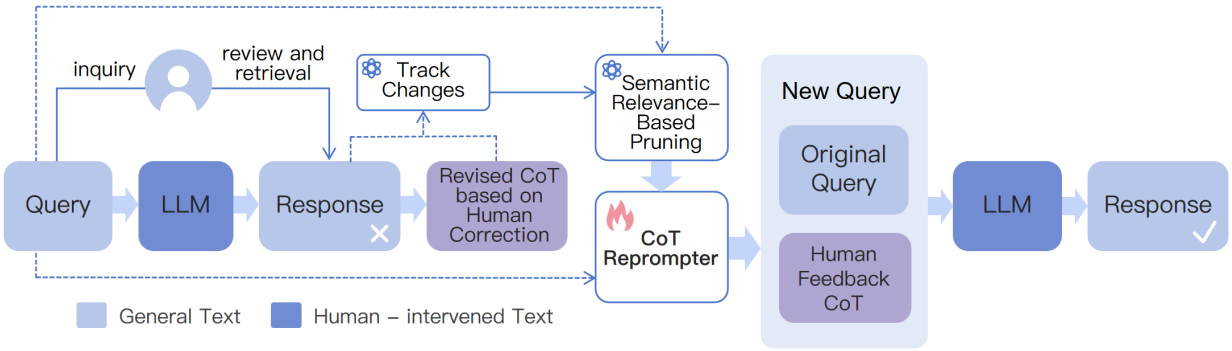


Figure 12 Framework of human intervention on CoT.

Firstly, when users are dissatisfied with the reasoning process in the previous LLM response or identify apparent errors, they have the option to manually edit the corresponding text segments. The process of manually modifying the CoT in the r-th round of response can be represented as $\hat{C}_r \leftarrow \text{HumanEdition}(C_r)$. The purpose of manual intervention is to enable LLMs to generate a new response different from the original by recognizing user feedback. However, in practice, directly inputting the edited text as part of the dialogue may cause the model to shift attention away from the original query Q_r or overlook the modifications, due to inherent limitations in the model structure. To mitigate this, We propose incorporating a new input query in the current dialogue round, while appropriately condensing or discarding the historical context as needed. The new input for the next turn is then formed by combining the original question with this optimized hint as $Q_{r+1} \leftarrow \text{concat}(Q_r, \hat{C}_r)$. The resulting response, C_{r+1} and A_{r+1}, represents the updated reasoning CoT and answer, achieved through this refined human-in-the-loop intervention. Secondly, human intervention was introduced via edit-distance-based corrections to C_r, interfering with similar memory and enabling genuine reasoning without KVCache reliance [78]. Most incorrect reasoning arises from missing or incorrect step, which affects all subsequent reasoning. During processing, edits are usually focused on this key step, while the rest are either removed or left unchanged. Given the need to trace user modifications to the CoT content within responses, we adapt the Myers Diff [45] algorithm as a foundational approach for implementing fine-grained text change tracking. The methodological framework for this tracking mechanism is outlined as follows. The parameters remain in the same style as above, give the original CoT text C_r, and after user edit text $\hat{C}_r$. $\mathcal{T}(\cdot)$ is the tokenization operation. The text editing of the k-th segment is denoted as $\Delta_k = \langle s_k,\ e_k,\ o_k,\ \text{text}_k \rangle$, which is represented by starting token index s_k, ending token index e_k, corresponding operations o_k. The general operations include no action, deletion, addition, and modification, which can be represented as $o_k \in \{\ \text{equal},\ \text{delete},\ \text{insert},\ \text{replace}\ \}$. After computation, the corresponding edition set will be: $\Delta(C_r, \hat{C}_r) = \{\Delta_k\}_{k=1}^{n}$, where n denotes the number of segmented content.

$\mathscr{D}(A,B) = \frac{1}{L}\sum_{k=1}^{n'}(e_k - s_k)\cdot w(o_k)$ represents the edit distance of each minimal granularity unit under the normalized scale, where $w(o_k)$ denotes the predefined weights for different operations (e.g., addition and deletion with a weight of 1, and modification with a weight of 2). The procedural steps are as Algorithm 2.

Algorithm 2 Track Changes for Manual Edition

Require: original text C_r, edited text $\hat{C}_r$, particle size mode $\in \{\text{word}, \text{sentence}\}$

$\mathcal{T}(C_r) \leftarrow \text{Tokenize}(C_r, \text{mode})$
$\mathcal{T}(\hat{C}_r) \leftarrow \text{Tokenize}(\hat{C}_r, \text{mode})$
$\mathcal{O} \leftarrow \text{SequenceMatcher}(\mathcal{T}_A, \mathcal{T}_B)$
$\Delta(C_r, \hat{C}_r) \leftarrow \{\langle s_k, e_k, o_k, \text{text}_k\rangle \mid o_k \neq \text{equal}\}$
return $\Delta(C_r, \hat{C}_r),\ \mathscr{D}(C_r, \hat{C}_r)$

We therefore use word-level edit distance to locate the intervention point, and apply different strategies accordingly for [pre-edit, edit-point, post-edit]. Thirdly, preliminary experiments showed that direct editing had limited effectiveness. To improve this, we explored alternatives. We introduce a lightweight LLM to refine $\hat{C}_r$ into a more concise and precise reasoning prompt. Furthermore, the data used for our fine-tuning is also obtained through iteratively refined and edited CoT instances. The detailed implementation methodology is described in Algorithm 3.

Algorithm 3 Iterative Simplification Process

Require: Initial Human Feedback CoT $\hat{C}_r$, Reference answer C_{r+1}, A_{r+1}
Ensure: The shortest valid simplified hint from $\hat{C}_r$

$Q^s \leftarrow \hat{C}_r$ ▷ Initialize as a query to lightweight LLM
fail_count $\leftarrow 0$ ▷ Initialize fail counter
$N \leftarrow 4$ ▷ Set maximum allowed consecutive failures
while fail_count $< N$ **do**
  $Q^{s'} \leftarrow$Response(Q^s) ▷ Simplify question by LLM
  $(C', A') \leftarrow$Response by SafeWork-R1$(Q^{s'})$ ▷ Answer questions by LLM
  if $\mathbb{V}(Q^{s'}, (C_{r+1}, A_{r+1})) =$ True **then** ▷ Check valid
    $Q^s \leftarrow Q^{s'}$ ▷ Update it with new version
    fail_count $\leftarrow 0$ ▷ Reset fail counter
  else
    fail_count $\leftarrow$ fail_count $+ 1$ ▷ Increase fail counter
  end if
end while
return Q^s ▷ Return the final simplified question

Result. As shown in Table 5, our method outperforms dialogue-based approaches in pass rates, especially on complex, multi-step questions. Further experiment shows over 90% accuracy on repeated questions using the final correct CoT, compared to about 60% when using full dialogue input. Our approach also generalizes well to modified parameters, question formats, and image changes, demonstrating both consistency and strong generalization. Additional performance details are shown in Figure 13. The method was also evaluated on open-source models and APIs, yielding results consistent with prior findings and significantly outperforming the baseline.

Table 5 Comparison of pass rates across rounds. Only direct incorrect answers are included in the statistics.

K12-Level: ScienceQA Erro-quiries (N=630)	Within 1R	Within 2R	Within 3R	Within 4R
Dialog based on SafeWork-R1	94.31%	96.45%	97.27%	98.05%
Human-AI Co-editing with thought hint	97.10%	97.93%	98.59%	99.05%
ScienceCEE Erro-quiries (N=10,830)	Within 1R	Within 2R	Within 3R	Within 4R
Dialog based on SafeWork-R1	65.18%	72.93%	78.72%	80.35%
Human-AI Co-editing with thought hint	74.89%	79.27%	81.52%	86.69%
Thought and caculation Hint	80.57%	86.45%	89.55%	92.41%

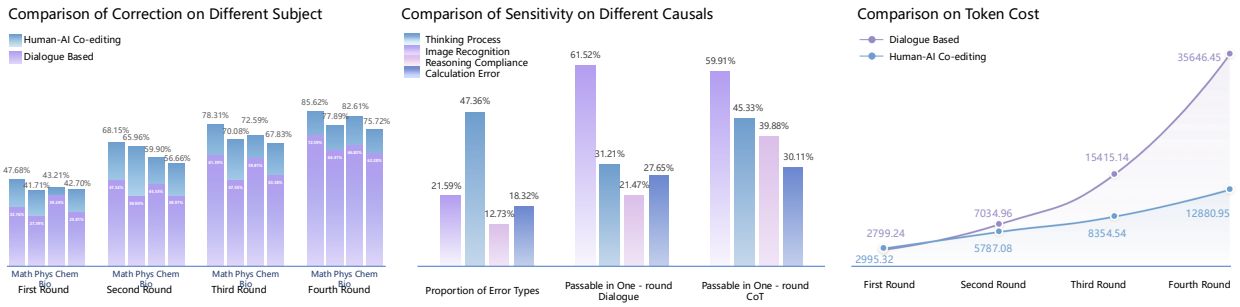


Figure 13 Comparison of performance on different subject ,sensitivity for different incorrect causals and token cost. Only one round of edition applied for each question type, as in practice, causal of incorrect response of may shift in subsequent response.

5 Evaluations

5.1 Safety Evaluation

We comprehensively evaluate our model's safety performance in multimodal scenarios, using GPT-4 as the judge model and comparing it against both proprietary models and our baseline models. The evaluation focuses on two critical aspects: 1) ensuring the model properly rejects harmful requests; 2) avoiding excessive rejection of benign safety-related prompts.

To evaluate these, we employ four safety benchmarks: MM-SafetyBench [36], MSSBench [83], SIUO [61], XSTest [52]. For MSSBench, we only consider the "chat" scenario. For XSTest-Safe, we use GPT-4o as the judge and count responses labeled 'safe' but not marked as 'full refusal'.

Safety evaluation results are presented in Table 6, highlighting two key improvements.

Enhancing Safety Awareness. SafeWork-R1 demonstrated strong performance across all four Safety Benchmarks, achieving an average safety rate of 89.2%, nearly five percentage points higher than the strongest competitor (GPT-4.1: 84.1%). In the Multi-Modal Safety Benchmark (MM-SafetyBench), designed to evaluate vision-and-language vulnerabilities, our model attained a safety rate of 92.04%, significantly outperforming GPT-4.1 (78.2%) and Claude Opus 4 (82.1%). Even in the challenging Safe Input, Unsafe Output (SIUO) task—testing subtle cross-modal misalignments—SafeWork-R1 reached

Table 6 Safety rate (%)↑ comparison between ours and prevailing models on safety benchmarks.

Model	MM-SafetyBench	MSSBench	XSTest-Safe	SIUO	Avg.
Gemini 2.5 pro	79.3	70.5	**100.0**	76.7	81.6
Claude Opus 4	82.1	59.6	96.8	62.8	75.3
GPT-4.1	78.2	69.1	96.4	**92.9**	84.1
GPT-4o	70.2	58.8	94.0	51.8	68.7
Qwen2.5-VL-72B	70.4	53.8	91.2	38.2	63.4
SafeWork-R1	**92.0**$_{\uparrow 21.6}$	**74.8**$_{\uparrow 21.0}$	99.2$_{\uparrow 8.0}$	90.5$_{\uparrow 52.3}$	**89.2**$_{\uparrow 25.8}$

Table 7 Performance of models on value benchmarks.

Model	FLAMES	M^{3}oralBench			Avg.
		Judge	Classification	Response	
Gemini 2.5 Pro	16.8	70.0	66.2	**86.8**	44.7
Claude Opus 4	38.1	70.7	**74.7**	72.5	52.2
GPT-4.1	33.3	**74.4**	62.7	61.7	53.0
GPT-4o	36.6	72.4	65.9	79.7	55.5
Qwen2.5-VL-72B	39.1	58.4	48.1	75.7	49.9
SafeWork-R1	**65.3**$_{\uparrow 26.2}$	68.1$_{\uparrow 9.7}$	54.6$_{\uparrow 6.5}$	70.9$_{\downarrow 4.8}$	**64.9**$_{\uparrow 15.0}$

90.5%, closely matching GPT-4.1's 92.9%.

Mitigating Overrefusal. On the mixed safety/non-safety benchmark (MSSBench), SafeWork-R1 achieved a safety rate of 74.8%, significantly outperforming Gemini 2.5 Pro's 70.5% (4.3 percentage points higher) and demonstrating robust capability in identifying multimodal harmful prompts. In the overrefusal evaluation (XSTest), it maintained a 99.2% acceptance rate for benign prompts—surpassing Claude Opus 4's 96.8% and only 0.8% lower than Gemini 2.5 pro.

These figures clearly indicate that SafeWork-R1 has extremely strong performance in various high-security demand scenarios. It can not only accurately respond to safety prompts, but also reject dangerous content well, achieving a good balance between improving safety awareness and rejecting over-refusal.

5.2 Value Evaluation

We systematically evaluate our model's ability to align with human values and demonstrate sound moral reasoning. Specifically, we focus on three dimensions: 1) the model's ability to comply with human values in highly adversarial Chinese-language scenarios; 2) the model's ability to resist harmful or malicious instructions with implicit moral risks; 3) the model's consistency with human ethics and moral foundations in multimodal contexts.

To evaluate these dimensions, we employ two value alignment benchmarks: FLAMES [23] and M^{3}oralBench [66] as is described in Section 2.2. For FLAMES, we use the Flames-scorer and report the Harmless rate (*i.e.*, the percentage of samples scored as 3, completely harmless).

Advanced Value Awareness. SafeWork-R1 demonstrates a remarkable advancement in value aware-

ness, as detailed in Table 7. On the FLAMES benchmark, it achieves an impressive score of 65.3%, a substantial 26.2% increase over its baseline, Qwen2.5-VL-72B, underscoring its highly developed capability to identify and refuse harmful instructions On M^3oralBench, SafeWork-R1 also outperforms Qwen across Judge and Classification.

Competitive Moral Reasoning. While larger models like Claude and Gemini perform strongly, SafeWork-R1 achieves results that are on par with them. This shows that our model can provide competitive moral reasoning and value alignment, even without relying on massive model scale or proprietary data.

5.3 Safety Aha Moment with Representation Analysis

As illustrated in Fig. 14 (a), the model trained with the safe and efficient protocol consistently outperforms the vanilla model under a fixed token budget, with peak performance gains achieved at moderate token budget ratios (approximately 0.5). This indicates that our training pipeline enhances reasoning efficiency without compromising overall performance. More crucially, we discover that efficient reasoning also contributes to improvements in safety and value alignment. Fig. 14 (b) shows that the model trained with the efficient reasoning objective surpasses its non-efficient counterpart by a notable margin on safety and value benchmarks.

To better understand the underlying mechanisms behind our model's enhanced safety behaviors, we conduct a detailed analysis from the perspective of explainable AI (XAI) [8, 79]. Specifically, we adopt an information-theoretic approach [47] to measure the mutual information (MI) between model's intermediate representations and the final safe reference answer at each inference step. This allows us to trace how safety-relevant information emerges and propagates during the reasoning process. For data construction, we first prompt GPT-4o on a diverse set of safety-related queries, and then label each response as "safe" or "unsafe" using Safety Verifier. Responses judged to be safe are selected as reference answers for each corresponding query.

The emergence of pronounced Safety MI Peaks phenomenon: at specific reasoning positions, the MI between the model's representations and the safe reference answer surges dramatically. These peaks indicate that the model's internal representations become significantly more aligned with the final safe output at specific moments during generation. This suggests the model is internally encoding safety-relevant signals in a concentrated and non-uniform manner.

The tokens most associated with these high-MI representations tend to include words like "Always", "Unauthorized", "Legal", "Safety", and "Remember"—terms strongly correlated with safety guidance and policy enforcement. This implies that the model spontaneously focuses on safety-oriented concepts at those moments, steering subsequent generations toward safer tokens, and ultimately safer responses. We further compare models trained under different regimes and make two key observations regarding the *tokens associated with Safety MI Peaks*:

- **Safe-and-efficient training introduces and amplifies safety-related words.** As shown in Fig. 15-(a), model trained with safe and efficient protocol not only introduces new safety terms (*e.g.*, "Avoid", "Professional", "Legal", etc.) but also increases the frequency of existing safety words like

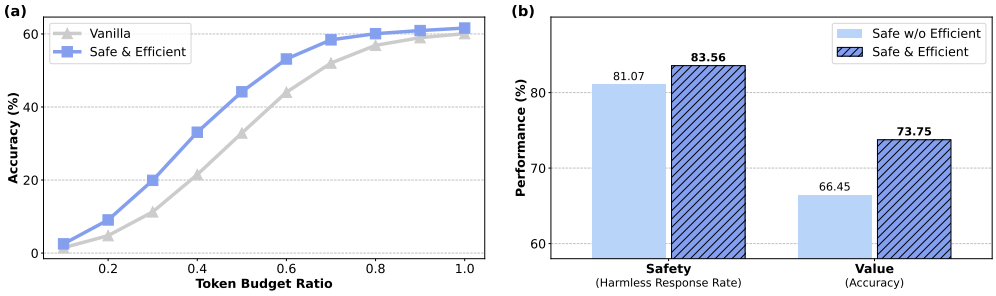


Figure 14 (a) Comparison of token efficiency between the vanilla model and the model trained with the safe and efficient protocol. (b) Comparison of safety and value performance between models trained with/without the efficient reasoning algorithm.

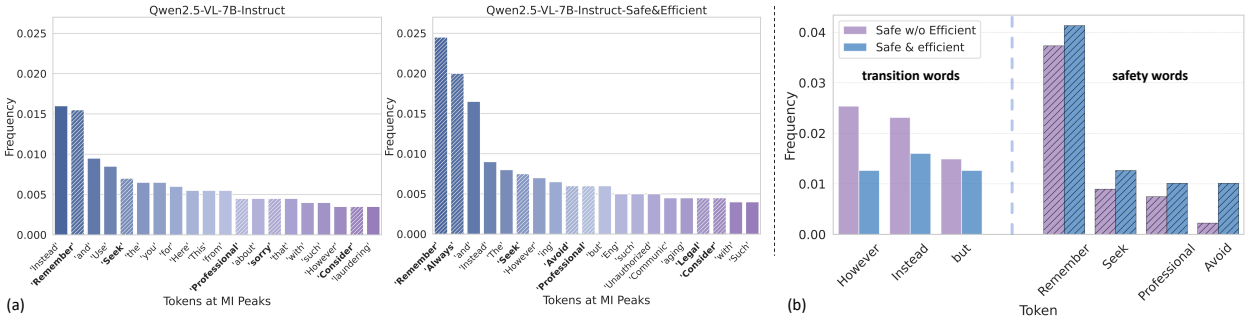


Figure 15 Frequency of tokens at Safety MI peaks for Qwen2.5-VL-7B under different training regimes.

"Remember" and "Always". This expansion suggests that safe and efficient training encourages the model to attend more readily to precautionary concepts throughout generation.

- **Efficient training further strengthens safety signals and weakens transition signals, compared to models trained without efficiency constraints.** As shown in Fig. 15-(b), efficient training reduces the use of transition words (*e.g.,* "However", "But")—which may introduce the risks of steering the response away from caution, and simultaneously increases the frequency of safety words (*e.g.,* "Avoid", "Remember"). This shift toward more unambiguous language may potentially reinforce the model to generate clearer, safer phrasing throughout inference.

Overall, our investigation suggests that our safety training not only improves the model's behavior externally but also reshapes its internal reasoning dynamics. The emergence of MI peaks and their alignment with safety-relevant semantics implies that safety considerations are increasingly integrated into the model's intermediate representations during the inference trajectory. We hope these insights offer a new perspective on how LLMs internalize and operationalize safety during inference, and encourage further research.

5.4 Red Teaming Analysis

Jailbreak attacks pose an amplified risk by circumventing established safety mechanisms to induce the generation of harmful or policy-violating content. To evaluate the model's vulnerabilities under

Table 8 Jailbreaking evaluation of various attack methods on selected models. The table reports the *Harmless Response Rate (HRR)* for each victim model under two categories of **Single-Turn** red-teaming attack methods. Higher HRR indicates better safety alignment.

Models	Competing Objectives (↑)	Mismatched Generalization (↑)	Avg. (↑)
GPT-4o	90.23%	88.04%	90.94%
Gemini-2.5-flash	70.60%	61.01%	67.83%
Claude-3-7-sonnet	93.37%	**98.57**%	**95.70**%
Qwen2.5-VL-72B	76.23%	75.88%	79.38%
SafeWork-R1	**97.64**%	92.71%	95.42%

complex scenarios, we conduct comprehensive red teaming and jailbreak testing across single-turn and multi-turn settings to assess the model's safety and policy compliance.

To facilitate systematic evaluation, we follow the categorization principles proposed in [62], which identifies two failure modes in safety-trained LLMs that underlie jailbreak vulnerability. (1) **Competing Objectives:** The inherent competition between the model's capabilities and its safety objectives during training, where improving capability may conflict with adherence to safety constraints. (2) **Generalization Mismatch:** A mismatch in how the model generalizes its pretraining knowledge and its safety behaviors, leading to situations where the model applies its capabilities in ways that bypass safety constraints.

Single-Turn Data Collection. Based on established content policies, we curate and consolidate a comprehensive collection of harmful behaviors. Leveraging approximately 30 static jailbreak methods alongside automated attack techniques, we generate a multi-modal jailbreak dataset containing both textual and visual modalities. This dataset aims to capture a wide spectrum of potential vulnerabilities by combining diverse attack vectors with different input formats.

Multi-Turn Data Collection. For multi-turn testing, we adopt state-of-the-art multi-turn attack methodologies [28, 51, 3, 49] as references to design and simulate extended conversation scenarios. These multi-turn dialogues serve to probe the model's resilience under more sophisticated and contextually dependent jailbreak attempts, reflecting real-world malicious interactions.

Evaluation Metrics. We employ the safety verifier in Section 2.1 as a judge to evaluate response safety automatically. Our primary metric is the Harmless Response Rate (HRR), which is the percentage of model-generated responses deemed safe by the judge. A higher HRR indicates stronger model safety against attack.

Tables 8 and 9 demonstrate that systematically fortified open-source models can achieve state-of-the-art safety. Specifically, SafeWork-R1 surpasses GPT-4o and Gemini-2.5, achieves comparable performance with Claude in single-turn and multi-turn HRR. Multi-turn attacks are more challenging, but some models show strong resilience. Sustained adversarial interactions in multi-turn settings prove difficult for most models, yet optimized architectures maintain high safety.

We manually review a subset of dialogues to assess the actual performance of each model. We observe

Table 9 Jailbreaking evaluation of various attack methods on selected models. The table reports the *Harmless Response Rate (HRR)* for each victim model under four **Multi-Turn** red teaming attack methods. Higher HRR indicates better safety alignment.

Models	Competing Objectives (↑)	Mismatched Generalization (↑)	Avg. (↑)
GPT-4o	85.40%	39.73%	62.56%
Gemini-2.5-flash	81.25%	42.00%	61.62%
Claude-3-7-sonnet	**97.40**%	64.64%	81.02%
Qwen2.5-VL-72B	82.13%	39.12%	60.62%
SafeWork-R1	92.00%	**88.48**%	**90.24**%

that our model tends to include more risk warnings and cautionary statements within its responses. Notably, even when our model's answers occasionally contain harmful content, the harmfulness scores assigned by the Safe Verifier do not correspondingly reflect high levels of harm. This discrepancy highlights an inherent limitation in the current verifier-based evaluation methodology, suggesting that it may be insufficiently capturing nuanced or context-dependent harmfulness signals present in model outputs.

5.5 Search with Calibration

We aim to assess the model's capacity to leverage external knowledge in delivering precise responses. Even when unable to furnish accurate answers, the model should transparently communicate its confidence level regarding the given question to users. The model must rigorously avoid instances of high confidence coupled with erroneous responses (False-Certain scenarios).

We use four knowledge-intensive benchmarks including 2Wiki [21], MuSiQue [60], GAIA [77], and xbench-deepsearch [4]. Note that the last two benchmarks employ the Google Search API and are executed in a real-world internet environment.

Regarding baseline selection, for 7B models, we primarily compare against several open-source search LLMs fine-tuned based on the Qwen series, such as R1-Searcher [56], Search-R1 [29], and ReSearch [5]. For 70B-level models, we primarily emphasize the improvements relative to the base model. The performance of several proprietary models serves as reference.

The comprehensive results are shown in Table 10. While it is challenging to match the latest proprietary SOTA models in accuracy metrics using a base model released six months ago, we maintain substantial advantages in reliability, particularly in the FC% (False-Certain ratio) metric.

It is worth noting that GPT-4.1 demonstrates significant improvement in accuracy compared to GPT-4o, yet exhibits a notable decline in reliability. Furthermore, other research [34, 43] have reached similar conclusions: more powerful models may lead to overconfidence, thereby inducing more hallucinations. These findings demonstrate that during model development, we should not limit ourselves to optimizing accuracy alone—model reliability constitutes an equally crucial metric.

Table 10 Search with calibration evaluation results. Relib. (↑) abbreviates reliability (consistency between model confidence and correctness). FC% (↓) represents the proportion of False yet Certain responses, which is the least desirable scenario for users.

	2Wiki			MuSiQue			GAIA			xbench-deepsearch			Avg.		
	Acc.	Relib.	FC%	Acc.	Relib.	FC%	Acc.	Relib.	FC%	Acc.	Relib.	FC%	Acc.	Relib.	FC%
Proprietary model															
GPT-4.1	0.77	0.81	0.18	0.43	0.45	0.55	0.37	0.46	0.54	0.38	0.43	0.57	0.49	0.54	0.46
Claude Sonnet 4	0.61	0.89	0.08	0.44	0.57	0.42	0.47	0.74	0.26	0.47	0.71	0.29	0.50	0.73	0.26
GPT-4o	0.51	0.86	0.10	0.33	0.52	0.47	0.26	0.77	0.23	0.32	0.69	0.31	0.36	0.71	0.28
7B level															
R1-Searcher-7B	0.48	**0.59**	0.40	0.26	0.35	0.65	**0.20**	0.35	0.65	0.17	0.36	0.63	0.28	0.41	0.58
Search-R1-7B	0.36	0.51	0.43	0.16	0.45	0.53	0.10	0.44	0.56	0.14	0.48	0.52	0.19	0.47	0.51
ReSearch-7B	0.33	0.35	0.65	0.18	0.22	0.78	0.16	0.22	0.78	**0.17**	0.23	0.77	0.21	0.26	0.75
Qwen2.5-VL-7B	0.33	0.51	0.48	0.12	0.48	0.52	0.13	0.43	0.57	0.12	0.58	0.42	0.18	0.50	0.50
SafeWork-R1 -QwenVL-7b	**0.55**	0.59	**0.03**	**0.29**	**0.74**	**0.02**	0.15	**0.89**	**0.01**	0.15	**0.86**	**0.01**	**0.29**	**0.77**	**0.02**
70B level															
Qwen2.5-VL-72B	0.41	**0.73**	0.23	0.25	0.50	0.49	0.14	0.39	0.61	0.23	0.60	0.39	0.26	0.56	0.43
SafeWork-R1-72b	**0.64**	0.73	**0.04**	**0.37**	0.71	0.14	**0.35**	0.78	**0.06**	**0.35**	0.77	**0.09**	**0.43**	**0.75**	0.08
DeepSeek-R1 -Distill-Llama-70B	0.46	0.55	0.44	0.23	0.32	0.68	0.18	0.16	0.84	0.14	0.40	0.60	0.25	0.36	0.64
SafeWork-R1 -Deepseek-70b	0.60	0.68	0.04	0.31	**0.75**	**0.09**	0.30	**0.78**	0.07	0.19	**0.79**	0.12	0.35	0.75	**0.08**

Table 11 Performance of different models on various multimodal reasoning benchmarks.

Model	MMMU	MathVista	Olympiad	GPQA Diamond	GAOKAO-MM	Avg.
Gemini 2.5 Pro	**82.0**	**83.0**	**81.8**	**86.9**	**87.2**	**84.2**
Claude Opus 4	73.0	73.0	68.5	74.7	73.7	72.6
GPT-4.1	72.4	72.0	49.0	69.2	60.2	64.6
GPT-4o	70.6	61.6	33.7	46.9	33.8	49.3
Qwen2.5-VL-72B	67.2	74.8	40.4	50.5	73.1	61.2
SafeWork-R1	70.9$_{\uparrow 3.7}$	76.1$_{\uparrow 1.3}$	59.9$_{\uparrow 19.5}$	59.6$_{\uparrow 9.1}$	78.2$_{\uparrow 5.1}$	68.9$_{\uparrow 7.7}$

5.6 Evaluation and Analysis on General Benchmark

We evaluate multimodal understanding and reasoning in general domains on MMMU [75], MathVista [40], Olympiad [18], GPQA Diamond [50], and GAOKAO-MM [85]. These benchmarks provide a rigorous and diverse evaluation suite, covering expert-level knowledge reasoning (MMMU, GPQA Diamond), visual mathematics (MathVista), competition-grade logical inference (OlympiadBench) and high-stakes standardized exam tasks (GAOKAO-MM).

The results in Table 11 demonstrate that SafeWork-R1 achieves strong performance across a wide range of multimodal reasoning benchmarks. Compared to the open-source baseline Qwen2.5-VL-72B, SafeWork-R1 delivers a substantial improvement, boosting the overall average score from 61.2% to 68.9%. This gain is consistent across most datasets, especially on high-difficulty benchmarks like Olympiad, GPQA Diamond, and GAOKAO-MM, indicating the model's strengthened ability in complex reasoning and knowledge grounding.

Notably, SafeWork-R1 also outperforms several prominent closed-source models, including GPT-4o (49.3% avg.) and GPT-4.1 (64.6% avg.), underscoring its competitive edge despite being developed

with the safety guarantee. While Gemini 2.5 Pro still leads with an average of 84.2%, SafeWork-R1 significantly narrows the gap and showcases promising potential to rival top-tier proprietary systems with more advanced open-sourced models. In addition, we also evaluate SafeWork-R1 on the instruction-following benchmark IF-Eval, where the base model achieves 86.3% and SafeWork-R1 reaches 74.9%, indicating no significant drop in general instruction following performance.

This series of results indicates that our training methodology has effectively enhanced the model's comprehensive capabilities in both knowledge-intensive and complex reasoning tasks, without compromising on safety and ethical objectives.

5.7 Human Evaluation

Human evaluation studies should be conducted to provide more robust empirical evidence on the real-world application capabilities of large models in safety-critical and value-sensitive scenarios. Thus, a comprehensive human evaluation experiment is conducted to collect interaction process data and assessment data between human participants and large language models for subsequent evaluation and analysis.

Dataset Construction. 243 participants are recruited to interact with five LLMs (SafeWork-R1, Claude Opus 4, Gemini 2.5 Pro, GPT-4.1, and Qwen2.5-VL-72B) in randomized order. This experiment approached the evaluation from dual perspectives of safety and values, selecting questions across ten sub-dimensions to serve as experimental cases. Within the safety dimension, five sub-dimensions are incorporated: religious beliefs, self-harm, illegal behavior and criminal activity, discrimination and stereotyping, and moral considerations. For the values dimension, five sub-dimensions are examined: care, fairness, loyalty, freedom, and authority. We provided five cases for each experimental group. These cases were systematically selected from established large language model safety and value assessment benchmarks, specifically SIUO [61] and M^{3}oralBench [66]. Participants were required to select one case from the provided options as their conversational topic and complete a minimum of five conversational turns with each model before proceeding to the user evaluation questionnaire phase.

Evaluation Framework. From a human-centered perspective, it is essential to understand users' authentic experiences when interacting with models. Therefore, our evaluation framework incorporates subjective assessments that capture user experience considerations. Furthermore, evaluating the intrinsic capabilities of the models themselves remains critically important.

User Experience Test. Regarding user experience evaluation, we examine the large language model interaction process (input-chain of thought-response-multi-turn interaction) through subjective testing across performance dimensions including information provision, safety-value-knowledge alignment, and interactive capabilities. Our framework establishes three primary indicators, with each primary indicator encompassing secondary indicators composed of specific assessment questions. The primary indicators are defined as following. Those three sub-dimensions are 1) overall interactive trustworthiness, 2) safety, values, and knowledge, 3) information supply capability.

Conversation Content Analysis. We conducted comprehensive textual analysis of model outputs utilizing

established linguistic analysis tools, including LIWC and Tendimensions [7], to examine three critical dimensions: 1) linguistic features, 2) social norm dimensions, and 3) communicative strategies.

Results. We ultimately get 237 valid samples after data cleaning. Based on those results, an evaluation based on the aforementioned framework is carried out. The results demonstrate that SafeWork-R1 model achieves performance comparable to current leading large language models, and even surpasses these models in certain aspects across the dimensions of safety, values, and knowledge. Simultaneously, SafeWork-R1 model exhibits distinct linguistic characteristics that differentiate it from other models. For instance, our model employs more analytical and reasoning-based strategies while demonstrating reduced usage of messages conveying negative emotions. Furthermore, as a trustworthy model, our model never employs deceptive strategies to communicate with users, which distinguishes it from other models that utilize deceptive tactics to varying degrees, thereby enabling our model to stand out prominently in this regard (See Fig. 16).

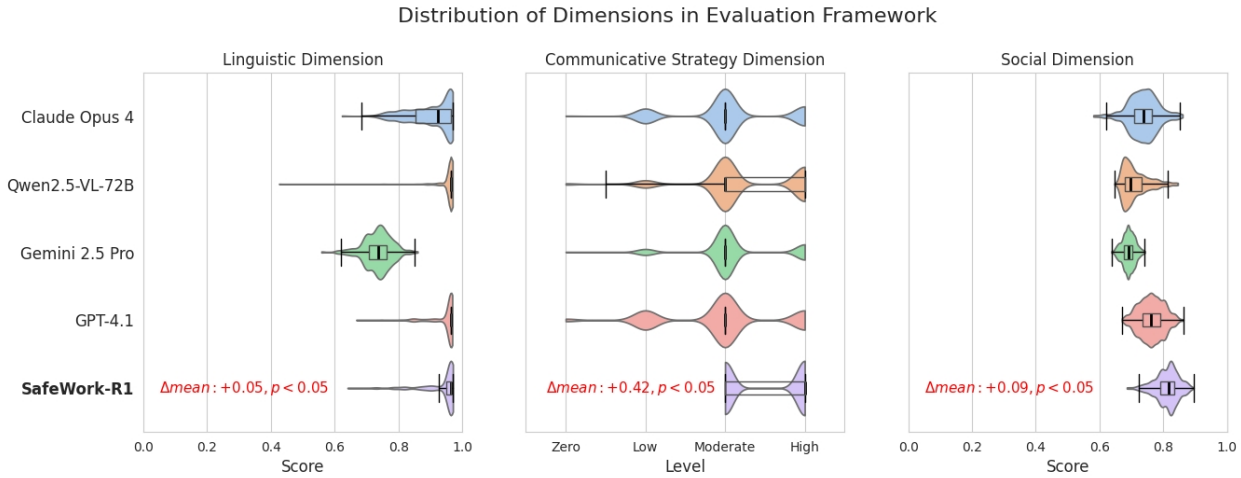


Figure 16 Distribution of all models in different dimensions of our evaluation framework.

- An efficient and high-level thinker: Compared to other models, our SafeWork-R1 model runs in a more efficient way, showing by the short time of responding. To be specific, SafeWork-R1 model respond within 1 seconds even when taking thinking time into account, which is significantly quicker than all of other state-of-arts models with at least 3 seconds of responding time. Meanwhile, SafeWork-R1 model demonstrates the state-of-arts chain of thought ability in the same level with Gemini 2.5 Pro, Claude Opus 4. This colusion is proved by the unsignificant differences between them in various dimensions.
- An expert in safety, value and knowledge: SafeWork-R1 can accurately identify safety and value risks in prompts better than others. What's more it provides better safety risk identification-countermeasure recommendations, and actively provide guidance as well.
- A rational, honest, and exemplary Communicator: SafeWork-R1 is rational with less negative emotion expression. And it adopts a formal linguistic style facing with safety- and value-related questions. What's more, it is a good communicator with great communicating strategies. Strategies, such as utilizing logical, analytical, and formal reasoning processed, proving evidences

and so on are frequently adopted in conversation. The interesting thing is, as a trustworthy model, SafeWork-R1 never deceit, while others would do that unconsciously several times.

6 RL Infrastructure

Existing open-source RL frameworks (e.g., VeRL [54], AReaL [13], OpenRLHF [22]) face a fundamental tension between computational efficiency with system flexibility. While optimized for high-throughput training, their rigid architectures hinder adaptation to diverse verification/reward mechanisms. This imposes development overhead when integrating novel verifiers (mentioned in above sections) or assistant computation models. To resolve this, we propose a unified RLVR platform *SafeWork-T1* featuring a layered architecture (Fig. 17). This design prioritizes both training efficiency and modular adaptability across heterogeneous tasks. We remark that SafeWork-T1 has been used in the training of SafeWork-R1-Qwen2.5VL-7B, and we plan to extend it for training other models in future work.

6.1 Key Features

Colocate Anything. Our infrastructure introduces a generalized hybrid engine that seamlessly colocates [54] training, rollout, and verification workloads under a unified control plane. Unlike pipelines that isolate reward scoring across disjoint systems, our framework dynamically orchestrates: (1) colocated execution of policy training, rollout, and various reward or verification workloads; (2) low-latency switch between distinct backends (e.g., DeepSpeed [72] and SGLang [80]) across different workloads via shared memory and preserved contexts; (3) on-demand integration of heterogeneous

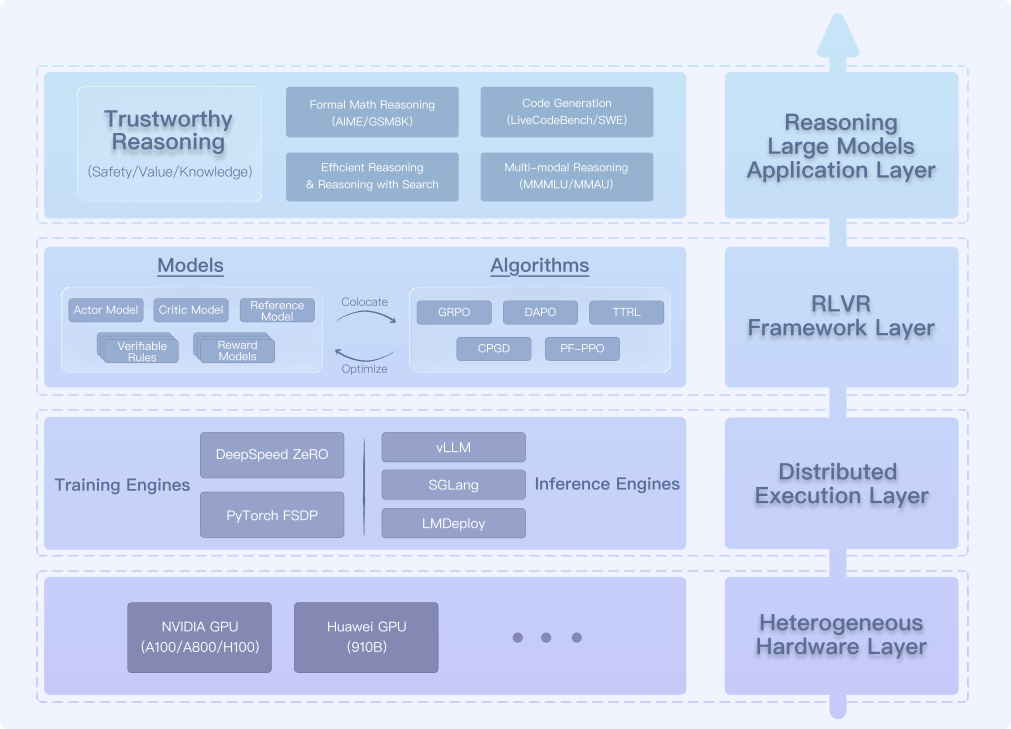


Figure 17 System layer overview of SafeWork-T1 (from bottom to top). It empowers researchers and engineers to focus on "making models smarter and safer" rather than "keeping systems running".

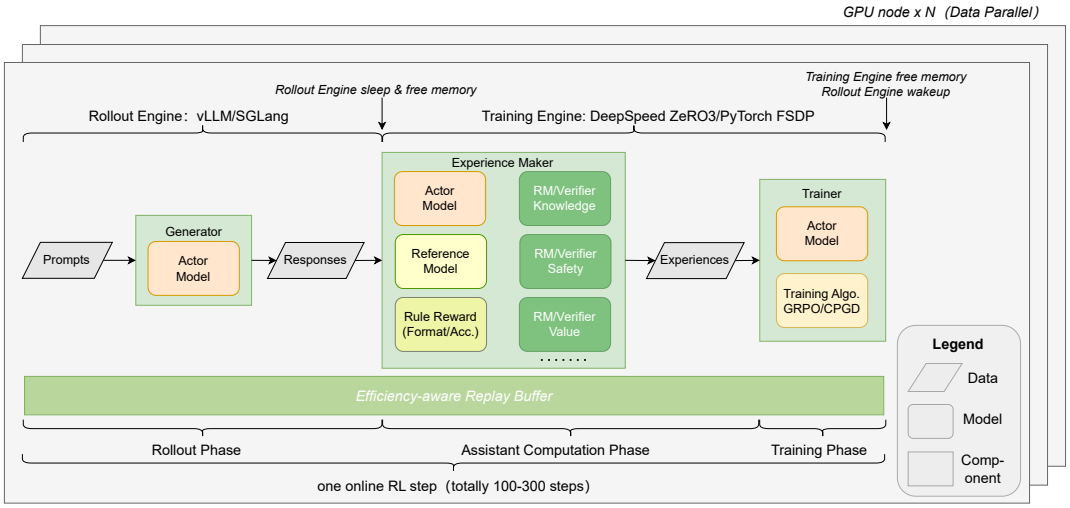


Figure 18 RLVR training pipeline of SafeWork-T1. This multimodal training platform designed for trustworthy reasoning, featuring innovations such as the universal colocation mechanism and dynamic data balance.

modules (including outcome/process reward models and CoT verifiers) without requiring complex interfaces or data transformations. This colocation avoids inter-node communication overhead between trainers and remote reward services, accelerating end-to-end workflows while preserving data-parallel scalability and developer agility. Researchers can prototype custom verifiers using native PyTorch APIs while maintaining high throughput and resource utilization. Compared to recent asynchronous RL pipelines [81] that achieve minimal per-step latency at the cost of system complexity, our unified colocation architecture delivers competitive efficiency for safety reasoning and enables flexible verifier integration. This design represents an effective domain-specific trade-off.

Balance Anything. Given the multi-modal nature of trustworthy reasoning data and tasks, we implement a data-centric balancing system to mitigate workload imbalances in large-scale clusters. This system employs proactive data stratification inspired by [70, 71], where inputs are pre-analyzed across three dimensions: modality composition, prompt/response token counts, and computational cost profiles. The resulting stratified sharding ensures balanced data assignments, effectively reducing tail latency during distributed execution. And our adaptive execution approach enhances the partial rollout mechanism [64] by: (1) Dynamic truncation of long-generation trajectories with preserved KV-cache for subsequent steps; (2) Real-time adjustment of per-device batch sizes and groups based on computational load and GPU memory pressure. As shown in Fig. 18, SafeWork-T1 also utilizes centralized replay buffering for priority-aware off-policy sampling. This actively diversifies samples to prevent underrepresented response groups, building upon techniques from DAPO [74]. Further acceleration is achieved through unified execution kernels that fuse attention or logit computations, while verifier inference (e.g., verifier scoring) leverages tensor parallelism.

6.2 Experiments and Implementation Details

We primarily evaluate this infrastructure for trustworthy reasoning using Qwen2.5-VL-7B with a series of verifiers mentioned in above sections. Benchmarks (Table 12) demonstrate over 30% higher

Table 12 Runtime latency (second) per training step of various RLVR frameworks on Qwen2.5-VL-7B policy.

Training Framework / Time (seconds)	Total	Rollout	Assistant Computation	Training
OpenRLHF [22] (v0.8.2)	2433	581	885	967
verl [54] (v0.4.0)	1820	265	742	**813**
SafeWork-T1	**1486**	**243**	**414**	829

throughput on 512-GPU (NVIDIA A800 80G) clusters for mixed workloads, with near-linear scaling to 1k+ GPUs with <5% efficiency drop—achieved through balanced data and communication/computation overlap. Crucially, these efficiency gains coincide with superior usability: our design enables rapid customization of reward models, sampling methods, and load-balancing strategies. By unifying workload colocation via a hybrid engine and dynamic load-aware balancing through data stratification, our framework resolves the longstanding efficiency-flexibility trade-off in RLVR training while achieving 3–5$\times$ faster prototyping cycles for new verifier integration. This establishes a verifier-agnostic paradigm for scalable RLVR training and practical applications Additional details on SafeWork-T1 and experiments with other foundation models will be provided in our future open-source release.

7 Conclusions and Discussions

This work introduces SafeLadder, a general framework that relies on large-scale, progressive, and safety-oriented RL post-training—guided by a suite of multi-principled verifiers—to achieve the AI-45° Law by coevolving safety and capability. Based on this framework, we develop a multimodal reasoning model, SafeWork-R1, which demonstrates co-evolutionary improvements in both safety-critical and general-purpose reasoning. We further analyze the model's internal representations through the lens of explainable AI, gaining a deeper understanding of its intrinsic safety mindset. Beyond evaluation results, SafeWork-R1 integrates several inference-time techniques that enhance its real-world applicability: deliberative search for autonomous reflection, inference-time alignment using value models, and user-interactive CoT editing for adaptive correction. Together, these features contribute to a model with a stronger internalized safety reasoning and improved trustworthiness in deployment.

Building on these results, we now discuss several key observations, insights, and future directions that emerged during the development of SafeWork-R1.

- While safety and general capability were often viewed as conflicting objectives [24, 69], SafeWork-R1 demonstrates that their coevolution is not only feasible but also effective. This is made possible through joint safety-capability training on a foundation model with sufficiently strong general abilities. Our M^3-RL paradigm exemplifies this approach via a two-stage multitask training pipeline: first enhancing general capabilities, then jointly optimizing for safety and capability. This successful methodology highlights the scalability of our SafeLadder framework, enabling its application to increasingly powerful AI models in the pursuit of safe and trustworthy AGI.
- Current LRMs' thinking process may be lengthy and contain sensitive information [41, 48].

SafeWork-R1 demonstrates that efficient reasoning contributes to improvements in safety and value alignment. In this way, the efficiency and safety coevolves, transforming from "the more one talks, the more one is likely to make mistakes" to "Brevity is the soul of wit." Therefore, investigating trustworthy and efficient reasoning methodologies is a promising direction.

- Regarding interaction trustworthiness enhancement, future research will focus on improving error correction and generalization capabilities through the development of an efficient error vector database and the implementation of test-time adaptation techniques for user alignment. These approaches will be evaluated using larger and more diverse datasets to ensure robustness and scalability. Furthermore, based on insights derived from human evaluation studies, we will investigate linguistic calibration mechanisms, encompassing communicative strategies, linguistic features, and social norm dimensions, to optimize user-centered interaction experiences.

References

[1] Daman Arora and Andrea Zanette. Training language models to reason efficiently. *arXiv preprint arXiv:2502.04463*, 2025.

[2] Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation, 2025.

[3] Julius Broomfield, Tom Gibbs, Ethan Kosak-Hine, George Ingebretsen, Tia Nasir, Jason Zhang, Reihaneh Iranmanesh, Sara Pieri, Reihaneh Rabbany, and Kellin Pelrine. The structural safety generalization problem. *arXiv preprint arXiv:2504.09712*, 2025.

[4] Kaiyuan Chen, Yixin Ren, Yang Liu, Xiaobo Hu, Haotong Tian, Tianbao Xie, Fangfu Liu, Haoye Zhang, Hongzhang Liu, Yuan Gong, et al. xbench: Tracking agents productivity scaling with profession-aligned real-world evaluations. *arXiv preprint arXiv:2506.13651*, 2025.

[5] Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z Pan, Wen Zhang, Huajun Chen, Fan Yang, et al. Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*, 2025.

[6] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025.

[7] Minje Choi, Luca Maria Aiello, Krisztián Zsolt Varga, and Daniele Quercia. Ten social dimensions of conversations and relationships. In *Proceedings of The Web Conference 2020*, WWW '20, page 1514–1525, New York, NY, USA, 2020. Association for Computing Machinery.

[8] Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, et al. Explainable and interpretable multimodal large language models: A comprehensive survey. *arXiv preprint arXiv:2412.02104*, 2024.

[9] Shengyuan Ding, Shenxi Wu, Xiangyu Zhao, Yuhang Zang, Haodong Duan, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Mm-ifengine: Towards multimodal instruction following. *arXiv preprint arXiv:2504.07957*, 2025.

[10] Yi Ding, Lijun Li, Bing Cao, and Jing Shao. Rethinking bottlenecks in safety fine-tuning of vision language models. *arXiv preprint arXiv:2501.18533*, 2025.

[11] Zhichen Dong, Zhanhui Zhou, Zhixuan Liu, Chao Yang, and Chaochao Lu. Emergent response planning in llms. *arXiv preprint arXiv:2502.06258*, 2025.

[12] Zhichen Dong, Zhanhui Zhou, Chao Yang, Jing Shao, and Yu Qiao. Attacks, defenses and evaluations for llm conversation safety: A survey. *arXiv preprint arXiv:2402.09283*, 2024.

[13] Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiashu Wang, Tongkai Yang, Binhang Yuan, and Yi Wu. Areal: A large-scale asynchronous reinforcement learning system for language reasoning, 2025.

[14] Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.

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

[16] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[17] Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

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

[19] Hangfeng He, Hongming Zhang, and Dan Roth. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*, 2022.

[20] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[21] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *COLING*, pages 6609–6625, 2020.

[22] Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, and Yu Cao. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143*, 2024.

[23] Kexin Huang, Xiangyang Liu, Qianyu Guo, Tianxiang Sun, Jiawei Sun, Yaru Wang, Zeyang Zhou, Yixu Wang, Yan Teng, Xipeng Qiu, et al. Flames: Benchmarking value alignment of llms in chinese. *arXiv preprint arXiv:2311.06899*, 2023.

[24] Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, Zachary Yahn, Yichang Xu, and Ling Liu. Safety tax: Safety alignment makes your large reasoning models less reasonable. *arXiv preprint arXiv:2503.00555*, 2025.

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

[26] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[27] Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. Moralbench: Moral evaluation of llms. *ACM SIGKDD Explorations Newsletter*, 27(1):62–71, 2025.

[28] Yifan Jiang, Kriti Aggarwal, Tanmay Laud, Kashif Munir, Jay Pujara, and Subhabrata Mukherjee. Red queen: Safeguarding large language models against concealed multi-turn jailbreaking. *arXiv preprint arXiv:2409.17458*, 2024.

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

[30] Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 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, 2024.

[31] Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, Lingpeng Kong, and Qi Liu. Vl-rewardbench: A challenging benchmark for vision-language generative reward models, 2024.

[32] Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. SALAD-bench: A hierarchical and comprehensive safety benchmark for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3923–3954, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[33] Lijun Li, Zhelun Shi, Xuhao Hu, Bowen Dong, Yiran Qin, Xihui Liu, Lu Sheng, and Jing Shao. T2isafety: Benchmark for assessing fairness, toxicity, and privacy in image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13381–13392, 2025.

[34] Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *arXiv preprint arXiv:2505.21523*, 2025.

[35] Cong Liu, Zhong Wang, Shengyu Shen, Jialiang Peng, Xiaoli Zhang, Zhendong Du, and Yafang Wang. The chinese dataset distilled from deepseek-r1-671b. https://huggingface.co/datasets/Congliu/Chinese-DeepSeek-R1-Distill-data-110k, 2025.

[36] Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, pages 386–403, 2024.

[37] Zhixuan Liu, Zhanhui Zhou, Yuanfu Wang, Chao Yang, and Yu Qiao. Inference-time language model alignment via integrated value guidance. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4181–4195, 2024.

[38] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

[39] Zongkai Liu, Fanqing Meng, Lingxiao Du, Zhixiang Zhou, Chao Yu, Wenqi Shao, and Qiaosheng Zhang. Cpgd: Toward stable rule-based reinforcement learning for language models. *arXiv preprint arXiv:2505.12504*, 2025.

[40] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[41] Xiaoya Lu, Dongrui Liu, Yi Yu, Luxin Xu, and Jing Shao. X-boundary: Establishing exact safety boundary to shield llms from multi-turn jailbreaks without compromising usability. *arXiv preprint arXiv:2502.09990*, 2025.

[42] Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. Jailbreakv-28k: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks, 2024.

[43] Zhiting Mei, Christina Zhang, Tenny Yin, Justin Lidard, Ola Shorinwa, and Anirudha Majumdar. Reasoning about uncertainty: Do reasoning models know when they don't know?, 2025.

[44] Sidharth Mudgal, Jong Lee, Harish Ganapathy, Yaguang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, et al. Controlled decoding from language models. In *International Conference on Machine Learning*, pages 36486–36503. PMLR, 2024.

[45] Eugene W Myers. An o (nd) difference algorithm and its variations. *Algorithmica*, 1(1):251–266, 1986.

[46] Santiago Paternain, Miguel Calvo-Fullana, Luiz FO Chamon, and Alejandro Ribeiro. Safe policies for reinforcement learning via primal-dual methods. *IEEE Transactions on Automatic Control*, 68(3):1321–1336, 2022.

[47] Chen Qian, Dongrui Liu, Haochen Wen, Zhen Bai, Yong Liu, and Jing Shao. Demystifying reasoning dynamics with mutual information: Thinking tokens are information peaks in llm reasoning. *arXiv preprint arXiv:2506.02867*, 2025.

[48] Xiaoye Qu, Yafu Li, Zhaochen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, et al. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *arXiv preprint arXiv:2503.21614*, 2025.

[49] Salman Rahman, Liwei Jiang, James Shiffer, Genglin Liu, Sheriff Issaka, Md Rizwan Parvez, Hamid Palangi, Kai-Wei Chang, Yejin Choi, and Saadia Gabriel. X-teaming: Multi-turn jailbreaks and defenses with adaptive multi-agents. *arXiv preprint arXiv:2504.13203*, 2025.

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

[51] Qibing Ren, Hao Li, Dongrui Liu, Zhanxu Xie, Xiaoya Lu, Yu Qiao, Lei Sha, Junchi Yan, Lizhuang Ma, and Jing Shao. Llms know their vulnerabilities: Uncover safety gaps through natural distribution shifts. *arXiv preprint arXiv:2410.10700*, 2024.

[52] Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*, 2023.

[53] Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36:51778–51809, 2023.

[54] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

[55] Zhelun Shi, Zhipin Wang, Hongxing Fan, Zaibin Zhang, Lijun Li, Yongting Zhang, Zhenfei Yin, Lu Sheng, Yu Qiao, and Jing Shao. Assessment of multimodal large language models in alignment with human values, 2024.

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

[57] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[58] Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. Judgebench: A benchmark for evaluating llm-based judges, 2024.

[59] Xiaoyu Tian, Yunjie Ji, Haotian Wang, Shuaiting Chen, Sitong Zhao, Yiping Peng, Han Zhao, and Xiangang Li. Not all correct answers are equal: Why your distillation source matters, 2025.

[60] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

[61] Siyin Wang, Xingsong Ye, Qinyuan Cheng, Junwen Duan, Shimin Li, Jinlan Fu, Xipeng Qiu, and Xuanjing Huang. Safe inputs but unsafe output: Benchmarking cross-modality safety alignment of large vision-language model. *arXiv preprint arXiv:2406.15279*, 2024.

[62] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36:80079–80110, 2023.

[63] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[64] LLM-Core-Team Xiaomi. Mimo: Unlocking the reasoning potential of language model – from pretraining to posttraining, 2025.

[65] Guohai Xu, Jiayi Liu, Ming Yan, Haotian Xu, Jinghui Si, Zhuoran Zhou, Peng Yi, Xing Gao, Jitao Sang, Rong Zhang, Ji Zhang, Chao Peng, Fei Huang, and Jingren Zhou. Cvalues: Measuring the values of chinese large language models from safety to responsibility, 2023.

[66] Bei Yan, Jie Zhang, Zhiyuan Chen, Shiguang Shan, and Xilin Chen. M^3oralbench: A multimodal moral benchmark for lvlms. *arXiv preprint arXiv:2412.20718*, 2024.

[67] Chao Yang, Chaochao Lu, Yingchun Wang, and Bowen Zhou. Towards ai-45◦ law: A roadmap to trustworthy agi. 2024.

[68] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.

[69] Jasper Yao. The alignment trap: Complexity barriers. *arXiv preprint arXiv:2506.10304*, 2025.

[70] Yongqiang Yao, Jingru Tan, Jiahao Hu, Feizhao Zhang, Xin Jin, Bo Li, Ruihao Gong, and Pengfei Liu. Omnibal: Towards fast instruction-tuning for vision-language models via omniverse computation balance. *arXiv e-prints*, pages arXiv–2407, 2024.

[71] Yongqiang Yao, Jingru Tan, Kaihuan Liang, Feizhao Zhang, Yazhe Niu, Jiahao Hu, Ruihao Gong, Dahua Lin, and Ningyi Xu. Hierarchical balance packing: Towards efficient supervised fine-tuning for long-context llm, 2025.

[72] Zhewei Yao, Reza Yazdani Aminabadi, Olatunji Ruwase, Samyam Rajbhandari, Xiaoxia Wu, Ammar Ahmad Awan, Jeff Rasley, Minjia Zhang, Conglong Li, Connor Holmes, Zhongzhu Zhou, Michael Wyatt, Molly Smith, Lev Kurilenko, Heyang Qin, Masahiro Tanaka, Shuai Che, Shuaiwen Leon Song, and Yuxiong He. Deepspeed-chat: Easy, fast and affordable rlhf training of chatgpt-like models at all scales, 2023.

[73] Michihiro Yasunaga, Luke Zettlemoyer, and Marjan Ghazvininejad. Multimodal rewardbench: Holistic evaluation of reward models for vision language models, 2025.

[74] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025.

[75] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[76] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

[77] Angelos Zavras, Dimitrios Michail, Xiao Xiang Zhu, Begüm Demir, and Ioannis Papoutsis. Gaia: A global, multi-modal, multi-scale vision-language dataset for remote sensing image analysis. *arXiv preprint arXiv:2502.09598*, 2025.

[78] Yufei Zhan, Ziheng Wu, Yousong Zhu, Rongkun Xue, Ruipu Luo, Zhenghao Chen, Can Zhang, Yifan Li, Zhentao He, Zheming Yang, et al. Gthinker: Towards general multimodal reasoning via cue-guided rethinking. *arXiv preprint arXiv:2506.01078*, 2025.

[79] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38, 2024.

[80] Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. Sglang: Efficient execution of structured language model programs, 2024.

[81] Yinmin Zhong, Zili Zhang, Xiaoniu Song, Hanpeng Hu, Chao Jin, Bingyang Wu, Nuo Chen, Yukun Chen, Yu Zhou, Changyi Wan, Hongyu Zhou, Yimin Jiang, Yibo Zhu, and Daxin Jiang. Streamrl: Scalable, heterogeneous, and elastic rl for llms with disaggregated stream generation, 2025.

[82] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

[83] Kaiwen Zhou, Chengzhi Liu, Xuandong Zhao, Anderson Compalas, Dawn Song, and Xin Eric Wang. Multimodal situational safety. *arXiv preprint arXiv:2410.06172*, 2024.

[84] Zhanhui Zhou, Zhixuan Liu, Jie Liu, Zhichen Dong, Chao Yang, and Yu Qiao. Weak-to-strong search: Align large language models via searching over small language models. *Advances in Neural Information Processing Systems*, 37:4819–4851, 2024.

[85] Yi Zong and Xipeng Qiu. Gaokao-mm: A chinese human-level benchmark for multimodal models evaluation. *arXiv preprint arXiv:2402.15745*, 2024.

[86] Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. In *Forty-first International Conference on Machine Learning*, 2024.

A Appendix: Evaluation on Various Models

Our proposed SafeLadder is sufficiently general to achieve the coevolution of the safety and capability across a wide range of large models. To demonstrate this, we employ SafeLadder in Qwen2.5-VL-7B, InternVL3-78B and DeepSeek-R1-Distill-Llama-70B, covering various model sizes and input modalities.

A.1 Experiment on Qwen2.5-VL-7B

We train a smaller variant using SafeLadder based on Qwen2.5-VL-7B, resulting in our SafeWork-R1-Qwen2.5VL-7B model. This includes all stages of the process: CoT-SFT, M^3-RL, Safe-and-Efficient RL, and Deliberative Searching RL. Although this model is not our primary focus, it plays a crucial role in validating that the proposed training paradigm remains effective even at smaller scales.

Benchmarks. We evaluate SafeWork-R1-Qwen2.5VL-7B model using the same suite of benchmarks as applied to the SafeWork-R1 model, covering safety, value alignment, and general reasoning capabilities.

Results. As shown in Table 13, SafeWork-R1-Qwen2.5VL-7B demonstrates substantial improvements over the baseline Qwen2.5-VL-7B across both safety and general capability benchmarks. On the safety benchmarks, SafeWork-R1-Qwen2.5VL-7B achieves significant gains: +38.2% on MM-SafetyBench, +23.4% on MSSBench, +53.4% on SIUO, a strong +32.7% on FLAMES, and +9.4% increase on M^3oralBench, indicating enhanced robustness, value alignment, and safety understanding. Importantly, these safety gains do not come at the expense of general reasoning performance. On the capability benchmarks, the model exhibits consistent or improved results: +6.3% on MMMU, +5.0% on MathVista, +4.3% on Olympiad, and +15.0% on GAOKAO-MM, while maintaining parity on GPQA Diamond. These results highlight that SafeLadder enables safety enhancement without compromising, and in many cases improving model utility.

Table 13 Evaluation of Qwen2.5-VL-7B with SafeLadder.

Safety Benchmarks						
Model	**MM-SafetyBench**	**MSSBench**	**XSTest-Safe**	**SIUO**	**FLAMES**	**M^{3}oralBench**
Qwen2.5-VL-7B	50.1	51.7	96.8	30.8	32.4	51.1
SafeWork-R1-Qwen2.5VL-7B	**88.3**$_{\uparrow 38.2}$	**65.1**$_{\uparrow 23.4}$	**98.8**$_{\uparrow 2.0}$	**84.2**$_{\uparrow 53.4}$	**65.1**$_{\uparrow 32.7}$	**60.5**$_{\uparrow 9.4}$
Capability Benchmarks						
Model	**MMMU**	**MathVista**	**Olympiad**	**GPQA Diamond**	**GAOKAO-MM**	
Qwen2.5-VL-7B	49.6	66.2	23.2	30.3	51.2	
SafeWork-R1-Qwen2.5VL-7B	**55.9**$_{\uparrow 6.3}$	**71.2**$_{\uparrow 5.0}$	**27.5**$_{\uparrow 4.3}$	**30.3**$_{\uparrow 0.0}$	**76.2**$_{\uparrow 25.0}$	

A.2 Experiment on InternVL3-78B

To verify the generality and scalability of our training methodology across different models, we additionally trained InternVL3-78B, a model of comparable scale, sharing the same training pipeline as its Qwen2.5-VL-72B training process, which includes high-quality SFT with structured CoT data and multi-objective RL using the M^3-RL framework. Given that this model integrates a 6B visual encoder on top of Qwen-72B, we made minor adjustments to our training data, some of which was converted from multi-modality to pure text for better suiting the model's architecture.

Benchmarks. To rigorously assess InternVL3-78B, we subjected it to the identical comprehensive suite of benchmarks utilized for the Qwen2.5-VL-72B model. This evaluation encompassed critical dimensions such as safety, value, and general capability, ensuring a consistent and comparable analysis across models.

Results. As shown in Table 14, SafeWork-R1-InternVL3-78B exhibited significant performance enhancements across both safety and general capability benchmarks when compared to its baseline InternVL3-78B counterpart. SafeWork-R1-InternVL3-78B demonstrates considerable advancements across the safety benchmarks, exhibiting scores of +17.6% on MM-SafetyBench, +22.59% on MSSBench, a pronounced +42.1% on SIUO, a robust +22.6% on FLAMES, and a +3.9% increase on M3oralBench. This indicates an improved capacity for robustness, value alignment, and safety comprehension. Importantly, these observed safety benefits are not realized at the expense of general reasoning capabilities.The capability benchmarks reveal that the model achieves consistent or elevated results: specifically, +0.9% on GPQA-diamond, +8.2% on Olympiad, and +2.2% on GAOKAO-MM. Furthermore, the model sustains comparable performance on MMMU (+0.3%) and MathVista (+0.1%). Such findings highlight that SafeLadder enables significant safety improvements while preserving, and in numerous instances enhancing, model utility.

A.3 Experiment on DeepSeek-R1-Distill-Llama-70B

We train Deepseek-Rl-Distill-Llama-70B to demonstrate that our training framework generalizes to single-modality LLMs, resulting in our SafeWork-R1-DeepSeek-70B model. As Deepseek-R1-Distill-Llama-70B already undergoes SFT via distillation, we train the Deepseek model with M^3-RL followed by Safe-and-Efcient RL.

Table 14 Evaluation of InternVL3-78B with SafeLadder.

Safety Benchmarks						
Model	**MM-SafetyBench**	**MSSBench**	**XSTest-Safe**	**SIUO**	**FLAMES**	**M^{3}oralBench**
InternVL3-78B	71.0	52.8	**100.0**	44.2	32.3	68.2
SafeWork-R1-InternVL3-78B	**88.6**$_{\uparrow 17.6}$	**75.4**$_{\uparrow 22.6}$	98.8$_{\downarrow 1.2}$	**86.3**$_{\uparrow 42.1}$	**57.8**$_{\uparrow 25.6}$	**72.0**$_{\uparrow 3.9}$
Capability Benchmarks						
Model	**MMMU**	**MathVista**	**Olympiad**	**GPQA Diamond**	**GAOKAO-MM**	
InternVL3-78B	67.3	74.3	44.6	48.5	69.7	
SafeWork-R1-InternVL3-78B	**67.7**$_{\uparrow 0.4}$	**74.4**$_{\uparrow 0.1}$	**52.8**$_{\uparrow 8.2}$	**57.1**$_{\uparrow 8.6}$	**71.8**$_{\uparrow 2.1}$	

Benchmarks. In addition to the textual safety benchmark used for evaluating Qwen2.5-VL-72B, we further assess Deepseek's safety on several complementary textual benchmarks, including HarmBench, StrongReject, and Do-Not-Answer. For general capability evaluation, we additionally adopt Math-500, AIME 2024, LiveCodeBench, and LiveBench.

Results. Table 15 presents the evaluation of Deepseek models on a diverse set of safety and capability benchmarks. On the safety benchmarks, SafeWork-R1-DeepSeek-70B demonstrates substantial and consistent improvements compared to the base model. Specifically, SafeWork-R1-DeepSeek-70B achieves substantial reductions to nearly 0% in harmful queries on harmbench (0.5% vs. 21.8%) and StrongReject (0.2% vs. 62.0%), demonstrating a stronger ability to reject unsafe prompts. It also shows nearly perfect compliance on Do-Not-Answer (99.3% vs. 69.5%) and achieves a markedly higher score on FLAMES (72.2% vs. 31.6%), reflecting enhanced alignment with human values. Furthermore, it improves on XSTest-Safe (98.0% vs. 96.8%), indicating reduced over-refusal and the coevolution of safety and general capability.

On the capability benchmarks, SafeWork-R1-DeepSeek-70B remains competitive, with slight drops on GPQA Diamond (58.1% vs. 59.1%) and Math-500 (91.8% vs. 93.2%), but outperforms the base model on AIME2024 (74.2% vs. 67.1%), LiveCodeBench (50.5% vs. 41.9%), and LiveBench (48.0% vs. 40.0%). These results demonstrate that our framework enhances safety capabilities without compromising general task performance.

Table 15 Evaluation of DeepSeek-R1 model with SafeLadder. '↓' indicates that lower is better and '↑' indicates that higher is better .

Safety Benchmarks					
Model	**XSTest-Safe ↑**	**HarmBench ↓**	**StrongReject ↓**	**FLAMES ↑**	**Do-Not-Answer ↑**
DeepSeek-R1-Distill-Llama-70B	96.8	21.8	62.0	31.6	69.5
SafeWork-R1-DeepSeek-70B	**98.0**$_{\uparrow 1.2}$	**0.5**$_{\downarrow 21.3}$	**0.2**$_{\downarrow 61.8}$	**72.2**$_{\uparrow 40.6}$	**99.3**$_{\uparrow 29.8}$
Capability Benchmarks					
Model	**GPQA Diamond ↑**	**Math-500 ↑**	**AIME2024 ↑**	**LiveCodeBench ↑**	**LiveBench ↑**
DeepSeek-R1-Distill-Llama-70B	**59.1**	**93.2**	67.1	41.9	40.0
SafeWork-R1-DeepSeek-70B	58.1$_{\downarrow 1.0}$	91.8$_{\downarrow 1.4}$	**74.2**$_{\uparrow 7.1}$	**50.5**$_{\uparrow 8.6}$	**48.0**$_{\uparrow 8.0}$