SHARP: Synthesizing High-quality Aligned Reasoning Problems for Large Reasoning Models Reinforcement Learning

Xiong Jun Wu[*1] **Zhenduo Zhang**[*1] **Zujie Wen**[1] **Zhiqiang Zhang**[1] **Wang Ren**[2]
Lei Shi[2] **Cai Chen**[2] **Deng Zhao**[2] **Qing Wang**[3] **Xudong Han**[3] **Chengfu Tang**[3]
Dingnan Jin[2] **Qing Cui**[2] **Jun Zhou**[1]
[1] AI Alignment at Ant Group [1,2] NextEvo, [3] Platform at Ant Group

Abstract

Training large reasoning models (LRMs) with reinforcement learning in STEM domains is hindered by the scarcity of high-quality, diverse, and verifiable problem sets. Existing synthesis methods, such as Chain-of-Thought prompting, often generate oversimplified or uncheckable data, limiting model advancement on complex tasks. To address these challenges, we introduce **SHARP**, a unified approach to **S**ynthesizing **H**igh-quality **A**ligned **R**easoning **P**roblems for LRMs reinforcement learning with verifiable rewards (RLVR). **SHARP** encompasses a strategic set of self-alignment principles—targeting graduate- and Olympiad-level difficulty, rigorous logical consistency, and unambiguous, verifiable answers—and a structured three-phase framework (Alignment, Instantiation, Inference) that ensures thematic diversity and fine-grained control over problem generation. We implement **SHARP** by leveraging a state-of-the-art LRM to infer and verify challenging STEM questions, then employ a reinforcement learning loop to refine the model's reasoning through verifiable reward signals. Experiments on benchmarks such as GPQA demonstrate that **SHARP**-augmented training substantially outperforms existing methods, markedly improving complex reasoning accuracy and pushing LRM performance closer to expert-level proficiency. Our contributions include the **SHARP** strategy, framework design, end-to-end implementation, and experimental evaluation of its effectiveness in elevating LRM reasoning capabilities.

1 Introduction

Large Reasoning Models (LRMs), such as OpenAI-O1, O3/O4 (OpenAI, a,b), Qwen3 (Qwen), and DeepSeek-R1 (DeepSeek-AI, 2025), have demonstrated remarkable capabilities in complex domains like mathematics and coding (Chen et al., 2025). However, mastering complex, multi-step reasoning, especially within STEM domains, remains a significant challenge (Hochlehnert et al., 2025; Rein et al., 2024; Lewkowycz et al., 2022). In these fields, models must not only understand the problem but also perform rigorous logical deductions to arrive at accurate answers. While techniques like Chain-of-Thought (CoT) prompting (Wei et al., 2022) encourage models to produce intermediate reasoning steps, the quality, complexity, and logical soundness of these generated paths can be inconsistent, often limited by the scale and quality of the underlying training data. Generating high-quality reasoning data for STEM is notoriously difficult. It requires domain expertise, careful problem construction to avoid ambiguity, and verifiable solutions (Lightman et al., 2023). Manually creating such datasets is expensive and slow, while existing automated methods may lack the necessary depth, diversity, or logical coherence required to train truly advanced reasoning models (DeepSeek-AI,

Preprint. Under review.

2025). This scarcity of suitable training data forms a critical bottleneck in advancing LLM reasoning capabilities towards expert-level or even superintelligence performance (Li et al., 2025).

To overcome these limitations and further enhance LRMs' performance on complex STEM reasoning tasks, particularly those requiring graduate- or Olympiad-level knowledge and reasoning skills (e.g., GPQA), we introduce a novel **SHARP**(**S**ynthesizing **H**igh-quality **A**ligned **R**easoning **P**roblems) approach. Specifically, the main components of **SHARP** approach include:
The SHARP Strategy: The strategy includes a set of self-alignment guiding principles covering problem difficulty (comparable in difficulty to graduate-level coursework or challenging Olympiads), reasoning consistency, answer format, authenticity, language, modality, structure, and output formatting, etc. This strategy focuses not only on alignment principles throughout the reasoning process, but also emphasizes answer verifiability and unambiguity.

The SHARP Framework: The framework comprise **Alignment, Instantiation, and Inference** phases. The **Instantiation** phase includes **Three-Tier Subject Categorization**, a hierarchical system (e.g., Subject $\rightarrow$ Category $\rightarrow$ Topic) enabling targeted generation of diverse samples across specific STEM sub-fields.

The SHARP Implementation: We first implement the **SHARP** framework leveraging an open-source state-of-the-art LRM (such as DeepSeek R1 DeepSeek-AI, 2025) itself to synthesize self-aligned generative challenging STEM problems systematically, their step-by-step problem-solving reasoning and reference answers, guided by the **SHARP** strategy. Then we evaluate these samples with general verifiers, such as Math-Verify (HuggingFace). to obtain the final ground truth. By utilizing these synthesized aligned high-quality and challenging samples, we train large reasoning models through reinforcement learning from zero (like DeepSeek R1 Zero (DeepSeek-AI, 2025)) and further enhance the model's reasoning capabilities in complex STEM problem-solving.

Extensive experiments demonstrate that our proposed **SHARP** strategy, particularly when coupled with the **SHARP** framework, through **SHARP Implementation**, can produce large-scale, high-quality samples capable of significantly boosting the complex reasoning performance of LLMs with reinforcement learning, pushing their reasoning capabilities closer to expert-level proficiency in STEM domains. Our main contributions are as follows:

- We propose a novel **SHARP** approach, comprising a set of carefully designed self-aligned core principles for synthesizing aligned generative complex and high-quality STEM reasoning samples.
- We detail the methodology in the **SHARP** framework, including **Alignment, Instantiation, and Inference** phases. The **Instantiation** phase includes a structured data fusion framework incorporating three-tier subject categorization for diverse and targeted sample generation.
- We implement the framework for LRMs with reinforcement learning for enhancing the complex reasoning capabilities of STEM problem-solving.
- Experiments demonstrate the effectiveness of the proposed approach in improving model performance on challenging STEM reasoning tasks and benchmarks through comprehensive evaluations.
- The **SHARP** offers a potential pathway to significantly enhance LRM performance on challenging STEM reasoning benchmarks like GPQA (Rein et al., 2024).

The remainder of this paper is organized as follows: Section 2 introduces background concepts. Section 3 details our proposed **SHARP** approach. Section 4 outlines the experimental setup. Section 5 presents experimental results and analysis. Section 6 discusses related work. Finally, Section 7 concludes the paper.

2 Background

LLMs often struggle with problems demanding true logical reasoning. Optimizing LLM reasoning to enable systematic, human-like logical thinking remains a key research direction. Several techniques are proposed to elicit reasoning from LLMs.

Chain-of-Thought (CoT): CoT prompting improves LLM performance on complex tasks by guiding them to generate intermediate reasoning steps (Wei et al., 2022; Li et al., 2024; Yeo et al., 2025). By mimicking human thought processes, CoT breaks down complex problems into smaller, manageable steps, aiding comprehension and solution derivation. Variants include Self-Consistency (Wang et al., 2023a), which samples multiple reasoning paths, and Tree-of-Thoughts (Yao et al., 2023) or Graph-of-Thoughts (Besta et al., 2024), which explore more complex reasoning structures. However, CoT has limitations: it can be highly dependent on precise prompt engineering. Crucially, the final generated answers cannot easily be verified or even are usually not accurate.

Self-Alignment in Large Reasoning Models (LRMs): Self-alignment utilizes an LLM's own capabilities to refine its behavior or training data (Wang et al., 2024b), aiming to reduce reliance on human annotation and improve data quality and diversity through model self-generation, evaluation, or correction (Dong et al., 2025). Samples include LLMs generating responses to unknown questions with explanations of unanswerability or using multi-round bootstrapping for self-improvement (Deng et al., 2024). Self-alignment offers a promising direction for training more powerful and reliable LLMs (Cao et al., 2024).

Reinforcement Learning for LLMs: The LRMs RL model OpenAI-O1, O3/O4 (OpenAI, a,b), Qwen3 (Qwen), and DeepSeek-R1 (DeepSeek-AI, 2025) involve self-play or self-critique mechanisms where the model learns from rewards generated based on its own outputs, akin to AlphaZero (Silver et al., 2017) but applied to text generation and reasoning.

In addition, several challenging benchmark datasets have been developed to evaluate LLM reasoning capabilities in STEM. GPQA (Graduate-Level Google-Proof Q&A Benchmark) (Rein et al., 2024) is designed by domain experts to be extremely difficult (PhDs achieve $\sim$65% accuracy). Its "Google-proof" nature makes it ideal for assessing deep understanding and reasoning, as answers are hard to find via web search. Performance on this benchmark serves as a crucial proxy for evaluating the effectiveness of our proposed **SHARP** approach.

3 SHARP: Synthesizing High-quality Aligned Reasoning Problems

Our proposed **SHARP** approach aims to systematically generate high-quality, complex STEM reasoning samples by guiding a state-of-the-art LRM (such as DeepSeek R1) instance-alignment reasoning inference through the **SHARP** framework 3.2 governed by the **SHARP** following strategy.

3.1 The SHARP Strategy

The starting point of the entire **SHARP** approach is to apply the **SHARP** strategy, and Fig.1 illustrates the **SHARP** strategy pipeline, including instance-level problem generation and alignment inference phases. This indicates that all subsequent steps, especially the **Instance-Alignment Reasoning Inference** in Fig.1 (described in **Instantiation Phase** 3.2), will strictly follow the self-alignment principles in the **SHARP** strategy.

Compared with conventional Direct QA and Chain-of-Thought (CoT) reasoning, the core objective of the **SHARP** strategy shown in Algo.1 is to ensure that generated samples possess high-quality and challenging samples, and precise reference answers. These synthesized aligned questions are not only of high difficulty and topic diversity, but also strictly follow the high consistency requirements of logic, ground truth, authenticity, language, structure, modality, and format. More importantly, the verified reference answers of these high-quality questions will strictly meet the Ground Truth consistency and complexity expansion requirements, that is, it will be an objectively verifiable single value (or a specified aggregation form) and follow the format specification.

Specifically, we formalize the **SHARP** self-alignment strategy as shown in Algorithm 1.

3.2 The SHARP Framework

Building upon the **SHARP** strategy, we introduce an enhanced **SHARP** data fusion framework specifically designed for synthesizing high-quality reasoning problems in STEM sub-disciplines. The core of this framework is the construction of the "**Seed Topics library**", which is built on a "Three-Tier Category" knowledge structure. This structure integrates the Magpie query generation approach (Xu et al., 2025b) with advanced semantic clustering and balanced sampling techniques, improving both the diversity and representativeness of the synthetic reasoning queries. Seed documents are meticulously curated from established benchmark question banks (we will not directly rephrase the

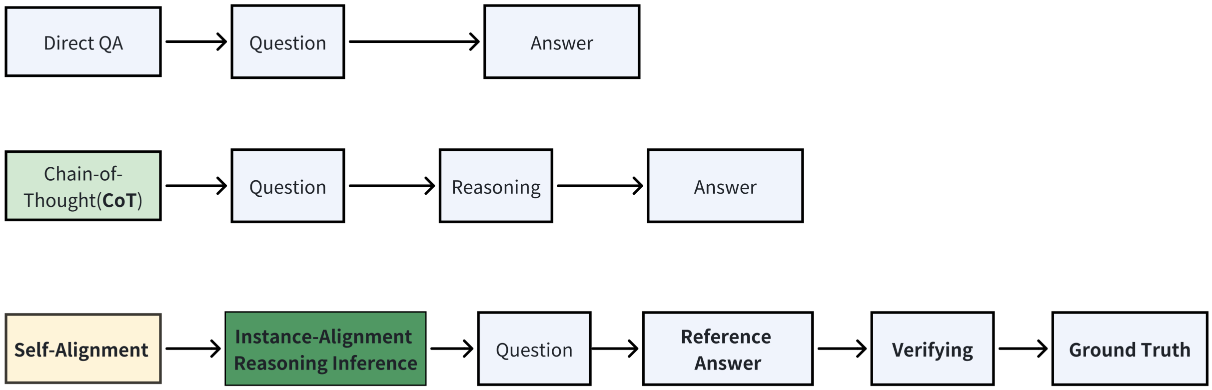

Figure 1: The **SHARP** Approach

Algorithm 1 SHARP Self-Alignment Problem Synthesis Strategy

Require: Seed topic set $S = \{s_1, s_2, \ldots, s_N\}$; alignment strategy constraints $\{x_v\}$; base LRM model; reasoning spec R_{spec}; verifier V

Ensure: Verified aligned question-answer pairs $Q = \{(q_1, a_1), \ldots, (q_m, a_m)\}$

Initialize $Q \leftarrow \emptyset$
for each seed topic $s_i \in S$ **do**
    ▷ Alignment Phase
    Configure alignment constraints: Alignment constraints $\{x_v\}$ (See Appendix A for details.)
    Construct reasoning blueprint (e.g., step-by-step, propose-verify)
    ▷ Instantiation Phase
    Generate prompt $p_i \leftarrow \textsc{InstantiatePrompt}(s_i, \{x_v\}, R_{\text{spec}})$
    Select reasoning structure using Three-Tier Category hierarchy.
    ▷ Inference Phase
    Query model with p_i to generate (q_i, r_i, a_i):
    $q_i \leftarrow$ question text, $r_i \leftarrow$ reasoning trace, $a_i \leftarrow$ final answer
    Format output using SHARP conventions:
    `<question start>` q_i `<question end>`
    reasoning: r_i, final answer: \boxed{{$answer}}
    ▷ Verifying Phase
    if $V(r_i, a_i)$ passes all alignment checks **then**
        $Q \leftarrow Q \cup \{(q_i, a_i)\}$
    end if
end for
return Q

query based on the validation set, but only analyze the topic keypoints covered by these benchmarks) and high-quality handcrafted corpora (STEM textbooks, papers, and data recalled through Common Crawl etc.), while cutting-edge LLMs, such as DeepSeek R1 and Qwen3, are employed to facilitate comprehensive topic extraction and "Three-Tier Category" generation, ensuring a broad coverage of critical reasoning domains. The clustering process, utilizing K-means algorithms (MacQueen, 1967) on BGE-m3 embeddings (Chen et al., 2024), in tandem with balanced sampling, addresses potential biases, ensuring uniform representation across a spectrum of reasoning topics.

Moreover, the integration of persona-based methodologies (Ge et al., 2025) and keypoint enhancements introduces a diverse array of reasoning contexts and enables the modulation of query difficulty levels, facilitating the generation of training data that reflects both problem complexity and cognitive challenges. This methodological approach ensures scalable reasoning problem synthesis that aligns closely with the depth and complexity required in STEM-related tasks. The **SHARP** framework, by leveraging sophisticated reasoning capabilities of LLMs like DeepSeek R1, synthesizes logically

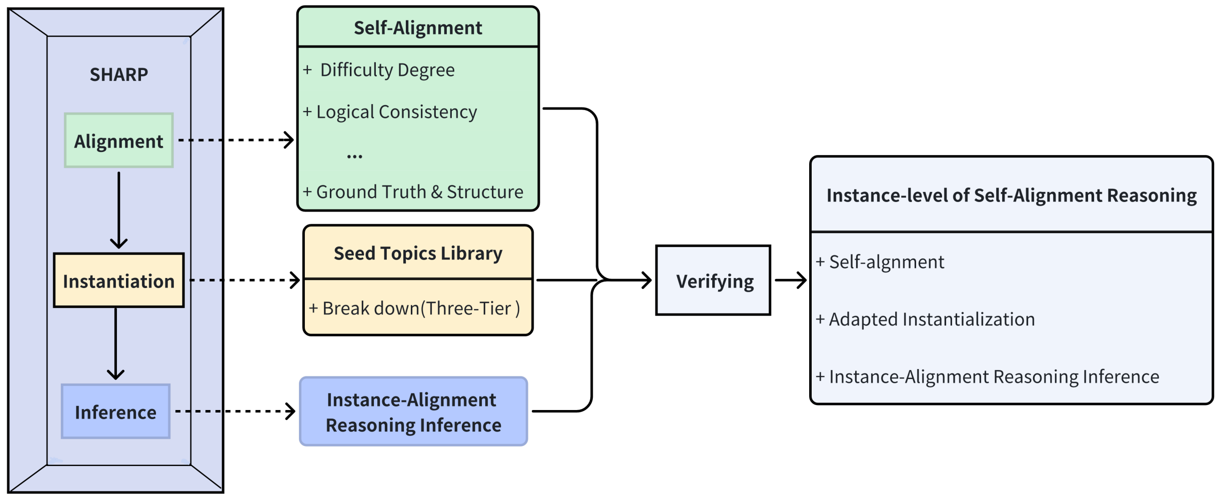

Figure 2: The **SHARP** Framework

coherent, complex reasoning problems that are carefully aligned with the nuanced demands of STEM disciplines. The primary objective of this framework is to generate high-quality, diverse training samples that drive the optimization of reinforcement learning (RL) models, especially in the context of high-difficulty STEM benchmarks.

The **SHARP** framework is underpinned by its core modules, prominently featuring the **Self-Alignment** block (depicted in Fig. 2). This block guarantees the adherence to stringent quality standards for the generated content, encompassing aspects such as question difficulty, reasoning consistency, and answer verifiability. It systematically encompasses three fundamental phases: **Alignment**, **Instantiation**, and **Implement**, thereby constructing a comprehensive reasoning alignment pathway from the initial **SHARP** strategy formulation to the generation of specific instances. This structured approach ensures that all internal operations are cohesively aligned with the **SHARP** strategy, ultimately facilitating the synthesis of high-quality, aligned reasoning problems tailored for reinforcement learning in large-scale reasoning models. The three phases of the **SHARP** framework are detailed as follows.

Alignment Phase: This phase initiates the **SHARP** approach, and serves as its implicit input of the overall goal of applying the **SHARP** Algo.1 strategy, corresponding to the "**Self-Alignment**" detail box on the right. The specific requirements set in this phase are the key manifestation of the **SHARP** Algo.1 strategy in sample generation, and all subsequent steps will align strictly to it. It inherits and strengthens the core advantages of **Self-Alignment**, especially the structural requirements for the reasoning process, ensuring the reasoning consistency and reliability of the generated samples, and helping the training model to form more standardized and reliable reasoning capabilities.

We begin this phase by operationalizing the **SHARP** principles into executable constraints, passing specific requirements (e.g., difficulty level, reasoning style, verification method, etc.) to the next phase. Then the **SHARP** plans a systematic reasoning framework or blueprint that meets logical consistency, ensuring that each step of deduction is supported by STEM theory or logic, eliminating jumps and intuitive guesses, and maintaining format requirements (such as the Math-Verify (HuggingFace)). It enforces that reasoning must be planned, orderly, and verifiable, rather than arbitrary heuristic deduction. For example, we set the "Difficulty Degree" to graduate- or Olympiad-level, mandate a "Step-by-Step" reasoning process, and employ a "Propose-verify" mechanism where the model internally proposes and verifies each reasoning step for validity and truthfulness. These standards are consistent with those in the **Verifying** stage.

Instantiation Phase: Building on the **SHARP** Algo.1 strategy of the **Alignment Phase** and a "Three-Tier Category" knowledge framework, a clear "reasoning structure" definition stage for the

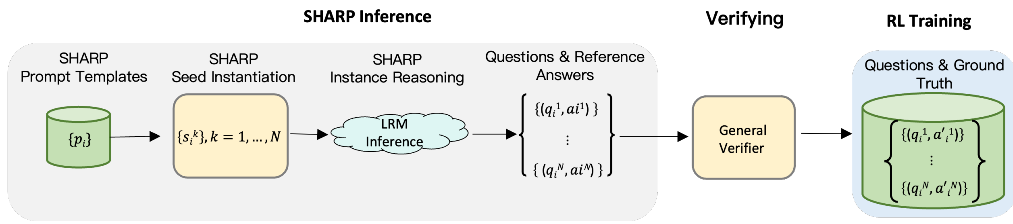

Figure 3: The **SHARP** Implementation for Large Reasoning Models Reinforcement Learning

instantiations is introduced, distinct from the relatively free "Reasoning" step in traditional CoT. This instantiation phase integrates a "Three-Tier Category" knowledge framework to instantiate the strategy to different subjects' characteristics and structures. The "Three-Tier Category" knowledge framework manages and organizes STEM knowledge hierarchically (e.g., Chemistry → Organic Chemistry, Spectroscopy → Elimination reactions, IR spectroscopy, Carbonyl compounds, Alcohols, Characteristic IR absorption frequencies). A comprehensive and easily expandable "**Seed Topics library**" (orange detail box) is organized in this hierarchy. This ensures the combination of broad thematic coverage and professional depth, enabling targeted generation of complex samples in specific STEM sub-fields. The structured topic information informs the **Verifying** stage for confirming instance topic attribution and generation relevance.

Inference Phase: Under the guidance of the "reasoning structure" defined in the **Instantiation Phase**, an instantiated aligned reasoning inference process generating a specific STEM sample that meets the **SHARP** strategy is performed. This involves leveraging the capabilities of the state-of-the-art LRMs model (such as DeepSeek R1) to generate aligned reasoning instances. Then these aligned reasoning instances are submitted to the next **Verifying** stage.

The **Verifying** stage is mainly responsible for quality control, strictly verifying whether the adapted instance fully complies with the **Self-Alignment** detail box, and whether it is consistent with the three-level category theme of **All Seed Topics** it claims, to ensure that the final output sample meets the preset high standards detailed in **SHARP** strategy. The final output of the entire process - high-quality samples are generated that can elicit complex reasoning for LRMs RL training.

Building upon **SHARP** Algo.1, **SHARP** framework 2 introduces innovative "instance-level" reasoning, where each sample constitutes a complete and self-consistent reasoning structure. This is achieved through a refined three-level subject classification adaptation mechanism, a robust inference and verification process. These meticulously refined samples are invaluable for the model, enabling it to learn fine-grained knowledge and complex reasoning patterns. By systematically generating STEM samples with ultra-high complexity at the sample level, this comprehensive approach provides unique value and significant potential for enhancing the complex reasoning capabilities of LRMs, particularly in improving top STEM reasoning benchmarks, such as GPQA.

3.3 The SHARP Implementation

Based on the **SHARP** strategy and framework, we implement the **SHARP** approach with a state-of-the-art LRM model (like DeepSeek R1) and generate complex reasoning samples at the sample level, shown in Fig. 2. Then these synthesized high-quality self-aligned samples are used to enhance the complex reasoning ability of the RL Zero model training (inspired by RL Zero (OpenAI, a; DeepSeek-AI, 2025; OpenAI, b)). The system mainly consists of three core stages: **SHARP Inference**, **Verifying** and **RL Zero Training** and are detailed as follows.

SHARP Inference:

1. **SHARP Prompt Templates** ($\{p_i\}$): As the starting point of the process, initial aligned question prompt templates that meet the **SHARP** strategy (such as high difficulty, plain text, single question, factual accuracy, etc.) are constructed.
2. **SHARP Seed Instantiation** ($\{s_i^k\}, k = 1, ..., N$, here N is the total number of topics): Extract an aligned prompt x_i from the **SHARP** prompt templates, and map the input general prompt x_i with the specific STEM knowledge points or topic "seeds" $\{s_i^k\}$ based on the

"**Seed Topics Library**" according to "Three-Tier Category" knowledge framework in the **SHARP** framework. This provides context for the subsequent generation of domain-specific and depth-specific reasoning, and the output is the prompt x_i and its associated instance with adapted specific topic $\{s_i^k\}$ from **Seed Topics Library** in the "**Instantiation Phase**".

3. **Questions and Reference Answers** ($\{(q_i^k, a_i^k), k = 1, ..., m\}$): Use the current best version of the reasoning model (e.g., DeepSeek R1) to generate candidate reasoning responses for the prompts and topic contexts, including high-quality questions $\{q_i^k\}$, its **SHARP** reasoning processes, and corresponding candidate reference answers ($\{a_i^k\}$)).

Verifying: Verification is performed according to the **SHARP** strategy and a scoring rubric combining model self-check confidence and a rule-based reward model for the generated question and answer pairs ($\{(q_i^k, a_i^k)\}$). The purpose is to further screen out data with poor quality, logical errors, unreliable reward signals, or data that does not meet the final requirements, and ensure that only the highest quality and most reliable samples enter the LRM RL training stage.

RL Training: Using the high-quality and verified samples created by the **SHARP Inference** and **Verifying**, RL Zero training is carried out based on an RL algorithm (e.g., PPO (Schulman et al., 2017), or GRPO (Shao et al., 2024)).

The unique value of **SHARP** inference system lies in its focus on quality and complexity, adaptive sample generation of higher quality and more complex reasoning samples, and thereby can train a more powerful LRM model through RL Zero and push the upper limit of the model's reasoning capabilities in STEM fields dimensions.

4 Experimental Setup

Our experimental setup was meticulously designed to rigorously evaluate the **SHARP** approach. The Training Data consisted of two primary sets: a baseline dataset generated using standard CoT prompting on existing STEM samples and a substantial dataset of 190,000 samples generated via the **SHARP** methodology. Our comparison models included distinct sets for distillation training and RL Zero training. For distill training, the **Qwen2.5-7B-Instruct-Distill** model served as a baseline, representing capable LRMs without specific STEM reasoning dataset training. This was compared against state-of-the-art **DeepSeek-R1-Distill-Qwen-7B** (DeepSeek-AI, 2025) and **SHARP-Qwen2.5-7B-Instruct-Distill** (Qwen et al., 2025), where the latter was the baseline model further distilled on **SHARP**-generated and verified samples. In the RL Zero Training comparison, **Open-Reasoner-Zero-7B** (Hu et al., 2025) was the baseline, evaluated against **SHARP-Open-Reasoner-Zero-7B**, which was trained using **SHARP**-generated problems through an RL Zero process. The training details for distillation involved standard procedures on the respective datasets. For **SHARP**-RL Zero training, we employed the GRPO algorithm, with a rule-based reward function, alongside specified hyperparameters and computational resources detailed in the appendix. Finally, evaluation metrics centered on model performance on the challenging GPQA STEM reasoning benchmark, using accuracy metrics like pass@k to compare models trained with and without **SHARP**-generated samples, thereby demonstrating the efficacy of our approach.

5 Experiments: Results and Analysis

Building upon the described experimental setup, our evaluations demonstrate the significant advantages of the **SHARP** methodology in enhancing large reasoning models. The experiments were conducted in two primary modes: high-quality complex reasoning knowledge distillation with supervised fine-tuning, and the utilization of challenging **SHARP**-generated samples to elicit complex reasoning capabilities in LRMs.

The results, presented in Table 1 and 2, are compelling. Approximately 190,000 STEM samples were constructed using the **SHARP** approach. In the distillation experiments (Table 1), the **SHARP-Qwen2.5-7B-Instruct-Distill** model, trained on **SHARP** data, achieved a GPQA Diamond score of 54.7. This represents an 8.3 percentage point improvement over the **Qwen2.5-7B-Instruct-Distill** baseline (46.4) and a 4.8 percentage point increase over the **DeepSeek-R1-Distill-Qwen-7B model** (49.9). This notable outperformance, even without RL refinement, underscores the superior quality of data generated by the structured **SHARP** approach. The **SHARP**-trained model also showed consistent improvements across GPQA Physics (71.1 vs. 60.6 baseline), Chemistry (38.8 vs. 31.3 baseline), and Biology (57.9 vs. 55.9 baseline).

Models	GPQA Physics	GPQA Chemistry	GPQA Biology	GPQA Diamond
Qwen2.5-7B-Instruct-Distill (Baseline)	60.6	31.3	55.9	46.4
DeepSeek-R1-Distill-Qwen-7B	70.1	31.9	43.4	49.9
SHARP-Qwen2.5-7B-Instruct-Distill	71.1	38.8	57.9	54.7

Table 1: Performance on GPQA benchmark (Diamond subset: most difficult tier), comparing distilled models trained with and without **SHARP**-synthesized data.

Models	GPQA Physics	GPQA Chemistry	GPQA Biology	GPQA Diamond
Open-Reasoner-Zero-7B (Baseline)	41.4	27.4	48.7	35.5
SHARP-Open-Reasoner-Zero-7B	44.6	26.3	54.9	37.0

Table 2: Performance on GPQA benchmark (Diamond subset: most difficult tier), comparing RL Zero models trained with and without **SHARP**-synthesized data.

In the RL-Zero reasoning training experiments (Table 2), the **SHARP-Open-Reasoner-Zero-7B** model, leveraging SHARP-generated STEM problems, achieved a GPQA Diamond score of 37.0, marking a 1.5 percentage point improvement over the **Open-Reasoner-Zero-7B** baseline (35.5). This outcome offers initial validation for the efficacy of **SHARP**-synthesized data in supporting RL-Zero reasoning training. Notably, performance enhancements were recorded in GPQA Physics (44.6 vs. 41.4 baseline) and GPQA Biology (54.9 vs. 48.7 baseline). Conversely, GPQA Chemistry exhibited a marginal decrease (26.3 vs. 27.4 baseline). We attribute this to the inherently high dependence of chemistry problems on deep, structured domain knowledge and nuanced symbolic reasoning, which may not be as effectively acquired through unsupervised RL Zero methods without pre-distilled domain-specific priors, as detailed in Appendix C.2. Detailed analyses of these discrepancies, including sample difficulty metrics (e.g., response length and reward signal distribution), are provided in Appendix B.2.

To further substantiate these findings and provide a more granular understanding, Appendix B includes comparative evaluations of **SHARP**-based distillation and ablation studies across STEM fields and mathematical data. Specifically, we present controlled experiments analyzing the performance impact of different **SHARP**-generated sub-corpora—physics, chemistry, biology—on both distillation and RL Zero models. Representative subject-level ablation examples further demonstrate how **SHARP**'s three-tier taxonomy enables precise control over difficulty and topic diversity. Additionally, Appendix C systematically evaluates the 190,000 **SHARP**-generated samples, showcasing balanced distributions across 600+ granular STEM subcategories and pass rate analyses that correlate with human expert assessments. Together, these results confirm that **SHARP**'s aligned and structured synthesis framework successfully generates high-difficulty, verifiable problems—spanning quantum mechanics to organic reaction mechanisms—that directly enhance LRMs' capacity for expert-level scientific reasoning.

Collectively, these findings highlight the **SHARP** approach's effectiveness in generating high-quality, complex training samples. The consistent performance gains observed across different models and evaluation subjects, particularly on the demanding GPQA-Diamond set, demonstrate that **SHARP** significantly enhances the capability of LRMs to tackle complex STEM reasoning tasks, pushing their performance closer to expert-level proficiency. The structured generation process, guided by **SHARP**'s self-alignment strategy, yields problems that is not only diverse and challenging but also logically rigorous and verifiable, directly contributing to the observed improvements in LLMs' reasoning abilities.

6 Related Work

LLM Reasoning Enhancement: As discussed in Section 2, numerous efforts focus on improving LLM reasoning via prompting (CoT, ToT, GoT) (Wei et al., 2022; Yao et al., 2023; Besta et al., 2024) or specialized fine-tuning (Trung et al., 2024; Lobo et al., 2025). However, scaling up verifiable signals for long CoT remains challenging due to the limited availability of high-quality, verifiable samples (Yeo et al., 2025). In contrast, **SHARP** generates verifiable, high-quality samples without relying on prompt engineering, enabled by structured self-alignment.

Synthetic Data for LLM Reasoning: Synthesizing data plays a crucial role in training large language models (LLMs) to enhance their reasoning abilities. Approaches like Self-Instruct and Alpaca (Wang et al., 2023b) have pioneered the use of generated instructional data to align LLM behaviors with desired outcomes. (Shao et al., 2023) introduced a method where a limited set of handcrafted samples prompts the model to autonomously create additional data, selectively incorporating high-quality demonstrations to bolster reasoning performance. Nemotron-CrossThink (Akter et al., 2025) leverages cross-thought reasoning to enable self-improvement within mathematical domains, while Qwen2.5-Math and Qwen2.5-Coder (Yang et al., 2024; Hui et al., 2024) focus on generating domain-specific data for mathematical problem-solving and coding tasks, respectively. Phi-4-Reasoning (Xu et al., 2025a; Abdin et al., 2025) demonstrates the effectiveness of compact architectures in handling complex reasoning tasks. (Goldie et al., 2025) introduced SWiRL for multi-step reward shaping; such signal shaping is partially mirrored in our SHARP RL reward design. Together, these studies highlight the significance of sophisticated data synthesis strategies in improving LLM reasoning capabilities. Unlike these approaches, **SHARP** specifically targets the synthesis of challenging STEM problems by enforcing a unique combination of explicit self-alignment principles for reasoning consistency, thematic diversity, and strict answer verifiability, aiming to overcome the limitations in generating consistently complex and reliable reasoning samples.

Self-Alignment. Self-alignment refers to training paradigms that utilize a model's own capabilities to assess, revise, or supervise its outputs—reducing the reliance on external human annotation. Existing approaches can be broadly categorized into three paradigms. First, *preference-based self-alignment* leverages internal comparisons or preference signals, often sparse or noisy, to guide alignment. For example, the Hummer framework (Wu et al., 2024) investigates competitive learning under weak preference supervision, emphasizing the importance of carefully designed reward structures for stable alignment. Second, *consistency-based self-alignment* uses internal logical coherence as a proxy for correctness. The SelfFeedback framework (Liang et al., 2024) enhances model reasoning by identifying and reinforcing internally consistent outputs, offering a form of self-improvement without external labels. Third, and most relevant to our work, are *verifiability-based self-alignment* approaches, which aim to provide strong and objective training signals by enforcing explicit correctness. Our proposed **SHARP** framework builds on the motivation of reducing human oversight but introduces a structured, multi-phase self-alignment strategy tailored to complex STEM domains. Unlike methods that rely on implicit or heuristic feedback, **SHARP** synthesizes *explicitly verifiable*, *logically rigorous*, and *thematically diverse* reasoning samples that are suitable for use as direct supervision targets. By encoding these properties through a principled three-phase pipeline—**Alignment, Instantiation, and Inference**—**SHARP** enables the systematic synthesis of reasoning problems that are not only logically rigorous and thematically diverse, but also explicitly verifiable. This structure supports high-fidelity, reward-aligned supervision signals, making it particularly effective for reinforcement learning from verifiable rewards (RLVR) in domains where correctness must be grounded in domain knowledge and formal reasoning. In this way, **SHARP** bridges the methodological gap between implicit alignment strategies and the stringent verification demands of complex STEM problem-solving.

In summary, **SHARP** offers a principled and extensible framework that advances the current state of self-alignment and synthetic reasoning for large language models. By tightly integrating explicit alignment objectives with a structured, verifiability-driven sample generation process, **SHARP** produces high-difficulty and semantically controlled reasoning data that overcomes the limitations of heuristic prompting, weak preference modeling, or unverified self-consistency. This makes **SHARP** especially well-suited for reinforcement learning settings in scientific domains, where both logical fidelity and verifiable correctness are essential. Ultimately, **SHARP** contributes a scalable, high-precision foundation for training advanced reasoning models to operate effectively in complex and rigorous STEM contexts.

7 Conclusion, Limitations and Future Work

We presented a novel **SHARP** approach to address the critical need for high-quality, complex, and verifiable training problems for enhancing the reasoning capabilities of LLMs, particularly in STEM domains. By employing **SHARP** inference and **Verifying** process, our approach systematically guides LRMs to generate challenging problems and logically sound, verifiable solutions efficiently and at scale, addressing the limitations of traditional CoT methods in producing difficult, diverse, and logically rigorous STEM reasoning samples. We presented the **SHARP** inference integrating with **Verifying** process, enabling iterative RL foundation model training and performance enhancement on complex reasoning tasks. Experimental results demonstrate significant performance gains on challenging STEM benchmark GPQA compared to baselines trained on CoT data and public STEM datasets, as well as substantial improvement over the state-of-the-art baseline model. For instance, SHARP-augmented distillation training resulted in an 8.3 percentage point improvement on the GPQA Diamond benchmark over the baseline. This validates the effectiveness of our proposed approach in enhancing the ability of large reasoning models to tackle complex STEM problems.

Future work could explore applying this approach to other domains and more complex reasoning tasks, and further optimizing the **SHARP** approach on various larger-scale RL reasoning foundation models. Besides, designing a reward function that weights principles from the **SHARP** strategy will be carried out. And distinctions among different subjects, such as chemistry and biology, have different subject attributes from physics and mathematics, which may involve the further improvement of logic, knowledge graph, and symbolic reasoning capabilities.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, Vibhav Vineet, Yue Wu, Safoora Yousefi, and Guoqing Zheng. Phi-4-reasoning technical report, 2025. URL `https://arxiv.org/abs/2504.21318`.

Syeda Nahida Akter, Shrimai Prabhumoye, Matvei Novikov, Seungju Han, Ying Lin, Evelina Bakhturina, Eric Nyberg, Yejin Choi, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Nemotron-crossthink: Scaling self-learning beyond math reasoning, 2025. URL `https://arxiv.org/abs/2504.13941`.

ArtofProblemSolving. Art of problem solving community. `https://artofproblemsolving.com/community/`.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690, March 2024. ISSN 2159-5399. doi: 10.1609/aaai.v38i16.29720. URL `http://dx.doi.org/10.1609/aaai.v38i16.29720`.

Book Industry Study Group. BISAC Subject Headings, 2025. URL `https://www.bisg.org/complete-bisac-subject-headings-list`.

Boxi Cao, Keming Lu, Xinyu Lu, Jiawei Chen, Mengjie Ren, Hao Xiang, Peilin Liu, Yaojie Lu, Ben He, Xianpei Han, Le Sun, Hongyu Lin, and Bowen Yu. Towards scalable automated alignment of llms: A survey, 2024. URL `https://arxiv.org/abs/2406.01252`.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.137. URL `https://aclanthology.org/2024.findings-acl.137/`.

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.

DeepScaleR. Deepscaler-preview-dataset, 2025. URL `https://huggingface.co/datasets/math-dataset/DeepScaleR-Preview-Dataset`.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Yang Deng, Yong Zhao, Moxin Li, See-Kiong Ng, and Tat-Seng Chua. Don't just say "I don't know"! self-aligning large language models for responding to unknown questions with explanations. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13652–13673, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.757. URL https://aclanthology.org/2024.emnlp-main.757/.

Guanting Dong, Keming Lu, Chengpeng Li, Tingyu Xia, Bowen Yu, Chang Zhou, and Jingren Zhou. Self-play with execution feedback: Improving instruction-following capabilities of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=cRR0oDFEBC.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas, 2025. URL https://arxiv.org/abs/2406.20094.

Anna Goldie, Azalia Mirhoseini, Hao Zhou, Irene Cai, and Christopher D. Manning. Synthetic data generation & multi-step rl for reasoning & tool use, 2025. URL https://arxiv.org/abs/2504.04736.

Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu, and Matthias Bethge. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility, 2025. URL https://arxiv.org/abs/2504.07086.

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, 2024. URL https://arxiv.org/abs/2405.11143.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model, 2025. URL https://arxiv.org/abs/2503.24290.

HuggingFace. Math-verify. https://github.com/huggingface/Math-Verify.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, Yunlong Feng, Xingzhang Ren, Xuancheng Ren, Jingren Zhou, and Junyang Lin. Qwen2.5-coder technical report, 2024. URL https://arxiv.org/abs/2409.12186.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of thought empowers transformers to solve inherently serial problems. *ArXiv*, abs/2402.12875, 2024.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhijiang Guo, Le Song, and Cheng-Lin Liu. From system 1 to system 2: A survey of reasoning large language models, 2025. URL https://arxiv.org/abs/2502.17419.

Xun Liang, Shichao Song, Zifan Zheng, Hanyu Wang, Qingchen Yu, Xunkai Li, Rong-Hua Li, Feiyu Xiong, and Zhiyu Li. Internal consistency and self-feedback in large language models: A survey. *CoRR*, abs/2407.14507, 2024. URL https://doi.org/10.48550/arXiv.2407.14507.

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

Elita Lobo, Chirag Agarwal, and Himabindu Lakkaraju. On the impact of fine-tuning on chain-of-thought reasoning. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11679–11698, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. URL https://aclanthology.org/2025.naacl-long.584/.

James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, volume 5, pp. 281–298. University of California press, 1967.

Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I Jordan, et al. Ray: A distributed framework for emerging {AI} applications. In *13th USENIX symposium on operating systems design and implementation (OSDI 18)*, pp. 561–577, 2018.

OpenAI. Introducing openai o1, a. URL https://openai.com/o1/.

OpenAI. Introducing openai o3 and o4-mini, b. URL https://openai.com/index/introducing-o3-and-o4-mini/.

OpenCompass. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023.

Qwen. Qwen3: Think deeper, act faster. URL https://qwenlm.github.io/blog/qwen3/.

Qwen. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL https://qwenlm.github.io/blog/qwq-32b/.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

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.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. Synthetic prompting: generating chain-of-thought demonstrations for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. Mastering chess and shogi by self-play with a general reinforcement learning algorithm, 2017. URL https://arxiv.org/abs/1712.01815.

THUDM. T1, 2025. URL https://huggingface.co/datasets/THUDM/T1.

Luong Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. ReFT: Reasoning with reinforced fine-tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7601–7614, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.410. URL `https://aclanthology.org/2024.acl-long.410/`.

Guanhua Wang, Heyang Qin, Sam Ade Jacobs, Xiaoxia Wu, Connor Holmes, Zhewei Yao, Samyam Rajbhandari, Olatunji Ruwase, Feng Yan, Lei Yang, et al. Zero++: Extremely efficient collective communication for large model training. In *The Twelfth International Conference on Learning Representations*, 2024a.

Haoyu Wang, Guozheng Ma, Ziqiao Meng, Zeyu Qin, Li Shen, Zhong Zhang, Bingzhe Wu, Liu Liu, Yatao Bian, Tingyang Xu, Xueqian Wang, and Peilin Zhao. Step-on-feet tuning: Scaling self-alignment of LLMs via bootstrapping. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*, 2024b. URL `https://openreview.net/forum?id=lAXNiTcMar`.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023a. URL `https://openreview.net/forum?id=1PL1NIMMrw`.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.754. URL `https://aclanthology.org/2023.acl-long.754/`.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

Yusen Wu, Li Jiang, Junwu Xiong, Jingqing Ruan, Yichuan Ding, Qingpei Guo, Zujie Wen, Jun Zhou, and Xiaotie Deng. Hummer: Towards limited competitive preference dataset. In *Proceedings of the Conference on Learning for Language Modeling (COLM)*, 2024. URL `https://openreview.net/forum?id=aKwQPRjdGa`. Under review.

Haoran Xu, Baolin Peng, Hany Awadalla, Dongdong Chen, Yen-Chun Chen, Mei Gao, Young Jin Kim, Yunsheng Li, Liliang Ren, Yelong Shen, Shuohang Wang, Weijian Xu, Jianfeng Gao, and Weizhu Chen. Phi-4-mini-reasoning: Exploring the limits of small reasoning language models in math, 2025a. URL `https://arxiv.org/abs/2504.21233`.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned LLMs with nothing. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL `https://openreview.net/forum?id=Pnk7vMbznK`.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement, 2024. URL `https://arxiv.org/abs/2409.12122`.

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

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms, 2025. URL https://arxiv.org/abs/2502.03373.

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. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 62557–62583. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/724be4472168f31ba1c9ac630f15dec8-Paper-Conference.pdf.

A SHARP Self-Alignment Strategy Constraints

SHARP Self-Alignment Strategy Constraints Details

Problem Difficulty & Thematic Diversity Alignment: Generate highly complex problems (graduate- or Olympiad-level) covering a wide range of STEM topics, covering expert-level AI themes. Difficulty is benchmarked against top exams and datasets (GPQA, etc.). Thematic coverage uses role-playing prompts template and a three-tier subject-category-topic framework.

Logical Consistency Alignment: Problem-solving must rely solely on rigorous reasoning or systematic derivation, avoiding pattern matching, heuristics, shortcuts, or fabrication. All intermediate steps require justification, preventing logical gaps or errors due to intuition.

Ground Truth & Structure Alignment: Answers must be single, verifiable numerical values (plain numbers, units, ratios, STEM formulas/equations). Avoid hard-to-verify formats (set operations, free text). For multi-solution problems, mandate a specific aggregation (e.g., sum or sum of squares, etc.) for a unique, objectively verifiable answer. Expand beyond single QA to include multi-solution problems (requiring summary values) (e.g., "calculate total moles of all possible products").

Problem Authenticity Alignment: Problems should be novel, based on authoritative knowledge, but not directly copied. They must be unambiguous, unbiased, accurate, and internally consistent, avoiding nonsensical or hallucinated scenarios.

Language Consistency Alignment: The entire generation process (problem statement, reasoning method, solution presentation) must use a single language (e.g., English or Chinese) to prevent multilingual confusion leading to reasoning errors or bad verification cases.

Problem Structure Consistency Alignment: Problems must contain only a single primary question, avoiding sub-questions, derivatives, or branching logic that leads to unverifiable cases.

Modality Consistency Alignment: Problems must be strictly text-based, describing any necessary complex structures (e.g., chemical molecules, genetic diagrams) textually.

Formatting Alignment: Use specific delimiters (e.g., `<question_start>`, `<question_end>`) for the problem statement and a standardized format (e.g., `\boxed{{$answer}}`) for the final answer.

The complete template including the **SHARP** self-alignment strategy constraints for constructing the challenge problem is shown in the following Table 8 and can also be found via this link.

B Performance Analysis of Distilling and RL Zero Model Reasoning Training with SHARP Samples

B.1 Distilling Training Model Performance Analysis

Fig.4 compares models trained on **SHARP-augmented Qwen2.5-7B-Instruct-Distill (Baseline)** and the strong benchmark **DeepSeek-R1-Distill-Qwen-7B** with samples generated by the **SHARP** approach across three STEM subjects: physics, chemistry, and biology (Our distilled models, including the RL Zero variants presented in this work, are evaluated on the GPQA benchmark using the widely adopted OpenCompass framework (OpenCompass, 2023), ensuring strict consistency with the official reported scores.). In addition, the physics, chemistry, and biology subjects all had positive improvements, and the chemistry and biology subjects compared with the DeepSeek chemistry subject improved significantly, indicating the effectiveness of our designed **SHARP** self-alignment strategy and reasoning training model, reflecting the improvement of the model in general knowledge and reasoning ability.

Also, model trained with **SHARP** problems only are significantly better than mathematical only distillation problems in improving the ability of physics, chemistry, and biology, as shown in 5 and

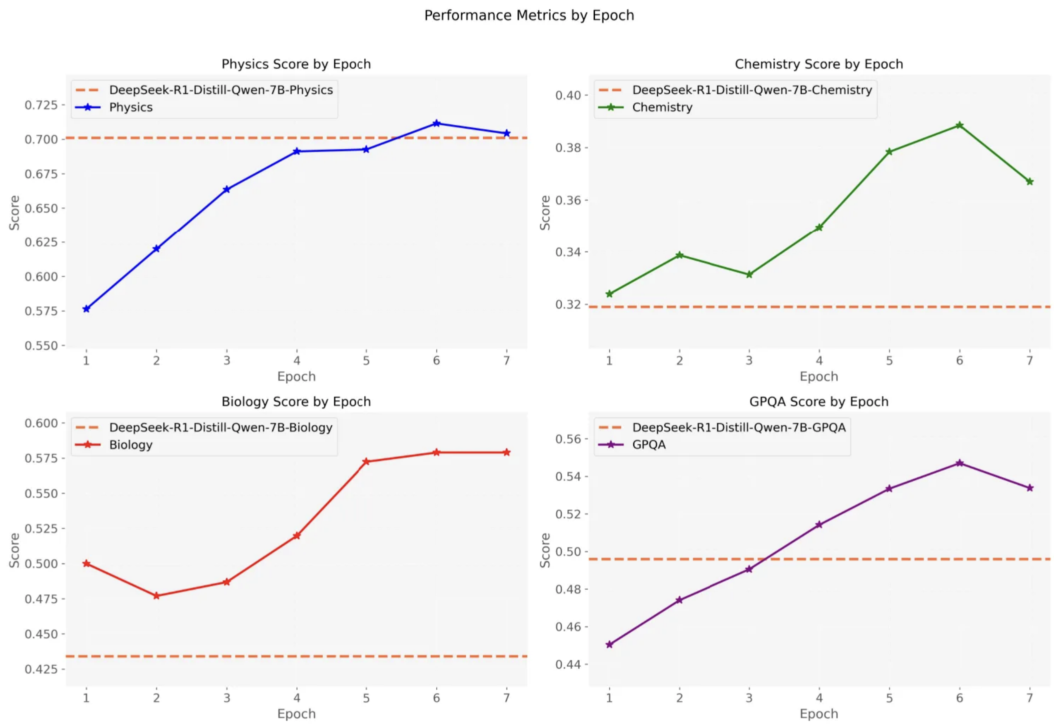

Figure 4: GPQA score improvement of single STEM disciplines (physics, chemistry, and biology) of **SHARP-Qwen2.5-7B-Instruct-Distill** relative to benchmark model **DeepSeek-R1-Distill-Qwen-7B** in overall ablation of STEM data generated by the **SHARP** approach and fused with some open-source mathematical data. (The x-axis represents the different epochs run during the training of the distill models, and the y-axis represents the GPQA score evaluation results corresponding to the checkpoints of the models generated at different epochs).

6, and thus significantly better than mathematical only distillation problems in the overall GPQA benchmark (Here, the mathematics data here accounts for 27.3%, mainly from (DeepScaleR, 2025; ArtofProblemSolving; THUDM, 2025), mathematics competition problems from all over the world, well-known universities, etc.). As seen from the Fig.6, the GPQA score of the distillation model is not as significant in improving the chemistry index in the pure mathematics data set as in physics and biology. This also shows, to some extent, that the attributes of chemistry and mathematical reasoning are relatively different.

We conduct these distillation model supervised finetuning across 10 epochs for all datasets and a learning rate of 5e-6. We employ a cosine learning rate scheduler, ensuring that the final learning rate reaches 1% of the peak value. We train these models with about 190,000 samples on 32 NVIDIA H800 GPUs for 10 hours. These core parameters for training are set as in Table 3:

Parameter Name	Value
max_length	16384
learning_rate	5e-6
lr_scheduler_type	cosine
warmup_ratio	0.01

Table 3: Distill Model Core Parameters.

Specific challenging problem samples generated by the **SHARP** approach used to train distill models can be found via this link.

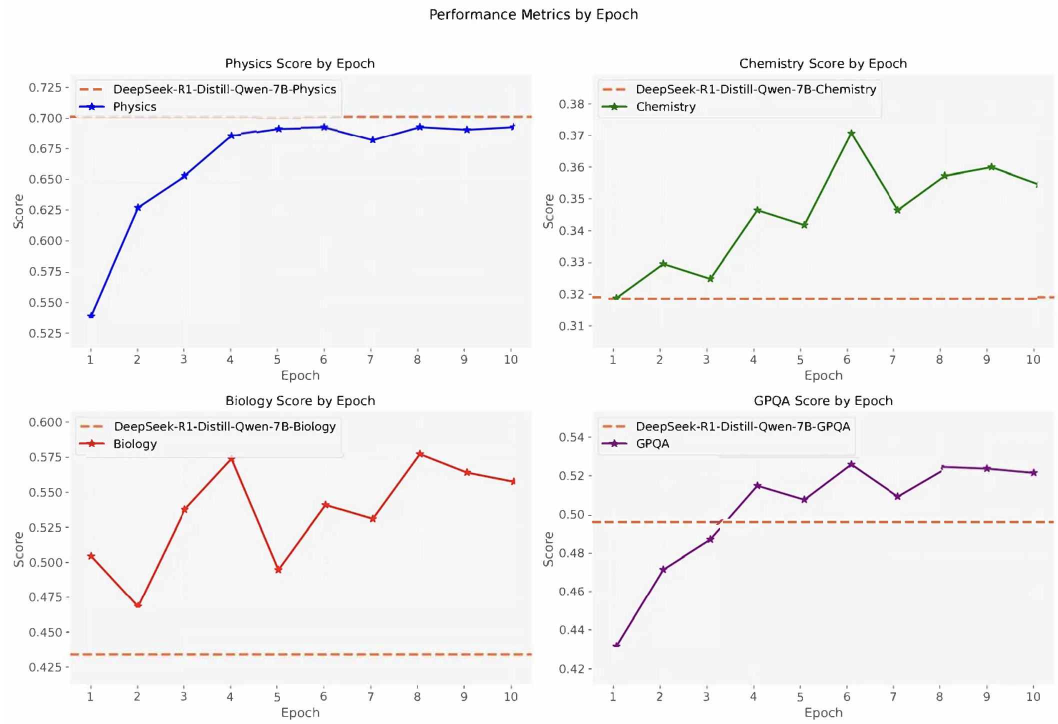

Figure 5: GPQA score improvement of single STEM disciplines (physics, chemistry, and biology) of **SHARP-Qwen2.5-7B-Instruct-Distill** relative to benchmark model **DeepSeek-R1-Distill-Qwen-7B** in ablation of STEM data generated by the **SHARP** approach. (The meanings of the x and y axes are the same as those in Fig.4.)

B.2 RL Zero Training Model Performance Analysis

As shown in Table 2 and Fig.7, after we added **SHARP** problems as the main training data (about 73%) for RL Zero enhanced reasoning training in **SHARP-Open-Reasoner-Zero-7B**, it has exceeded the pure mathematics RL Zero mathematical reasoning model **Open-Reasoner-Zero-7B (Baseline)** by about 4.22%, and the single subjects of physics and biology have exceeded the **Open-Reasoner-Zero-7B (Baseline)** model, and the chemistry subject is basically the same, which shows that the **SHARP** self-alignment strategy and inference training system implemented have improved the pure complex reasoning ability of the model. Especially for chemistry, we compare two key metrics for evaluation RL Zero training model: the response length (which usually is used to indicate the complexity of the problems), as shown in Fig.8 and reward value (whose values are usually used to indicate the difficulty degree of the problems) as shown in Fig.9 in three different problems datasets, 1) problems synthesized through traditional COT, 2) problems augmented synthesized referencing to real challenging chemistry exercises and 3) problems synthesized **SHARP** approach. Through the experimental comparison of each stage, the difficulty of the sample problems generated by our **SHARP** approach has significantly increased the response length for the correct answer, and the distribution of rewards has shown a significant downward trend. Although the GPQA score of the chemistry subject has not improved significantly, through combined with the gradual and significant improvement of the experimental evaluation indicators, it demonstrates the effectiveness of the **SHARP** approach in improving the complex reasoning ability of the model, and also indicates that increasing the model's own complex problems to obtain the groundtruth can further significantly increase the effect of the model.

Specific challenging problems generated by the **SHARP** approach used to train RL Zero models are shown in the Table 4, and other samples can be found via this link.

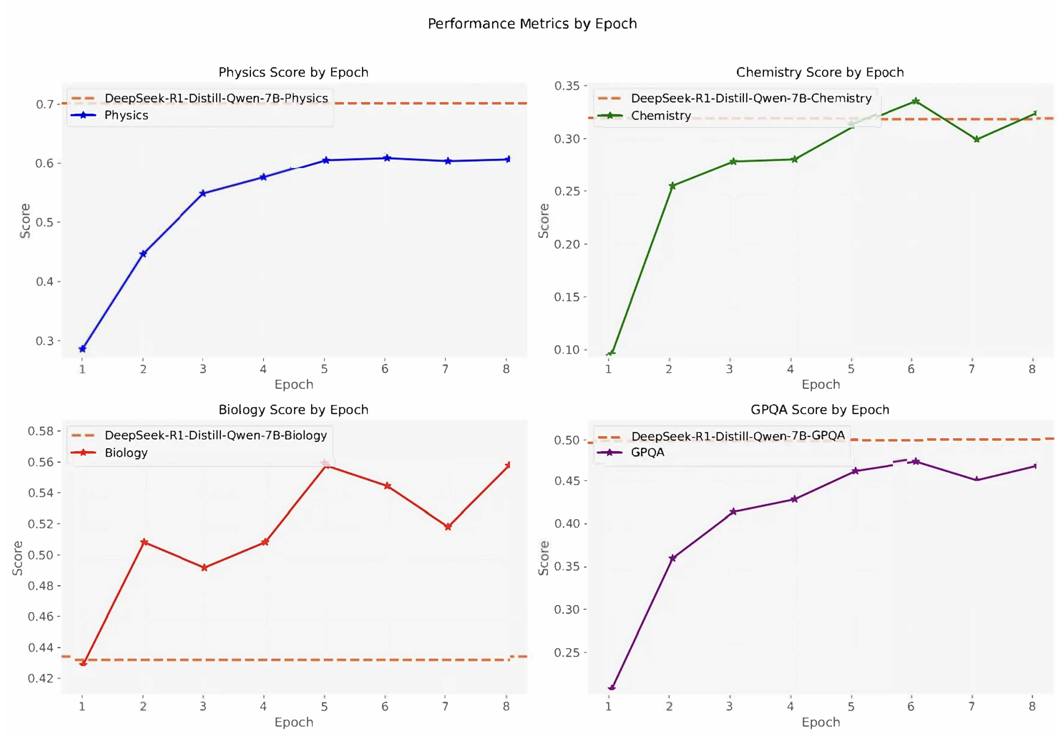

Figure 6: GPQA score improvement of single STEM disciplines (physics, chemistry, and biology) of **SHARP-Qwen2.5-7B-Instruct-Distill** relative to benchmark model **DeepSeek-R1-Distill-Qwen-7B** in ablation of some open-source mathematical data. (The meanings of the x and y axes are the same as those in Fig.4.)

To initialize RL Zero training, we employ the verified **SHARP**-generated dataset containing input questions and corresponding verified ground truth answers. We adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024), a memory-efficient reinforcement learning method well-suited to **SHARP**'s batch-verifiable training data. GRPO bypasses the need for a separate critic by estimating baselines from group-level sample scores. For each problem q, GRPO samples a group of outputs $o_1, o_2, \ldots, o_G$ from the old policy $\pi_{\theta_{\text{old}}}$ and then optimizes the policy model π_θ by maximizing the following objective:

$$
\begin{aligned}
\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\Big[& q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(O|q) \\
& \times \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left(\min\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t}, \text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\right) \\
& -\beta D_{\text{KL}}(\pi_\theta||\pi_{\text{ref}})\Big)\Big]
\end{aligned}
$$

$$
D_{\text{KL}}[\pi_\theta||\pi_{\text{ref}}] = \frac{\pi_{\text{ref}}(o_{i,t}|q,o_{i,<t})}{\pi_\theta(o_{i,t}|q,o_{i,<t})} - \log\frac{\pi_{\text{ref}}(o_{i,t}|q,o_{i,<t})}{\pi_\theta(o_{i,t}|q,o_{i,<t})} - 1. \quad (1)
$$

where ϵ and β are hyperparameters, and $\hat{A}_{i,t}$ is the advantage, computed using a group of rewards $\{r_1, r_2, ..., r_G\}$ corresponding to the outputs within each group:

$$
\hat{A}_{i,t} = \frac{r_i - \text{mean}(\{r_1, r_2, ..., r_G\})}{\text{std}(\{r_1, r_2, ..., r_G\})}
$$

Inspired by prior work (DeepSeek-AI, 2025), we employ a simplified reward function $\mathcal{R}$_acc grounded in binary accuracy. Unlike prior methods, **SHARP** enables reward assignment based

Figure 7: **Open-Reasoner-Zero-7B** performance in ablation of STEM data generated by the **SHARP** approach. (The x-axis represents the different running steps during the training of the reinforcement learning reasoning model, and the y-axis represents the GPQA score evaluation results corresponding to the checkpoints of the models generated at different steps.)

solely on correctness, owing to its verified outputs. This design simplifies the reward signal while preserving alignment quality. Here, the accuracy reward $\mathcal{R}$_acc evaluates correctness based on whether the model's response a_i is similar to the ground truth solution a_i' to satisfy the correctness criteria:

$$\mathcal{R}_{\text{acc}}(a_i, a_i') = \begin{cases} 1, & \text{if equal}(a_i, a_i'), \\ 0, & \text{otherwise.} \end{cases}$$

We implement GRPO training on RL Zero models using OpenRLHF (Hu et al., 2024), an open-source RL framework built atop Ray (Moritz et al., 2018), vLLM (Kwon et al., 2023), ZeRO-3 (Wang et al., 2024a), and HuggingFace Transformers (Wolf et al., 2020). We train these models with about 190,000 samples on 256 NVIDIA H800 GPUs for 48 hours. Key algorithm parameters for RL Zero models training are set as in Table 5. For each prompt, we generate 64 diverse completions to support robust group-based reward estimation. The KL divergence constraint coefficient is fixed at 0.001 across all experiments. Additionally, we mix problems from various STEM domains during model training to ensure diverse learning. We report accuracy by averaging the results over greedy decoding across 16 independent inference runs, which ensures statistical stability while preserving inference consistency for the GPQA benchmark evaluation. Answers are extracted from the standardized `\boxed{$Answer$}` format to verify against the ground truth solutions to ensure correctness and ensure alignment with **SHARP**'s verifiability constraints during automatic evaluation.

These results affirm that **SHARP**-aligned training samples, combined with GRPO and rule-based accuracy rewards, significantly enhance RL Zero model performance in complex STEM reasoning tasks.

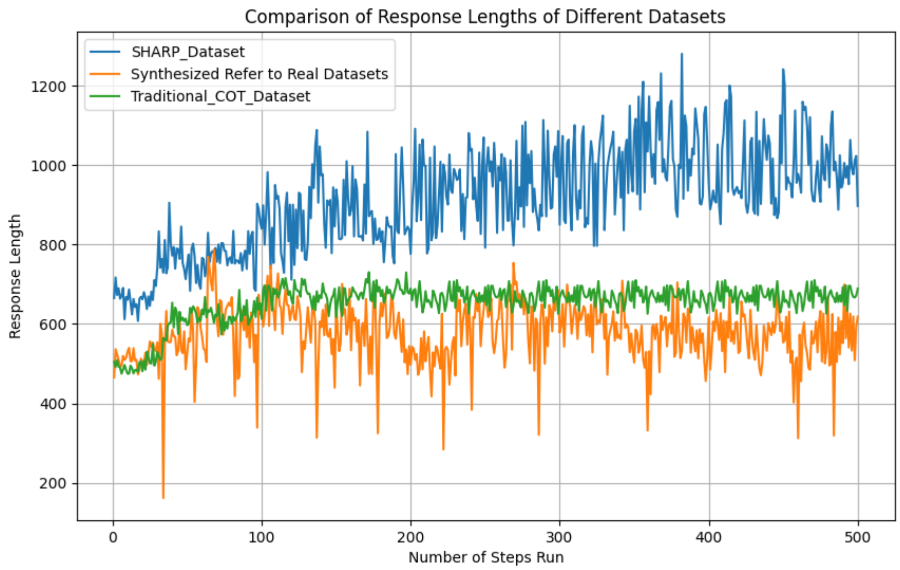

Figure 8: The response length of RL Zero model in ablation of chemistry data from three different problems datasets, 1) problems synthesized through traditional COT, 2) problems augmented synthesized referencing to real challenging chemistry exercises and 3) problems synthesized **SHARP** approach. (The x-axis represents the different running steps during the training of the reinforcement learning reasoning model, and the y-axis represents the response length when the models runs at corresponding steps.)

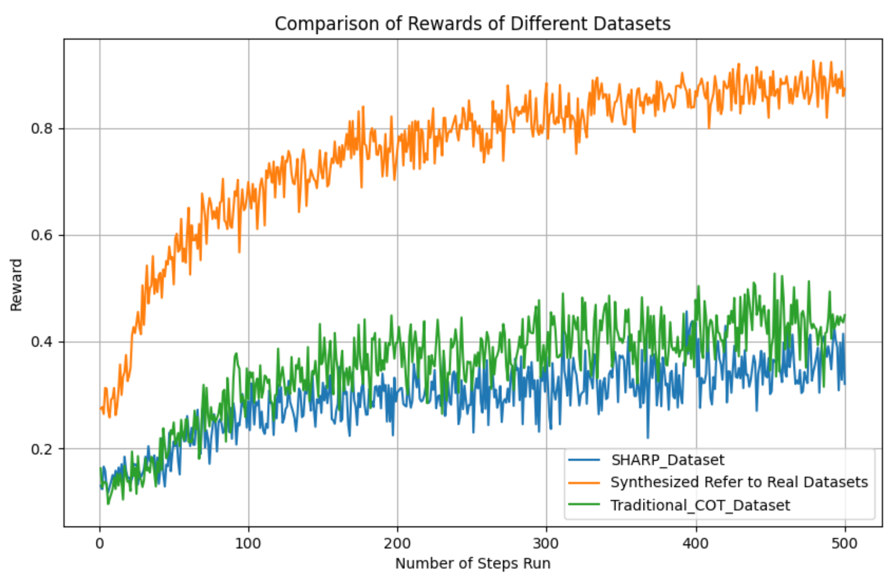

Figure 9: The reward RL Zero model in ablation of chemistry data from three different problems datasets, 1) problems synthesized through traditional COT, 2) problems augmented synthesized referencing to real challenging chemistry exercises and 3) problems synthesized **SHARP** approach. (The x-axis represents the different running steps during the training of the reinforcement learning reasoning model, and the y-axis represents the reward when the models runs at corresponding steps.)

Subject Name	Problem and Reference Answer
Particle Physics, High Energy Physics	"problem": "Solve the following chemical problem step by step. The last line of your response should be of the form `\boxed{$Answer$}` (without quotes) where $Answer$ is the answer to the problem. A core-collapse supernova at a distance of 1 kiloparsec (3.086×10^{19} meters) releases 3×10^{46} J of energy, with 99% of this energy emitted as neutrinos. Each neutrino has an average energy of 10 MeV (1.6×10^{-12} J). A spherical lead detector with a radius of 10 meters is used to observe these neutrinos. Lead has a density of 11,340 kg/m^3 and an atomic mass of 207.2 g/mol. The neutrino-nucleus interaction cross section is 1×10^{-43} cm^2 per nucleus. Assuming neutrinos are emitted isotropically and all physical quantities are uniform, calculate the total number of neutrino interactions in the detector. **Constants and Formulas:** Avogadro's number: $N_A = 6.022 \times 10^{23}$ mol^{-1} Sphere volume: $V = \frac{4}{3}\pi r^3$ Neutrino flux at Earth: $\Phi = \frac{N_\nu}{4\pi d^2}$ Interaction rate: $N_{\text{interactions}} = \Phi \cdot \sigma \cdot N_{\text{nuclei}}$ Please reason step by step, and put your final answer within `\boxed{$Answer$}`.", "ref_answer": "$\boxed{2140}$".
Organic Chemistry	"problem": "Solve the following chemical problem step by step. The last line of your response should be of the form `\boxed{$Answer$}` (without quotes) where $Answer$ is the answer to the problem. An impure sample of zinc carbonate ($ZnCO_3$) undergoes thermal decomposition, releasing carbon dioxide gas. The mass loss due to CO_2 emission is measured as 2.64 g. The resulting zinc oxide (ZnO) is then reduced using excess carbon, producing 5.89 g of zinc metal. 1. Write the balanced equation for the decomposition of $ZnCO_3$. 2. Write the balanced equation for the carbon reduction of ZnO. 3. Determine the percentage purity of zinc in the original impure sample. Assume all reactions proceed to completion, and impurities do not participate in any reactions. (Atomic masses: Zn = 65.38 g/mol, C = 12.01 g/mol, O = 16.00 g/mol) Remember to put your final answer within `\boxed{$Answer$}`", "ref_answer": "$\boxed{58.9\%}$".
Molecular Biology, Virology	"problem": "Solve the following biological problem step by step. The last line of your response should be of the form `\boxed{$Answer$}` (without quotes) where $Answer$ is the answer to the problem. The SARS-CoV-2 genome is a single-stranded RNA virus with a genome length of 29,903 nucleotides. The spike (S) protein gene constitutes 12.73% of the genome. Each S protein monomer consists of amino acids with an average molecular weight of 110 Da. A single virion contains 2.5 femtograms (fg) of S protein. Calculate the total number of S protein trimers on the virion's surface. Use Avogadro's number (6.022×10^{23} mol^{-1}) for your calculations. Remember to put your final answer within `\boxed{$Answer$}`", "ref_answer": "$\boxed{3586}$".

Table 4: The challenging problems of physics, chemistry, and biology generated by the **SHARP** approach.

C SHARP Challenging Problem Datasets Analysis

In this section, we first provide a detailed supplementary explanation of the overall dataflow process of the **SHARP** approach. Then we further analyze the coverage of the subject categories related to

Parameter Name	Value	Description
algorithm	GRPO	Reinforcement Learning Algorithm Used
actor_lr	1e-6	Learning Rate of The Actor Network
rollout_bs	256	Total Batch Size Used for Experience Collection
train_bs	16384	Total Batch Size Used During Parameter Updates
micro_train_bs	8	Batch Size for a Single Forward Pass During Training
micro_rollout_bs	8	Batch Size for a Single Forward Pass During Experience Collection
sample_k	64	Number of output samples (G) generated per prompt by the policy for group reward estimation
lambda	1.0	Regularization Coefficient
gamma	1.0	Discount Factor
kl	0.001	KL Divergence Constraint Coefficient
max_len	8192	Maximum Sequence Length
temperature	1.0	Sampling Temperature

Table 5: RL Zero Algorithm Core Parameters. Parameter values were tuned based on ablations to balance training stability, efficiency, and model performance.

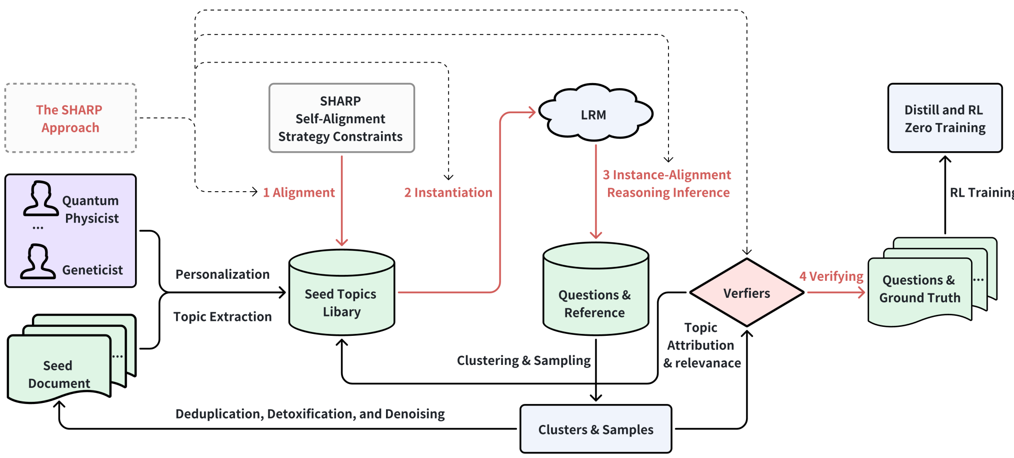

Figure 10: The overall dataflow process of the **SHARP** approach.

the data flow and the difficulty of the STEM challenging problems generated by the **SHARP** method based on this category.

The overall dataflow diagram of the construction of the **Seed Topics Library** and "Three-Tier Category" knowledge structure for STEM problems in the **SHARP** approach is shown in Fig.10. As mentioned, they were mainly built by methods combined with the persona method (Ge et al., 2025) and the Magpie-like method(Xu et al., 2025b) to generate a large number of personalized target topic query problems. Furthermore, in order to ensure the diversity and balance of the generated problems, a clustering strategy is designed, and these questions are distributed and balanced. In this way, we ensure that the generated problems cover a wide enough range of topics and have enough diversity under each topic, so as to provide comprehensive and balanced training problems for training LRMs.

Persona-driven (Ge et al., 2025) prompts simulate domain experts with distinct problem-creation styles (e.g., a theoretical physicist vs. an organic chemist), ensuring varied problem framing and difficulty levels. Based on the persona method, we further improved the Magpie method to generate a large number of target topic query questions. We first designed a new "Three-Tier Category" knowledge structure with reference to the BISG category organization (Book Industry Study Group, 2025) and subject characteristics to ensure that the first-level sub-disciplines, second-level self-disciplines, and basic concepts of each discipline are covered. Then we built high-quality seed

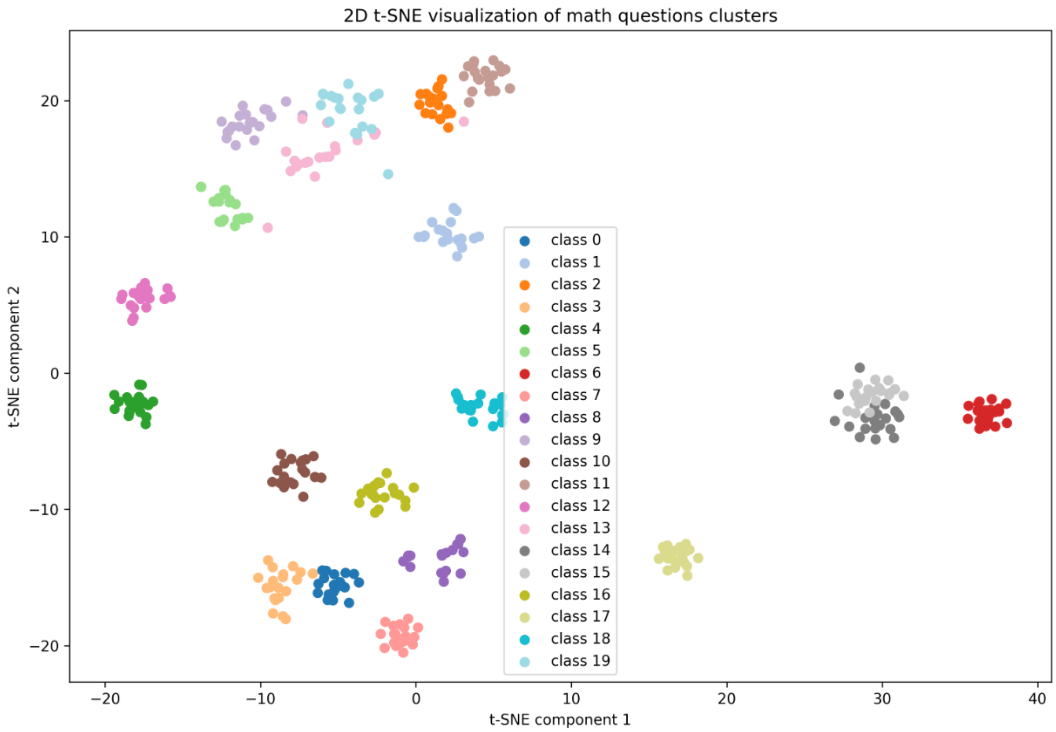

Figure 11: The K-means clustering results based on question embedding features extracted using BGE-m3 from problems generated by the **SHARP** approach.

documents to supplement and improve the themes of the "Three-Tier Category" knowledge structure of **SHARP** through the following two aspects. On the one hand, we analyzed and extracted topic keypoints based on high-quality open source training sets and question bankmarks such as **GPQA** (here, we only extracted topic keypoints without quoting or rewriting problems to prevent data leakage). On the other hand, we recall seed documents based on high-quality STEM textbooks, academic papers, Common Crawl, etc., and extract topics through the latest reasoning models such as Deepseek R1 and Qwen3 to obtain better topic diversity, thereby improving the model's generalization ability. Through the above series of methods, we ensure that the query problems have sufficient coverage while having expert persona characteristics, and at the same time ensure that the generated problems are consistent with the distribution of the current benchmark but have sufficient diversity and depth, so as to provide comprehensive, rich and challenging problems training dataset support for the complex reasoning of LLMs. In addition, we use BGE-m3 (Chen et al., 2024) to extract embedding features from the generated problems, and then use the K-means (MacQueen, 1967) algorithm for clustering. We specify about 1,000 clusters via elbow method analysis on BGE-m3 embeddings to ensure that the number of clusters can cover most of the query problems, while ensuring that the queries within each cluster have a certain similarity and that there is sufficient difference between clusters. While clustering, each class is uniformly sampled to ensure the class balance of samples, and then an appropriate number of samples is extracted from each class for training. Fig. 11 illustrates the clustering results based on query embeddings, where we visualize a representative subset of 20 clusters. Each cluster exhibits strong intra-cluster cohesion, with samples tightly grouped in the embedding space. This suggests that queries within the same cluster share high semantic similarity. Moreover, clusters are well-separated from one another, indicating low semantic overlap across different groups. The clear inter-cluster boundaries highlight the effectiveness of our clustering pipeline in capturing meaningful semantic distinctions. Finally, the clustering and sampling results are processed for data deduplication, detoxification, and decontamination. Specific examples of the "Three-Tier Category" knowledge structure in the **Seed Topics Library** are shown in Table 6.

The "Three-Tier Category" structure, integrated with persona-driven prompts and clustering, ensures thematic diversity and logical consistency in SHARP-generated problems, directly contributing to

First-level Discipline	Second-level Discipline	Basic Knowledge-points
Theoretical Physics, High Energy Physics	Quantum Mechanics, Particle Physics	Energy levels, Heisenberg uncertainty principle, Lifetime-energy uncertainty relation, Energy resolution, Quantum states
Organic Chemistry	Stereochemistry	Hydrogenation, Epoxidation, Nucleophilic substitution, Esterification, Limonene, Peracids, DCC coupling
Basic Biology	Molecular Biology, Cancer Biology, Genetics, Epigenetics	Tumor suppressor genes, Gene expression, Epigenetic regulation, Gene silencing, Mouse models, Cancer cells

Table 6: The "Three-Tier Category" structure examples of physics, chemistry and biology in the **SHARP Seed Topics Library**.

enhanced model performance. Based on these data flow processing, the **SHARP** approach first combines the self-alignment strategy as shown in Algorithm 1 to generate problems that help guide the reinforcement reasoning model at multiple levels. By integrating the persona (Ge et al., 2025) role, the "Three-Tier Category" structure, and the **SHARP** self-alignment strategy, the following template for creating problems is designed, as shown in the Table 8. Then, the problem template is fused with the actual "Three-Tier Category" structure and knowledge framework through instantiation reasoning, thereby generating complex reasoning problems with self-alignment conditions and their corresponding reference answers. These complex questions and reference answers are then further verified, and finally, a high-quality, challenging question set and ground truth for complex reasoning are generated. The generated problems are not only conducive in characteristic disciplines and enhancing the generalization reasoning capabilities, but also generating difficult and logically consistent problems and corresponding verifiable answers that are conducive to LRMs through reinforcement learning via verifiable rewards (RLVR).

To enable scalable generation of **SHARP**-aligned STEM reasoning problems, we deploy the DeepSeek R1 model using the SGLang inference framework (Zheng et al., 2024), selected for its high-throughput serving capabilities, long-context support, and compatibility with structured output formatting.

This deployment is integrated into the **SHARP Instantiation** and **Inference** phases, where DeepSeek R1 is queried using templated prompts instantiated from persona roles, seed topics, and self-alignment constraints. Table 7 summarizes the server and inference configurations. Each request follows the **SHARP** prompting strategy (Table 8), with the `messages` field encoding the relevant persona roles, subject categories, and alignment directives.

The system, deployed on 8 NVIDIA H20 GPUs, supports 16K-token contexts and accommodates up to 48 concurrent requests, enabling efficient generation of high-difficulty, verifiable reasoning problems across diverse STEM domains. In total, 229,452 question–answer pairs were generated over 168 hours and subsequently evaluated through the **SHARP Verifying** phase for quality assurance prior to integration into RL Zero training.

The raw generated data samples for creating the challenging problems using the **SHARP** approach can be found via this link.

Next, we conduct a detailed analysis of the coverage of evaluated benchmarks (mainly on GPQA as an example), subject categories related to the dataflow and the difficulty of the STEM challenging problems (mainly on physics, chemistry, and biology) generated by the **SHARP** method.

C.1 Key Knowledge Point Distribution

Through statistical analysis of the distribution of labels and knowledge points, we found that the GPQA benchmark is unevenly distributed, relevant key points distributions shown as in Fig.13, 15, 17 and basic knowledge points shown in Fig.14, 16, 18. Therefore, in the SHARP data synthesis method, we sampled according to the distribution of disciplines and the corresponding knowledge

Parameter Name	Value	Description
tp	8	Tensor Parallelism
max-running-requests	48	Concurrency
mem-fraction-static	0.94	Memory Allocation
context-length	16384	Context Length
temperature	0.6	Temperature
max_tokens	4192	Maximum Output Tokens
repetition_penalty	1.05	Repetition Penalty
top_p	0.8	Top-p Sampling

Table 7: Key SGLang server and inference parameters for each **SHARP** instantiated problem generation prompt.

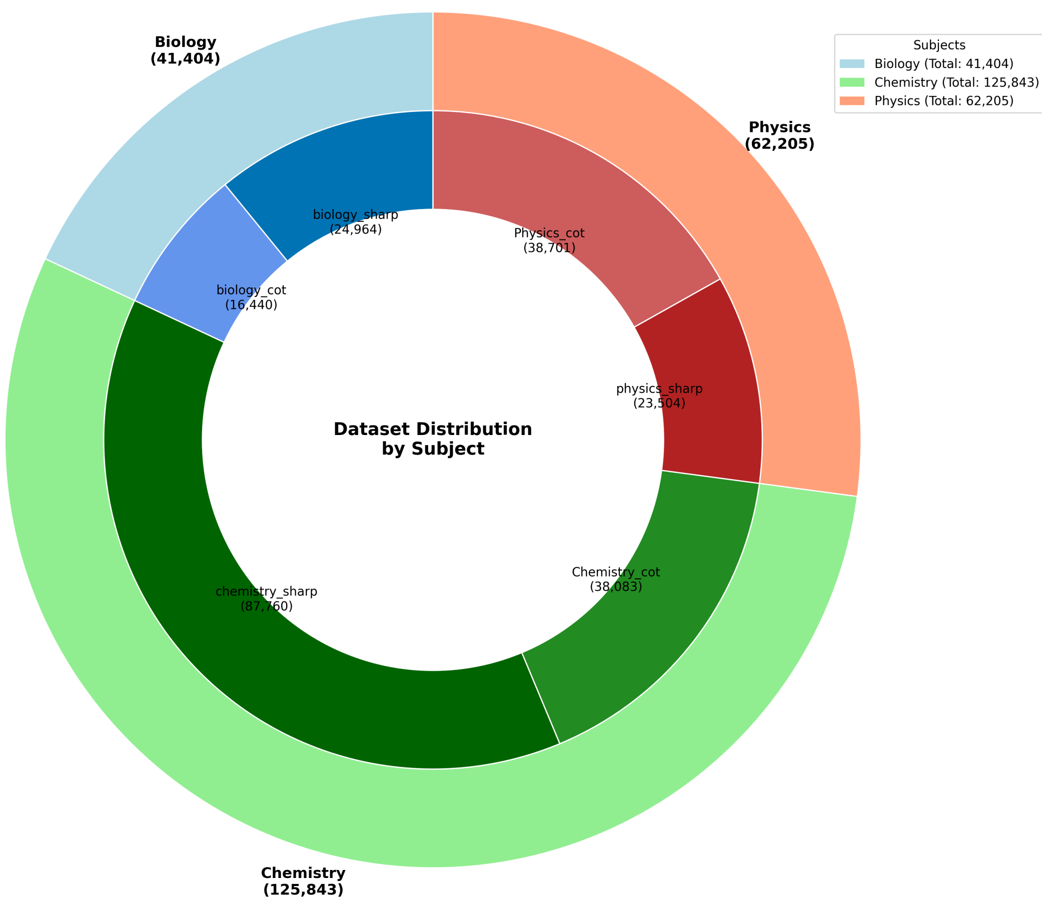

Figure 12: The overall subjects distribution of 229,452 question-answering problems generated by the **SHARP** approach and the baseline traditional CoT method.

points to ensure that the synthetic data fully covers the relevant knowledge points of GPQA in terms of distribution.

C.2 Category Distribution Analysis

We carefully analyzed the 229,452 STEM question-answering problems generated by the **SHARP** approach and the baseline traditional CoT method (about 190,000 questions remained after disinfection, deduplication, and decontamination, and pass ratio filtering), with the subject distribution

The SHARP Approach Prompt

<Role_Start>
To test the < **Subject_Name:** **{subject_name}**> reasoning and complex problem-solving skills of talented graduate students across various <**Subject_Name:{subject_name}**> disciplines, you, a <**Persona_Role**: **{persona_role}** > at a world-renowned institution, are creating a graduate- or Olympic-level challenging problem.
<Role_End>
<Task_Description_Start>

- You MUST refer to the following resources: **<SUB> Subject_Name:** **{subject_name}** **Subdisciplines**: **{subdisciplines}<SUB>**, **<BC>Basic Concepts**: **{basic_concepts}<BC>**.
- You MUST randomly choose one or more items from the **<SUB> Subject_Name:** **{subject_name}** **Subdisciplines**: **{subdisciplines}<SUB>**, and then select several related concepts from the **<BC>Basic Concepts:** **{basic_concepts}<BC>** according to the subdisciplines to form an outline for the problem. Finally, create a calculation problem.

<Task_Description_End>
<Requirements_and_Expectations_Start>
Note: The problem must satisfy the following self-alignment constraints:

- **Problem Difficulty & Thematic Diversity Alignment**: Generate highly complex problems (graduate- or Olympiad-level) covering a wide range of STEM topics. Difficulty is benchmarked against top exams and datasets (GPQA, etc.). Thematic coverage uses role-playing prompts template and a three-tier subject-category-topic framework.
- **Logical Consistency Alignment**: Problem-solving must rely solely on rigorous reasoning or systematic derivation, avoiding pattern matching, heuristics, shortcuts, or fabrication. All intermediate steps require justification, preventing logical gaps or errors due to intuition.
- **Ground Truth & Structure Alignment**: Answers must be single, verifiable numerical values (plain numbers, units, ratios, STEM formulas/equations). Avoid hard-to-verify formats (set operations, free text). For multi-solution problems, mandate a specific aggregation (e.g., sum or sum of squares, etc.) for a unique, objectively verifiable answer. Expand beyond single QA to include multi-solution problems (requiring summary values) (e.g., "calculate total moles of all possible products").
- **Problem Authenticity Alignment**: Problems should be novel, based on authoritative knowledge, but not directly copied. They must be unambiguous, unbiased, accurate, and internally consistent, avoiding nonsensical or hallucinated scenarios.
- **Language Consistency Alignment**: The entire generation process (problem statement, reasoning method, solution presentation) must use a single language (e.g., English or Chinese) to prevent multilingual confusion leading to reasoning errors or bad verification cases.
- **Problem Structure Consistency Alignment**: Problems must contain only a single primary question, avoiding sub-questions, derivatives, or branching logic that leads to unverifiable cases.
- **Modality Consistency Alignment**: Problems must be strictly text-based, describing any necessary complex structures (e.g., chemical molecules, genetic diagrams) textually.
- **Formatting Alignment**: Use specific delimiters (e.g., <question_start>, <question_end>) for the problem statement and a standardized format (e.g., \boxed{{$answer}}) for the final answer.

<Requirements_and_Expectations_End>

Table 8: The **SHARP** prompt to synthesize high-quality aligned reasoning problems for LRMs reinforcement learning. The colored variables with curly braces in the prompt template are the variables corresponding to the algorithm framework, which will be instantiated with specific values for problem generation.

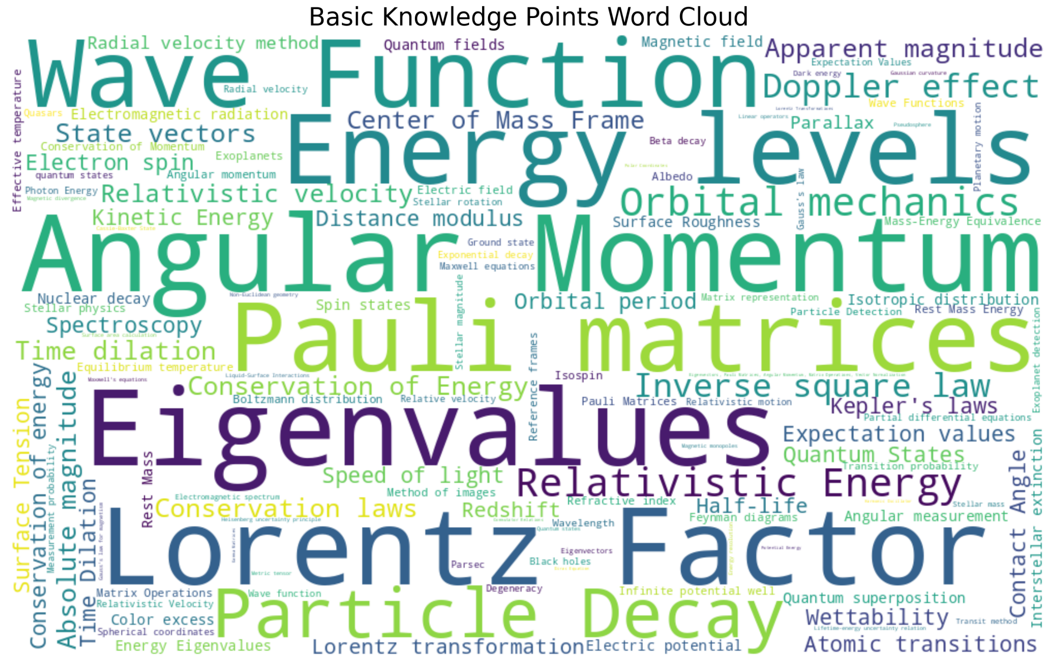

Figure 13: The physics subject distribution of basic knowledge points word cloud of GPQA benchmark.

across physics, chemistry, and biology shown in Fig.12. The distribution of subject categories for each subject of physics, chemistry and biology is described below.

Physical Category Distribution Analysis The category distribution of the synthetic dataset is presented in the following Fig.19. As observed, the data adheres to a scientifically structured three-level taxonomy. The first-level categories "Theoretical Physics", "Mechanics", and "Electromagnetism" are the top three categories, encompassing critical second-level disciplines such as Quantum mechanics, Fundamental mechanics, and Electrodynamics. These branches further decompose into highly specialized third-level categories like Theoretical Mechanics, Wave Functions and Schrodinger Equations, Electrostatic Fields, Laws of Thermodynamics—domains particularly effective for evaluating models' reasoning and computational capabilities. Notably, the dataset maintains substantial diversity despite this concentration, boasting over 200 distinct third-level categories. This comprehensive coverage across diverse physics domains ensures robust training signals, enabling models to develop balanced proficiency in both dominant and niche scientific reasoning tasks.

Biology Category Distribution Analysis The category distribution of the synthetic dataset is presented in the following Fig.20. As observed, the data adheres to a scientifically structured three-level taxonomy. The first-level category "Fundamental Biology" dominates with over half of the samples, encompassing critical second-level disciplines such as molecular biology, genetics, and cell biology. These branches further decompose into highly specialized third-level categories like molecular genetics, gene expression, and DNA repair mechanisms—domains particularly effective for evaluating models' reasoning and computational capabilities. Notably, the dataset maintains substantial diversity despite this concentration, boasting over 100 distinct third-level categories. This comprehensive coverage across diverse biological domains ensures robust training signals, enabling models to develop balanced proficiency in both dominant and niche scientific reasoning tasks.

Chemistry Category Distribution Analysis The analysis of category distribution within the synthetic chemistry dataset is illustrated in the accompanying Fig.21. A careful examination reveals that the data adheres to a rigorously structured three-tier category. At the first level, the category of "Organic Chemistry" is predominant, representing more than 75% of the total samples. This primary category encompasses significant second-level disciplines, including unsaturated hydrocarbons, pericyclic reactions, and the methodologies for characterizing organic compounds. These second-level classi-

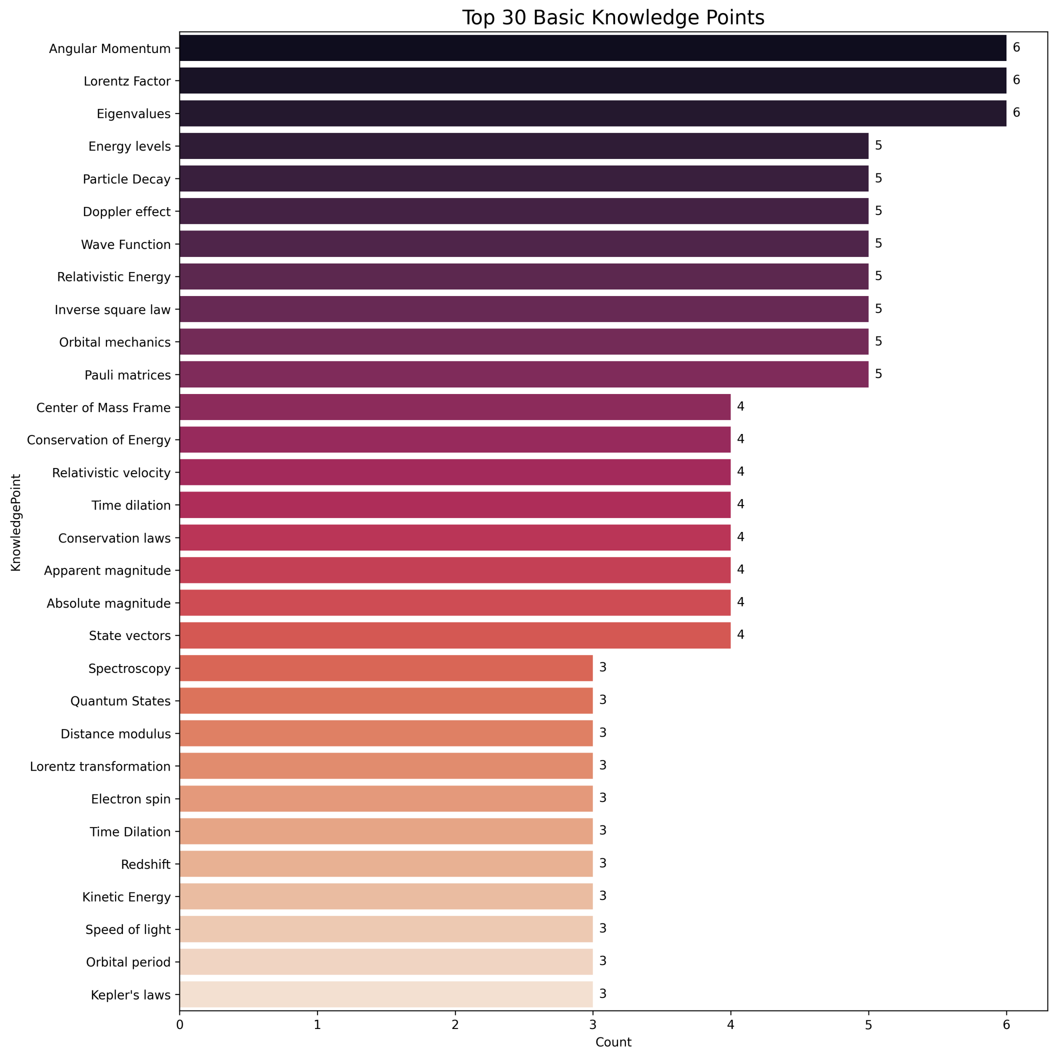

Figure 14: The top 30 basic knowledge points of the physics subject of the GPQA benchmark.

fications are further delineated into specialized third-level categories, such as olefins, electrocyclic reactions, and H-NMR nuclear magnetic resonance spectroscopy, which are particularly effective in assessing the reasoning and computational capabilities of the models employed. Importantly, notwithstanding the dominance of "Organic Chemistry," the dataset exhibits a commendable level of diversity, with over 300 distinct third-level categories represented. This extensive range of coverage across various domains of chemistry fosters robust training signals, thereby facilitating the development of models that exhibit balanced proficiency in both mainstream and niche scientific reasoning tasks.

C.3 Problem Datasets Difficulty Degree Analysis

In this section, we present a comprehensive analysis of the difficulty of the problems generated by the **SHARP** method.

STEM Pass Rate Distribution Comparison Fig.22, 23 and Fig.24 illustrate the pass rate distributions of three subjects among three different datasets: the open-source dataset (here refers to the open source data that is mainly based on real data in the industry, with a small amount of open source synthetic data, which is high-quality and challenging after being cleaned, deduplicated and decontaminated), the traditional CoT synthetic dataset, and our **SHARP** synthetic dataset. The

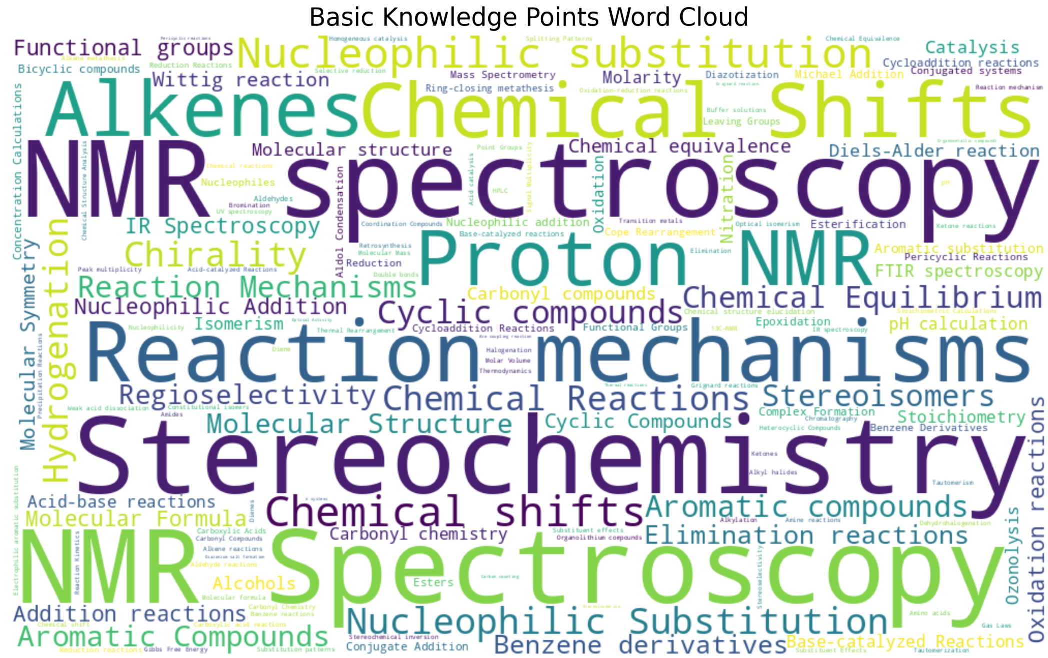

Figure 15: The chemistry subject distribution of basic knowledge points word cloud of GPQA benchmark.

pass rate is defined as the percentage of correct answers generated by the Qwen2.5-32B-Instruct model over five attempts, where a lower pass rate indicates a higher difficulty level of the question. As shown in the figure, the difficulty distribution of the **SHARP** synthetic dataset closely aligns with that of the real-world open-source dataset, making it a viable extension for enhancing the diversity and representativeness of real data. In contrast, the traditional CoT synthetic dataset exhibits an imbalanced difficulty distribution, with a skewed concentration of either very easy or very challenging questions. Furthermore, the **SHARP** dataset demonstrates a well-distributed pass rate across intermediate difficulty levels, providing a multi-level difficulty spectrum for model training. This balanced distribution enables hierarchical enhancement of the model's reasoning capabilities, ensuring progressive learning and robust performance across tasks of varying complexity.

Physics Pass Rate Distributions on the SHARP Dataset Fig.25 illustrates the pass rate distributions of two models on the **SHARP** dataset: Qwen2.5-32B-Instruct (based on ten independent responses) and QwQ-32B (Qwen, 2025) (based on five responses). As shown, the QwQ-32B model, which exhibits stronger reasoning capabilities, achieves a significantly higher overall pass rate compared to the Qwen2.5-32B-Instruct model. This is evidenced by a notable reduction in the proportion of questions with a pass rate of 0 and a corresponding increase in the proportion of questions with a pass rate of 1. These results demonstrate the effectiveness of the **SHARP** dataset in distinguishing the reasoning capabilities of models. By clearly differentiating between models of varying strengths, the **SHARP** problems dataset can be used to enhance the performance of LLMs' reasoning models in complex tasks.

Biology Pass Rate Distributions on the SHARP Dataset Fig.26 illustrates the pass rate distributions of two models on the **SHARP** dataset: Qwen2.5-32B-Instruct (based on five independent responses) and QwQ-32B (based on a single response). As shown, the QwQ-32B model, which exhibits stronger reasoning capabilities, achieves a significantly higher overall pass rate compared to the Qwen2.5-32B-Instruct model. This is evidenced by a notable reduction in the proportion of questions with a pass rate of 0 and a corresponding increase in the proportion of questions with a pass rate of 1. These results demonstrate the effectiveness of the **SHARP** dataset in distinguishing the reasoning capabilities of models. By clearly differentiating between models of varying strengths, the training dataset generated by **SHARP** can be used to enhance the performance of reasoning models in complex tasks.

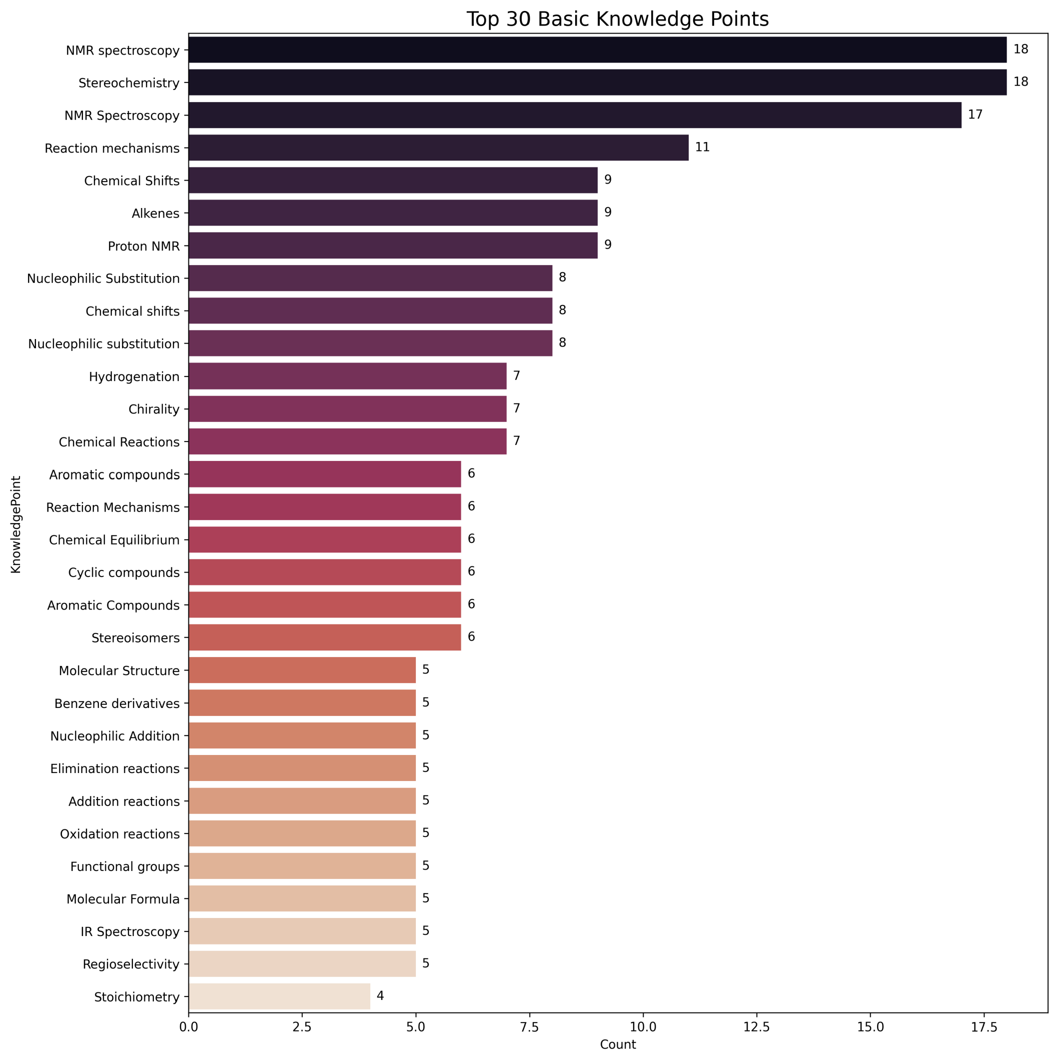

Figure 16: The top 30 basic knowledge points of the chemistry subject of the GPQA benchmark.

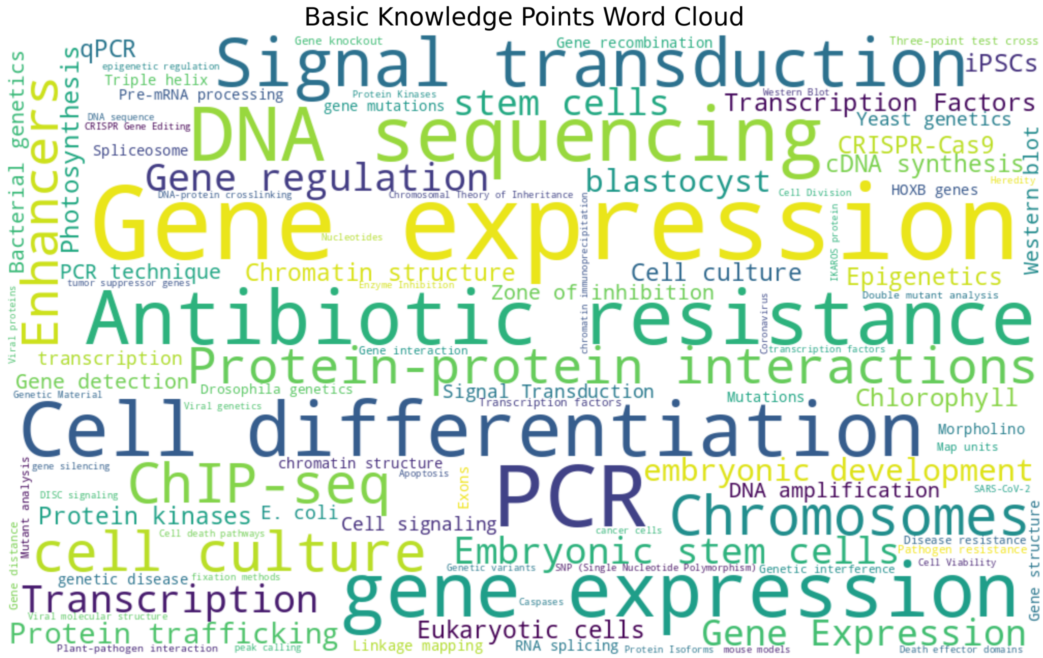

Figure 17: The biology subject distribution of basic knowledge points word cloud of GPQA benchmark.

Chemistry Pass Rate Distributions on the SHARP Dataset The following Fig.27 presents a comparative analysis of the pass rate distributions for two distinct models evaluated on the **SHARP** dataset: Qwen2.5-32B-Instruct, which is based on ten independent responses, and QwQ-32B, which relies on five singular responses. The data indicates that the QwQ-32B model, characterized by superior reasoning capabilities, achieves a markedly higher overall pass rate in comparison to the Qwen2.5-32B-Instruct model. This is evidenced by a significant decrease in the proportion of questions that registered a pass rate of 0, alongside a corresponding increase in the proportion of questions attaining a pass rate of 1. These findings underscore the efficacy of the **SHARP** dataset in differentiating between the reasoning capabilities of various models. By effectively distinguishing between models with disparate strengths, the **SHARP** dataset can be used to further enhance the reasoning capabilities of LLMs engaged in complex tasks.

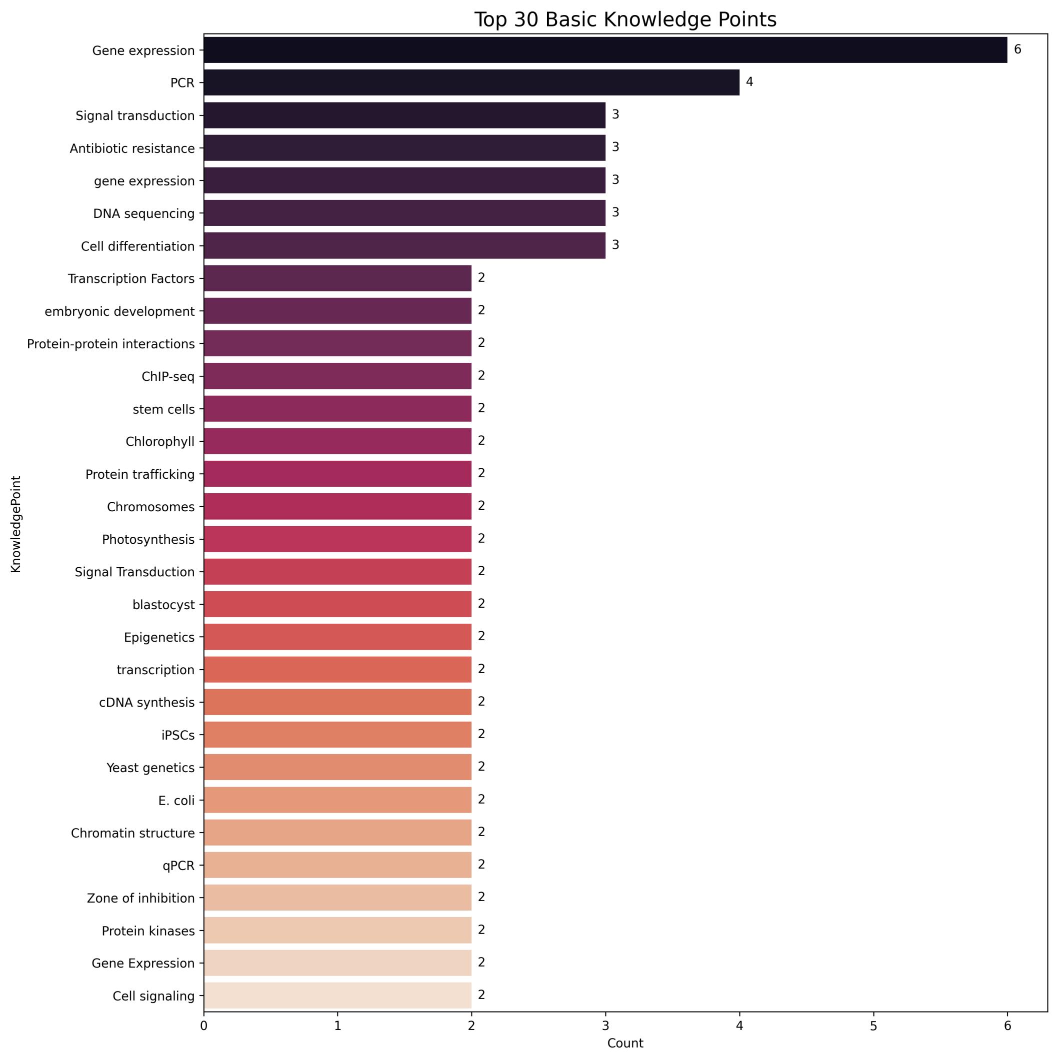

Figure 18: The top 30 basic knowledge points of the biology subject of the GPQA benchmark.

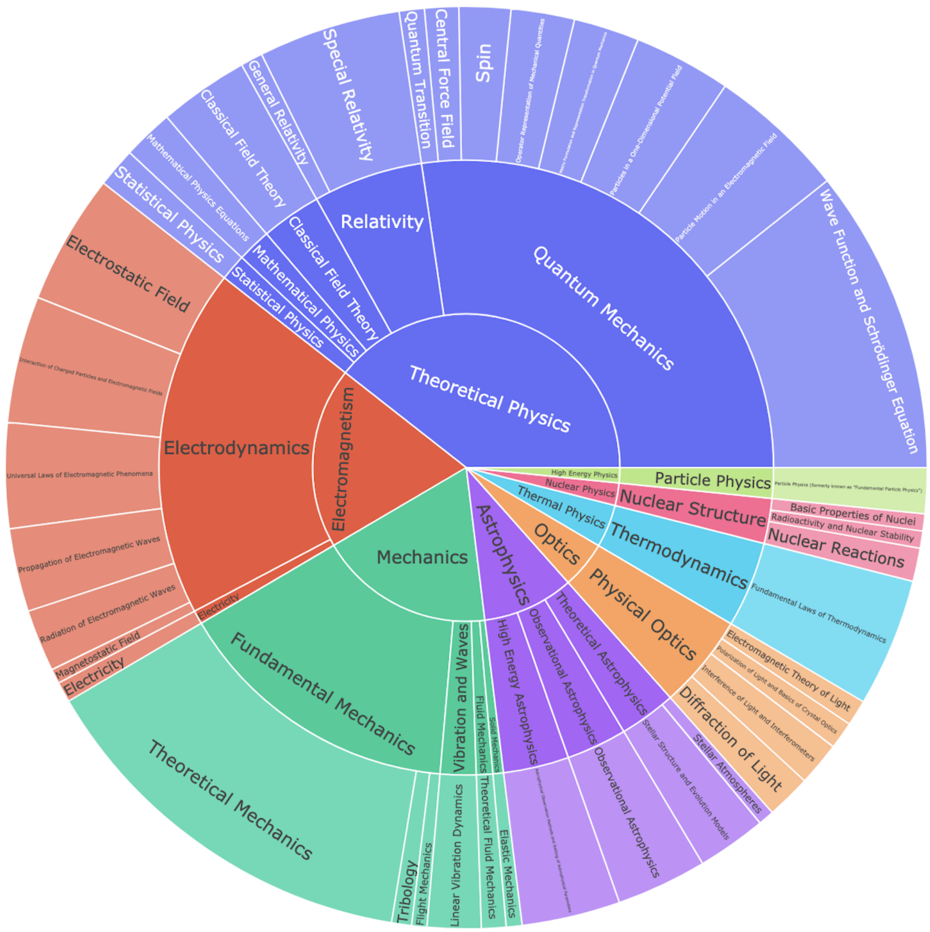

Figure 19: The "Three-Tier Category" category distribution of physics subject for problems generated by the **SHARP** approach.

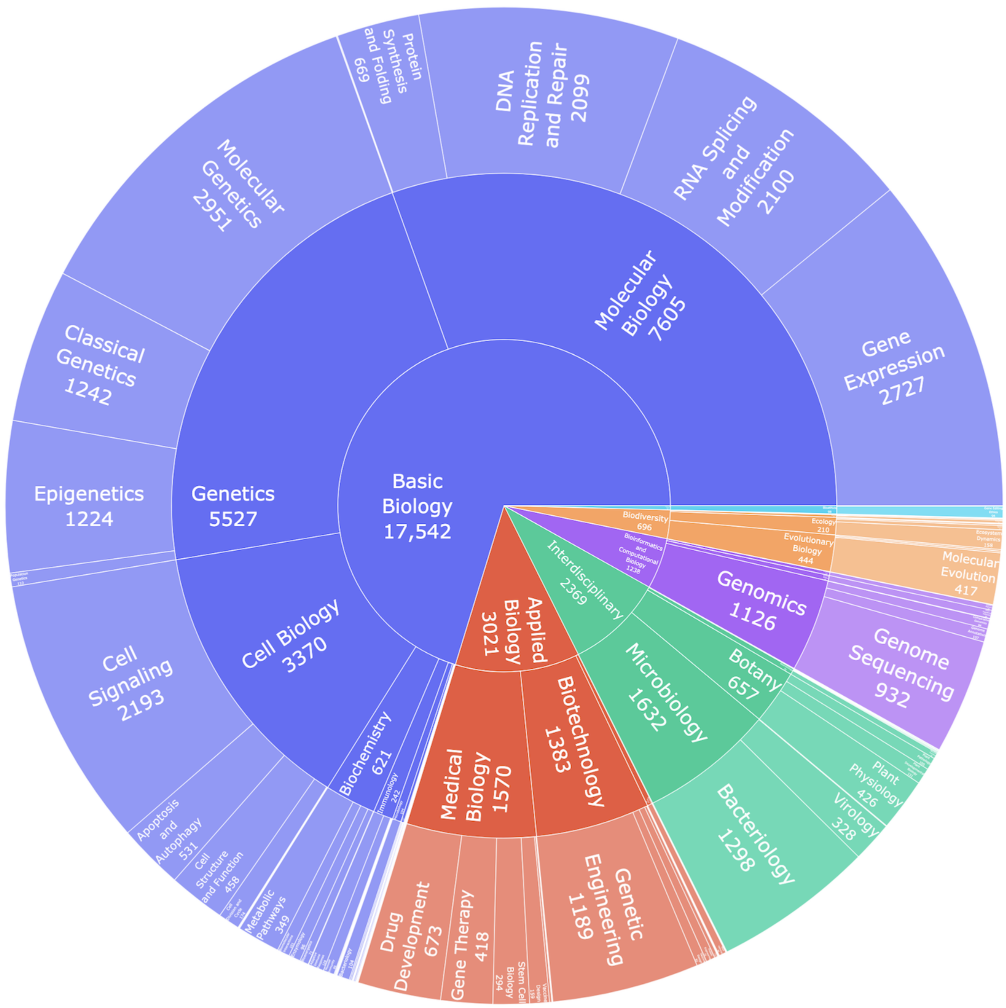

Figure 20: The "Three-Tier Category" category distribution of biology subject for problems generated by the **SHARP** approach.

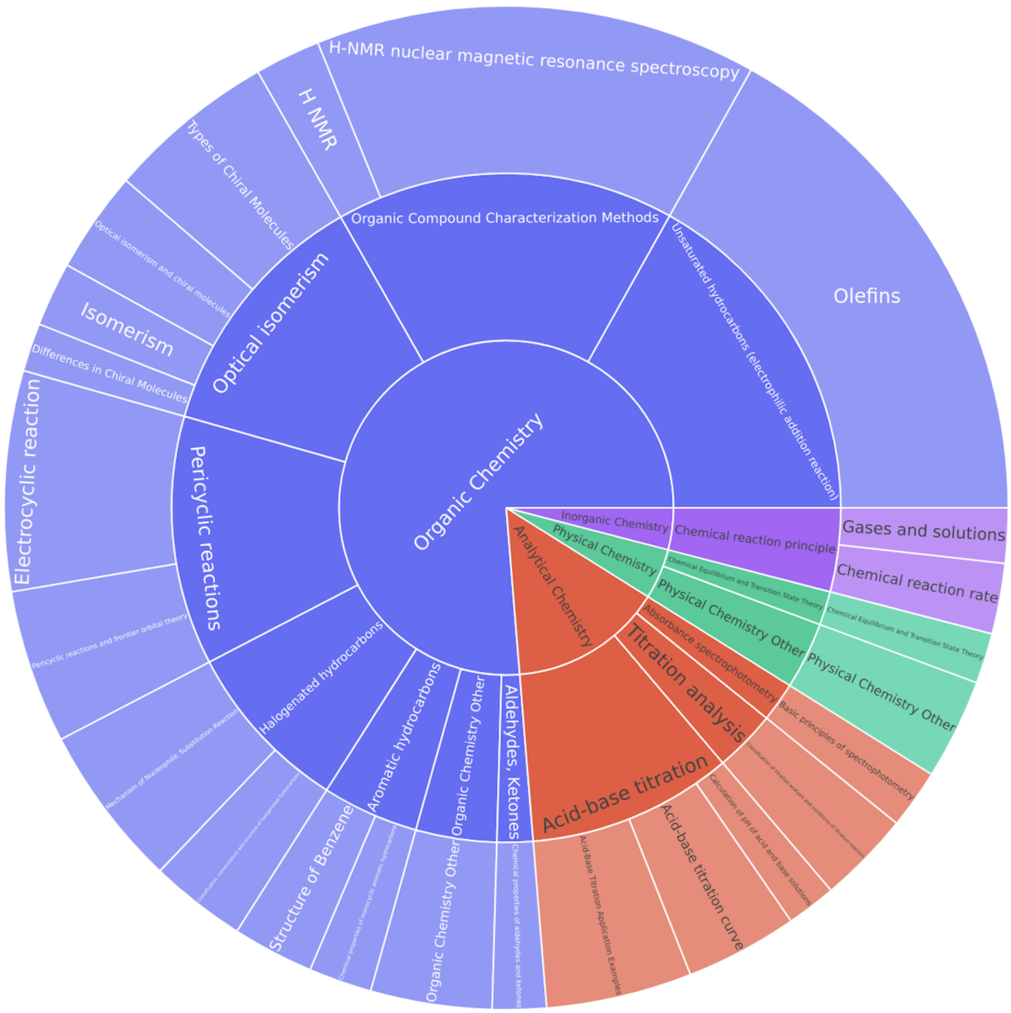

Figure 21: The “Three-Tier Category” category distribution of chemistry subject for problems generated by the **SHARP** approach.

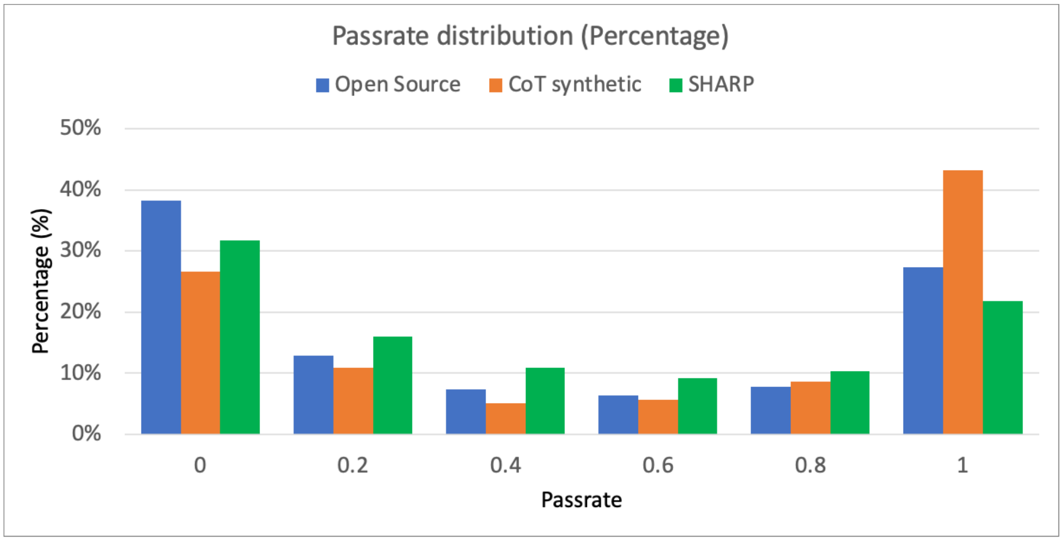

Figure 22: The passrate distribution of the physics problems from open-source , the traditional CoT synthetic, and generated by the **SHARP** approach.

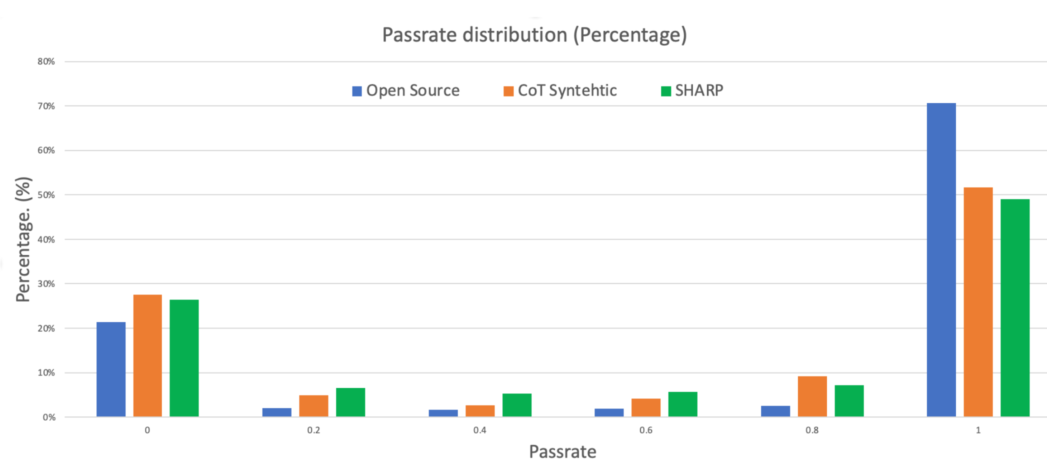

Figure 23: The passrate distribution of the chemistry problems from open-source , the traditional CoT synthetic, and generated by the **SHARP** approach.

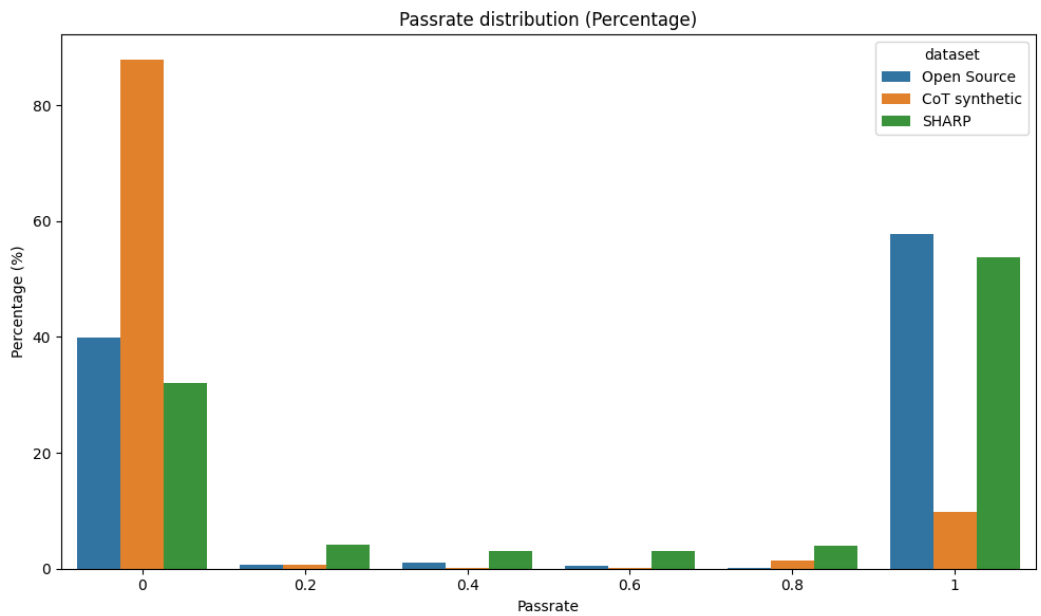

Figure 24: The passrate distribution of the biology problems from open-source , the traditional CoT synthetic, and generated by the **SHARP** approach.

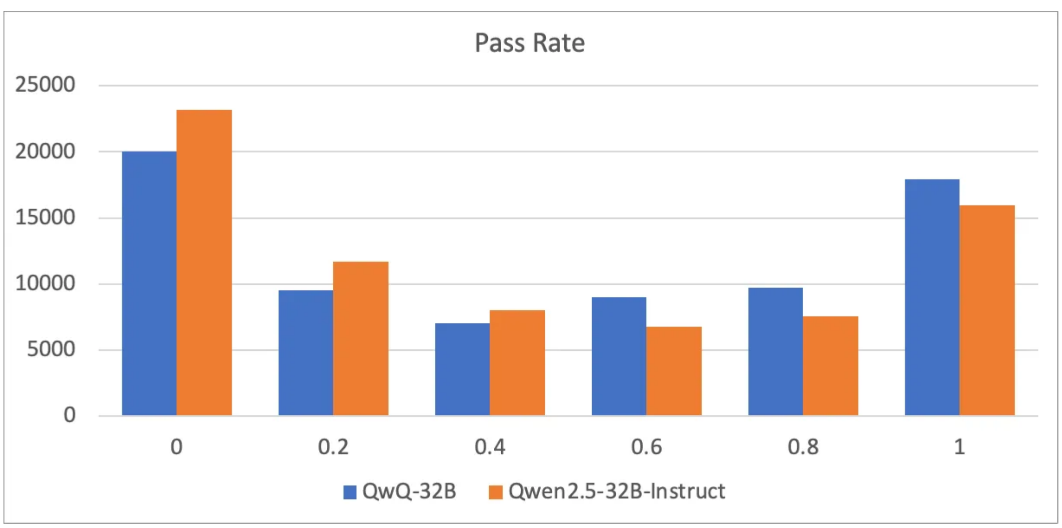

Figure 25: The passrate distribution of physics problems generated by the **SHARP** approach.

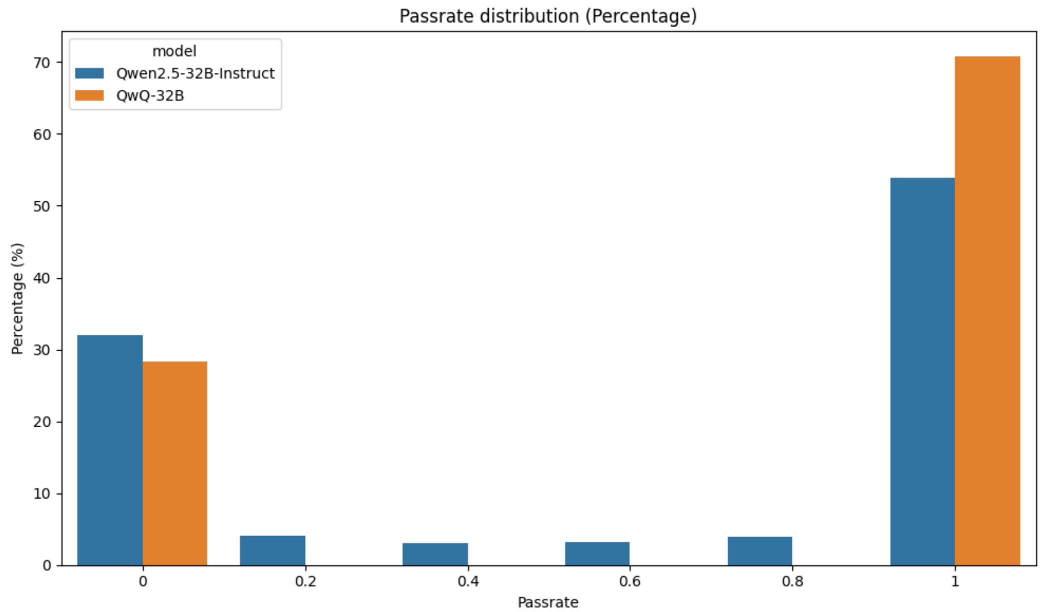

Figure 26: The passrate distribution of biology problems generated by the **SHARP** approach.

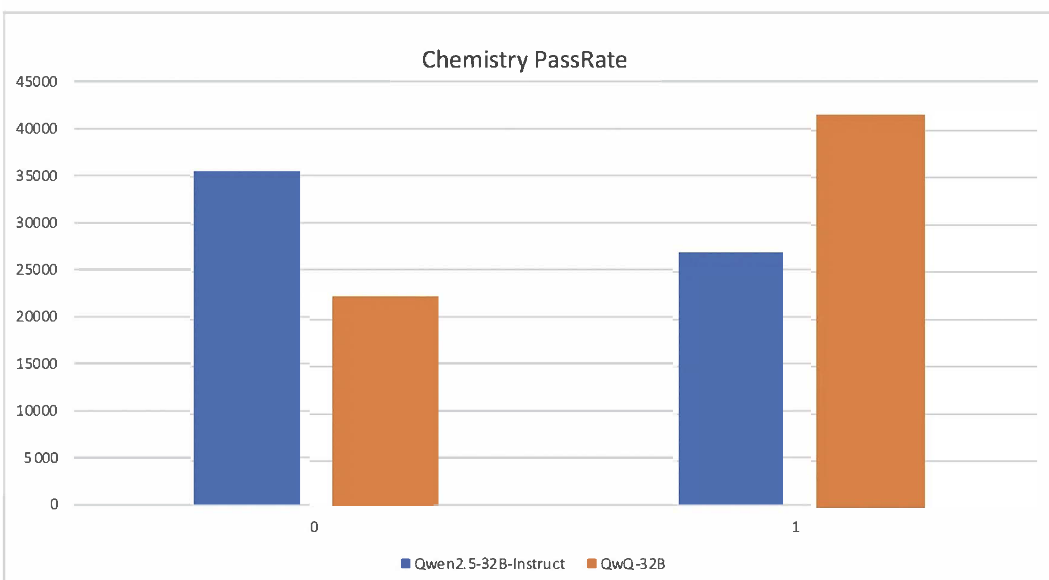

Figure 27: The passrate distribution of chemistry problems generated by the **SHARP** approach.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract states that this paper introduces **SHARP**, a unified approach for synthesizing high-quality reasoning problems for LRMs reinforcement learning with verifiable rewards (RLVR). It claims **SHARP** encompasses self-alignment principles and a three-phase framework. The abstract also claims experiments demonstrate **SHARP**-augmented training substantially outperforms existing methods. The introduction reiterates these points, highlighting **SHARP**'s aim to overcome limitations in generating complex STEM reasoning problems and its main components: the **SHARP** strategy, framework, and implementation. These claims appear to be consistent with the detailed descriptions of the **SHARP** strategy, framework, implementation, and experimental results presented.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper identifies several limitations in the end. Future work could explore applying this approach to other domains and more complex reasoning tasks, and further optimizing the **SHARP** approach on various larger-scale RL reasoning foundation models, designing a reward function that weights principles from the **SHARP** strategy and diving into the distinctions among different subjects, etc. Besides, this paper acknowledges a marginal decrease in GPQA Chemistry performance for the RL-Zero model and attributes it to the nature of chemistry problems and the limitations of unsupervised RL Zero methods without pre-distilled domain-specific priors.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper introduces a new approach (**SHARP**) and a framework, supported by experimental results. It does not appear to present new theoretical results in the form of theorems or mathematical proofs that would require a separate section for assumptions and proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the training data, including the baseline dataset and the **SHARP**-generated dataset of 190,000 samples. It specifies the comparison models used for distillation and RL Zero training. It mentions that training details for distillation involved standard procedures, and for **SHARP**-RL Zero training, the GRPO algorithm was used with a rule-based reward function, with hyperparameters and computational resources detailed in the appendix. The evaluation metrics are centered on the GPQA STEM reasoning benchmark using accuracy metrics like pass@k. The appendix also provides further details, including ablation studies and analysis of the generated dataset's distribution.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed

instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Essential **SHARP** strategy prompts template and specific problems for different subjects are included in the appendix. Moreover, we will open-source all necessary codes and related data for industry use during the review period.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the training data used (baseline CoT samples and 190,000 SHARP-generated samples). It names the models used for comparison. For SHARP-RL

Zero training, the GRPO algorithm and a rule-based reward function hyperparameters, and computational resources are detailed in the appendix. The evaluation benchmark (GPQA) and metrics (accuracy, pass@k) are also clearly stated. The appendix further details some of these aspects.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The tables presenting the main experimental results (Table 1 and Table 2) show performance scores (e.g., GPQA Diamond scores, scores for Physics, Chemistry, Biology) but do not include error bars, confidence intervals, or mention statistical significance tests. If needed, we will include them in the camera-ready version.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The computational resources are detailed in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research involves algorithmic development and evaluation on standard optimization benchmarks. It does not involve human subjects or obviously ethically sensitive applications, and we assume it conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper focuses on the technical contributions of the **SHARP** approach in enhancing LRM reasoning capabilities, particularly in STEM domains. It discusses the potential to push LRM performance closer to expert-level proficiency and superintelligence in STEM. However, it does not contain a dedicated section or explicit discussion on broader societal impacts, either positive or negative, beyond the advancement of AI reasoning capabilities.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper focuses on generating challenging STEM problems and does not explicitly state it is releasing a pre-trained language model or a dataset scraped from sources that would pose a high risk for misuse in the sense described by the guidelines (e.g., generating deepfakes). The generated data consists of STEM problems. While advanced AI models could have dual-use potential, the paper does not discuss releasing models or data in a way that would necessitate specific safeguards as outlined.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: The paper properly cites the sources for existing assets like baseline models (e.g., DeepSeek R1, Qwen models) and benchmarks like GPQA. However, the specific licenses and terms of use for these assets are not explicitly mentioned in the paper text or the appendix.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The main new asset introduced is the **SHARP** methodology and the dataset of 190,000 STEM problems generated using this methodology. The paper provides extensive documentation on the **SHARP** strategy (Algorithm 1), the **SHARP** framework (Alignment, Instantiation, Inference phases, Three-Tier Category knowledge structure), and the **SHARP** implementation. Appendix B provides a detailed analysis of the 190,000 **SHARP**-generated samples, including distributions across STEM subcategories and pass rate analyses. This constitutes detailed documentation of the new asset (the problem generation methodology and the resulting dataset characteristics).

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not describe any crowdsourcing experiments or research involving human subjects as participants in studies. The process involves using LLMs to generate and verify problems, and then training other LLMs.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: As the paper does not involve research with human subjects, IRB approval is not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The core methodology of **SHARP** heavily involves the use of Large Reasoning Models (LRMs). The paper states, "We implement SHARP by leveraging a state-of-the-art LRM to infer and verify challenging STEM questions" and "Our proposed **SHARP** approach aims to systematically generate high-quality, complex STEM reasoning samples by guiding a state-of-the-art LRM (such as DeepSeek R1) instance-alignment reasoning inference through the **SHARP** framework". The use of LRMs is central to the problem generation and refinement process, making it an important and original component of the research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.