

AsyncFlow: An Asynchronous Streaming RL Framework for Efficient LLM Post-Training

Zhenyu Han* †, Ansheng You* ‡, Haibo Wang†, Kui Luo†, Guang Yang†, Wenqi Shi‡, Menglong Chen†, Sicheng Zhang†, Zeshun Lan†, Chunshi Deng†, Huazhong Ji†, Wenjie Liu†, Yu Huang†, Yixiang Zhang†, Chenyi Pan†, Jing Wang†, Xin Huang†, Chunsheng Li†, Jianping Wu† §

† Huawei

‡ Individual Researcher

wujianping5@huawei.com

Abstract

Reinforcement learning (RL) has become a pivotal technology in the post-training phase of large language models (LLMs). Traditional task-located RL frameworks suffer from significant scalability bottlenecks, while task-separated RL frameworks face challenges in complex dataflows and the corresponding resource idling and workload imbalance. Moreover, most existing frameworks are tightly coupled with LLM training or inference engines, making it difficult to support custom-designed engines. To address these challenges, we propose AsyncFlow, an asynchronous streaming RL framework for efficient post-training. Specifically, we introduce a distributed data storage and transfer module that provides a unified data management and fine-grained scheduling capability in a fully streamed manner. This architecture inherently facilitates automated pipeline overlapping among RL tasks and dynamic load balancing. Moreover, we propose a producer-consumer-based asynchronous workflow engineered to minimize computational idleness by strategically deferring parameter update process within staleness thresholds. Finally, the core capability of AsyncFlow is architecturally decoupled from underlying training and inference engines and encapsulated by service-oriented user interfaces, offering a modular and customizable user experience. Extensive experiments demonstrate an average of 1.59× throughput improvement compared with state-of-the-art baseline. The presented architecture in this work provides actionable insights for next-generation RL training system designs.

1 Introduction

Large language models (LLMs) have demonstrated remarkable linguistic capabilities through unsupervised next-token prediction trained on massive nature language corpora [19], paving the way for the pursuit of artificial general intelligence (AGI). Fueled by readily accessible web-scale nature language resources, model parameters have grown exponentially from millions [1, 30] to trillions [3, 9], accompanied by continued improvement of intelligence. However, as the

cornerstone of large language model development, these scaling laws [7] have encountered critical data constraints as pre-training corpora are approaching exhaustion [12, 31]. Estimates indicate that the effective stock of publicly available text data will be depleted within three years (by 2028) [31], threatening to halt the progress of pre-training scaling.

To enhance LLM capabilities, instruction tuning [33] and reinforcement learning from human feedback (RLHF) [17] are introduced to align models with human preferences and societal values, termed as post-training process [37]. Beyond alignment, reinforcement learning (RL) also offers a promising pathway to overcoming data scarcity by leveraging self-generated high-quality responses and reward signals as data flywheels [14], enabling iterative performance improvements. The emergence of reasoning models [4, 16, 29], which exhibit human-level reasoning capabilities through RL-driven optimization, unveiling the post-training scaling laws [25]. Different from the pre-training process that only incorporates a single LLM, RL workflows involve several models and tasks that pose significant challenges for training system design. For instance, the widely used Proximal Policy Optimization (PPO) algorithm [23] in LLM post-training incorporates six distinct tasks: *actor rollout*, *reference inference*, *critic inference*, *reward inference*, *actor update*, and *critic update*. The data interdependence along with complex parallelism strategies making it hard to design efficient and scalable RL post-training systems.

Existing RL post-training frameworks can be categorized as **task-located** or **task-separated**, depending on the task placement strategies. A typical **task-located** framework (e.g., DeepSpeed-Chat [35]) assigns all the tasks to the same set of devices. During training, only one task runs at a time, occupying all computational resources until completion. This paradigm suffers from several critical drawbacks. *Memory inefficiency*. Storing all model parameters in GPU memory introduces excessive memory overhead, either limiting training efficiency or incurring costly memory offloading overhead. *Resharding overhead*. Actor rollout and actor update require distinct parallelism strategies. Transitioning between these tasks necessitates model resharding, introducing significant latency. *Suboptimal Parallelism*. When handling models of varying sizes, smaller models may fail to

* These authors contribute to this research equally.

§ Corresponding authors.

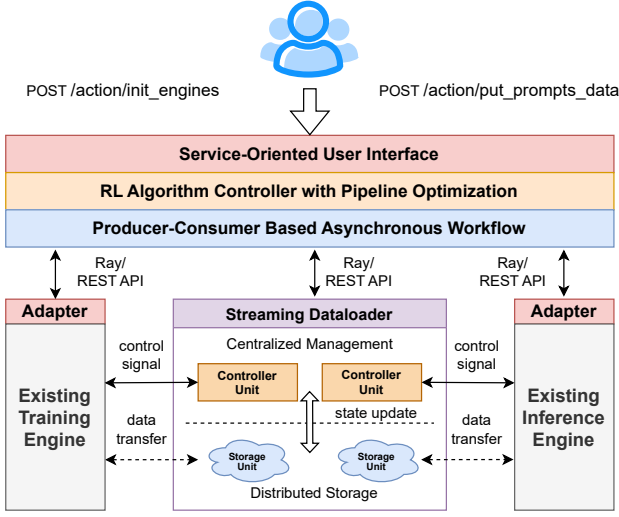


Figure 1. System overview of AsyncFlow framework.

fully utilize all the computational resources due to scaling inefficiencies.

While for **task-separated** frameworks, distinct RL tasks are assigned to different resource pools proportionally sized to their workload demands. In typical task-separated frameworks (e.g., OpenRLHF [5]), distributed resource management and execution tools (e.g., Ray [11]) are introduced to schedule RL workflows. However, inherent data dependencies in RL algorithms restrict full parallelization: tasks such as *reference forward* and *critic update* must wait for *actor rollout* completion, causing prolonged resource idling that delays post-training process. Compounding this challenge, the nested parallelism strategies of LLMs complicate cross-task dataflow between minimal instances, further degrading computational efficiency and bringing extra difficulty in algorithm development.

Another critical obstacle lies in framework usability. Most existing RL frameworks are tightly coupled with specific training and inference engines. For example, NeMo-Aligner [26] can only operate on Megatron-LM [28] and TensorRT-LLM [15], while OpenRLHF requires DeepSpeed [21] and vLLM [8] as mandatory backends. This design paradigm significantly restricts flexibility and adaptability, particularly for industrial users who often rely on pre-existing training and inference clusters built on customized backends. Ideally, a RL framework should provide high-level abstractions that seamlessly integrate with heterogeneous training and inference engines, thereby accommodating diverse customization demands. Furthermore, the absence of a unified algorithm controller exacerbates developing overhead and impedes academic research workflows that require frequent algorithm modifications.

To address these challenges, we propose AsyncFlow, an asynchronous RL framework with task-separated architecture, built atop the Ascend-powered post-training framework MindSpeed-RL [6]. Specifically, we design a distributed data storage and transfer module named TransferQueue that optimizes dataflow across instances. It collects all the responses produced by the inference engine, and dynamically dispatches these samples to instances of downstream RL tasks upon requests, enabling a streamed dataflow control. It eliminates the need to explicitly define cross-instance data dependency chains, thereby enabling automated load-balancing and pipeline overlapping among RL tasks. Furthermore, to fully unlock the architectural potential of task-separated framework, we propose a producer-consumer-based asynchronous workflow optimization algorithm. By deferring the parameter update process for rollout instances, we achieve a stable asynchronous workflow that minimizes the idle time with controllable staleness. Combined with dynamic load-balancing strategies, the efficiency of post-training can be significantly enhanced. Moreover, AsyncFlow serves as a high-level scheduling layer of RL algorithms, encapsulated by minimal service-oriented interfaces to suit various backend engines. This design principle enables an efficient and customizable user experience, which would reconcile research flexibility with industrial deployment requirements, bridging the gap between theoretical exploration and industrial applications. We will integrate the proposed optimization into MindSpeed-RL, which is publicly accessible at <https://gitee.com/ascend/MindSpeed-RL>.

Our contributions are summarized as follows:

- We design a distributed data management module that enables automated load-balancing and pipeline overlapping among RL tasks, which greatly improves the training efficiency of task-separated RL framework.
- We propose a producer-consumer-based asynchronous workflow that can well-balancing the training efficiency and the convergence of RL algorithm.
- We present a user-friendly service-oriented API that supports various training/inference backend engines, bridging the gap between academic research requirements and industrial deployment scalability.
- We conduct extensive experiments against the state-of-the-art RL post-training framework, and observe a maximum 2.03× throughput improvement over the baseline in large-scale clusters.

2 System Overview

We present the architecture design of AsyncFlow, an asynchronous streaming reinforcement learning framework for scalable post-training in Fig.2. As its foundation, the resource layer leverages Ray to manage computing resources, with hardware allocation pre-optimized through an execution time simulator to ensure efficient training. Building upon

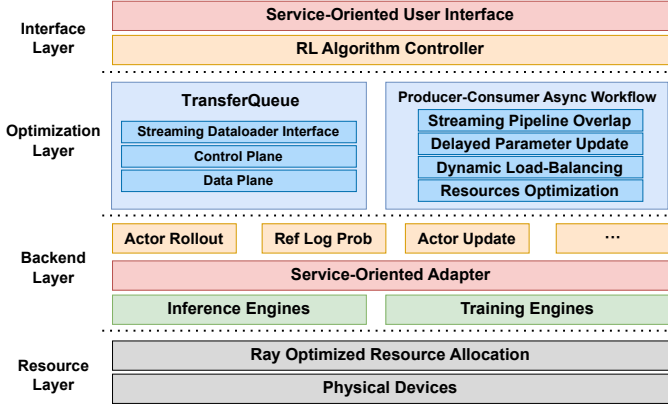


Figure 2. Hierarchical architecture design of AsyncFlow framework.

this, the backend layer provides modular adapters for varying heterogeneous training and inference engines. Therefore, RL tasks can be instantiated on these backend engines while maintaining engine-agnostic. In the optimization layer, we try to address two critical challenges in dataflow management and resource utilization for task-separated frameworks. We introduce TransferQueue as a streaming dataloader to dynamically schedule complex dataflows across diverse parallelism strategies of RL tasks. Besides, we propose a producer-consumer-based asynchronous workflow to achieve high computational resource utilization. Combined with streaming pipeline overlapping and delayed parameter update mechanism, we can significantly reduce the idle time in task-separated frameworks. For usability design, the interface layer provides a unified algorithm entry point. It serves as a single controller that satisfies the requirement of algorithm research. Complementing this, AsyncFlow provides a set of service-oriented API for industrial workflows, which enables seamless integration with existing training and inference clusters. In summary, AsyncFlow aims to bridge the gap between algorithmic research and industrial deployment in LLM post-training by delivering a unified framework that harmonizes flexibility with scalable efficiency.

3 TransferQueue: High-Performance Asynchronous Streaming Dataloader

During the RL post-training process, data dependencies among tasks pose a significant challenge for task-separated framework design. Existing implementations only provide basic data storage and transfer capabilities for the entire training dataset, leading to substantial resource idling for downstream tasks. In AsyncFlow, we introduce TransferQueue—a centralized data management module with distributed storage capability that serves as an asynchronous streaming

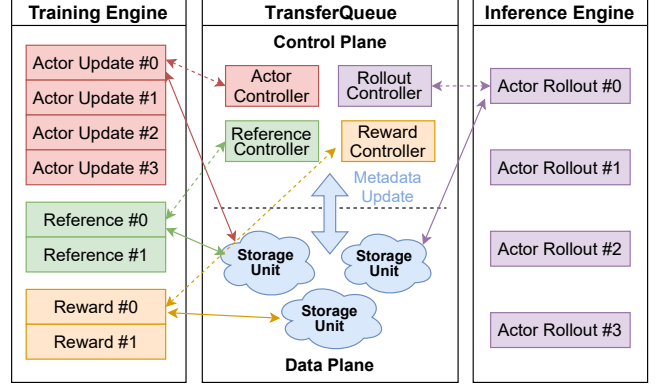


Figure 3. Architecture design of TransferQueue. Each DP group interacts with its corresponding TransferQueue controller to get the metadata upon request, then execute the read/write operation with storage units in the data plane. The dashed line represents metadata communication, while the solid line depicts the communication process with real data. All these interaction processes are encapsulated as a distributed streaming dataloader to seamlessly integrate into training and inference engines.

dataloader. In task-separated RL framework, this design facilitates streaming pipeline overlapping by enabling downstream tasks access part of training samples for computation, rather than waiting for the entire dataset to be ready.

For data management capability, we aim to provide a centralized view of data status for each RL task. This feature eliminates the need to manually define all the dataflows between DP groups across several RL tasks. This design differs us from existing solution such as OpenRLHF [5] and StreamRL [38], providing a more flexible and efficient programming paradigm. This centralized view enables better load-balancing policies.

For data storage and transfer, we aim to support high-concurrency, asynchronous requests in large-scale post-training. Inspired by software-defined networking (SDN), we decouple the data plane from the control plane, and instantiate multiple controllers and data storage objects in each plane. This design alleviates potential I/O and network bottlenecks and supports varying storage systems, enabling scalable post-training.

3.1 Architecture Overview

As illustrated in Fig.3, TransferQueue acts as a streaming data scheduler bridging the training and inference clusters, managing the entire dataflow in RL post-training process. Specifically, each RL task is equipped with a dedicated TransferQueue controller, which maintains metadatas for training samples. Specifically, the metadata includes storage location, data status and consumption status. These controllers operate independently, as RL tasks inherently avoid algorithmic

interference with each other. For data storage, each unit is responsible for a subset of samples within current global batch. These samples are assigned global indices to ensure accurate addressing by all the distributed controllers. When new data are written to a storage unit, it triggers an update notification to all controllers to update their metadata. This mechanism allows each RL task to dynamically access newly available data upon request.

When a DP group requires new data, it initiates the process by sending a read request to its corresponding controller. The controller dynamically assembles a batch of samples from currently available data and returns their metadata to the requester. The DP group then retrieves the corresponding data from distributed storage units using the provided metadata. To ensure each DP group has access to distinct samples without duplication, the controller tracks consumption records of each sample, guaranteeing that only one DP group within the RL task accesses the sample’s metadata. Similarly, after a DP group completes the computation, it writes results back to the storage units atomically through metadata to guarantee consistency across distributed components.

To ensure compatibility with the input formats of training and inference engines, we encapsulate the aforementioned interaction logic into a PyTorch DataLoader. This allows users to seamlessly integrate TransferQueue as a distributed streaming dataloader, eliminating the need to understand the underlying implementation details.

3.2 Data Plane: Distributed Data Transfer and Storage

3.2.1 Data Structure. To address the diverse data requirements of RL tasks, we adopt a 2D columnar data structure, which is illustrated in Fig.4. In this design, columns correspond to task-specific data components, such as actor responses and reference log probabilities. Rows represent complete training samples, each can be uniquely addressed by a global index. This structure enables RL tasks to retrieve only the specific data components by algorithm requirements. For instance, the reference model needing input token IDs and responses can exclusively request the relevant columns, minimizing unnecessary data transfer. Additionally, this organization supports concurrent read/write operations across distinct positions, enhancing scalability in high-concurrency scenarios.

Considering the potential I/O and bandwidth bottlenecks, we store the training samples in a distributed manner. Each of the storage unit maintains a subset of data entries (i.e., rows) to amortizing storage and communication overhead across the system.

3.2.2 Metadata Notification. When new data has been written to a data storage unit, it triggers a notification to controllers to update their metadata. As demonstrated in Fig.5, upon completion of the writing process, data storage units



Figure 4. Data structure of TransferQueue. Each row represents a complete data sample, while each column represents the task-specific data component. Through the global index, we can accurately address the data samples across different storage units for concurrent read/write operations.

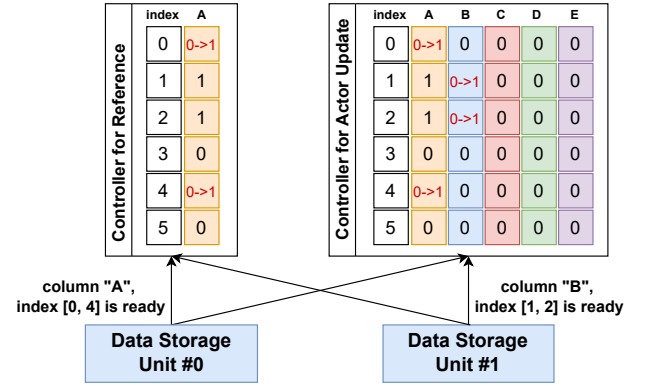


Figure 5. Metadata notification process of TransferQueue. Data storage units will broadcast the metadata (including the global index and corresponding data columns) to all the controllers.

broadcast the corresponding row indices and column identifiers to all the controllers (registered during initialization). This mechanism ensures controllers are immediately aware that the data at these locations is ready for consumption.

3.3 Control Plane: Centralized View of Data Management

The nested parallelism strategies inherent in LLM training, combined with multiple concurrent RL tasks lead to complicated dataflows during post-training. To disentangle these complex data dependencies, we leverage the control plane of TransferQueue as a centralized data management module, ensuring coherent coordination across distributed workers for each task.

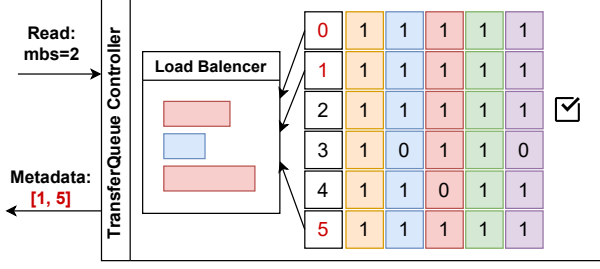


Figure 6. Scheduling process of TransferQueue controller. Red indexes demote that the corresponding data samples satisfies the RL task requirement, and the checkmark means the corresponding data has been consumed in earlier requests.

We initialize distinct TransferQueue controllers for each RL task. The controller maintains the data status metadata scoped to its corresponding task, containing all entries of the required columns. The data status metadata uses binary status indicators: status 0 indicates data unavailability, while status 1 denotes the data is ready for retrieval. As described in Section 3.2.2, these metadata entries are dynamically updated whenever new data is written to the data storage unit.

As illustrated in Fig. 6, upon receiving a read request, TransferQueue controller scans the metadata to identify entries that satisfy the RL task’s requirements—specifically, entries where all column statuses equal to 1 and no prior consumption records by other DP groups. If the currently available data exceeds the requested micro-batch size, the controller selects and packs the metadata into a micro-batch according to a load-balancing policy. These samples are then marked as consumed to prevent duplication by other DP groups within the same RL task. Leveraging these metadata, the consumer communicates with distributed data storage units to execute the actual data retrieval process.

A key advantage of our architecture lies in its centralized data management. By dynamically scheduling all available data through the controller rather than pre-allocating them to DP groups, TransferQueue enables advanced load-balancing strategies. For instance, faster instances (due to hardware heterogeneity or shorter response lengths) can dynamically request more data, improving overall efficiency. Furthermore, proactive load-balancing can be implemented to ensure equitable distribution of processed tokens across DP groups, thereby minimizing computational idling for *actor update* process. In summary, TransferQueue offers critical insights into the development of next-generation dataflow management systems.

3.4 Interaction Interface

To simplify the usage of TransferQueue, we have encapsulated its capabilities into a PyTorch DataLoader. As demonstrated in Code 1, users first initialize TransferQueue as a

streaming dataloader by specifying the current RL task, required input data columns, and micro-batch size. During each forward step, the required data can then be easily retrieved via an iterator wrapper. This interface simplifies TransferQueue’s integration, ensuring seamless compatibility with existing training workflows.

```

1 # In class RolloutWorker(BaseWorker):
2 def generate_sequences(self):
3     # define the data columns and initialize the
4     # TransferQueue
5     experience_consumer_stage = 'actor_rollout'
6     experience_columns = ['prompts', '
7     prompt_length']
8     experience_count = self.rl_config.
9     rollout_dispatch_size
10
11     data_loader = self.create_stream_data_loader(
12         experience_consumer_stage=
13         experience_consumer_stage,
14         experience_columns=experience_columns,
15         experience_count=experience_count,
16         use_vllm=True,
17         pad_to_multiple_of=self.generate_config.
18         infer_tensor_parallel_size,
19     )
20     data_iter = iter(data_loader)
21
22     for batch_data, index in data_iter:
23         # do inference
24         prompts_data = batch_data['prompts']
25         responses = self.rollout.
26         generate_sequences(prompts_data)
27         # write generated responses to
28         TransferQueue
29         self.collect_transfer_queue_data(responses
30         , index)
    
```

Code 1. TransferQueue usage example

3.5 High-Concurrency Design

TransferQueue is designed to address large-scale post-training challenges through three steps. First, its decoupled control plane and data plane architecture enables scalable data storage unit expansion for high-concurrency operations. When I/O or network bandwidth bottlenecks emerge, additional data storage units can be easily initialized to expand total bandwidth and reduce system latency. This design also supports easily switch to other storage backends in the future, such as Mooncake Store [18], Redis [22] or other storage systems tailored to LLM training. Furthermore, the scheduling process in the control plane and I/O operations in the data plane execute concurrently, creating a pipelined workflow that efficiently handles multiple incoming requests. Second,

TransferQueue optimizes communication efficiency by designating a single rank within each DP group to interface with the system, considering the fact that each rank in a DP group should receive the same data (when not considering sequence parallelism). Retrieved data is then broadcast to other ranks within the DP group. This approach significantly reduces the volume of direct requests to TransferQueue in large-scale post-training scenarios. Third, we eliminate unnecessary padding during data storage and transmission. TransferQueue inherently supports variable-length data transfers. For device-to-device Huawei Collective Communication Library (HCCL) communication, tensors are concatenated along the sequence dimension for broadcast operation, then we restore the received tensor using the length metadata. This strategy minimizes redundant communication overhead caused by padding, particularly in settings with large micro-batch sizes. Collectively, these designs alleviate data storage and transfer bottlenecks, ensuring robust high-concurrency performance for large-scale post-training workloads.

4 Producer-Consumer-Based Asynchronous Workflow Optimization

A critical challenge in task-separated framework stems from pipeline bubbles—idle hardware resources caused by data dependencies for tasks that are deployed across different sets of devices. To overcome this challenge, AsyncFlow implements a series of pipeline optimization techniques to maximize hardware utilization.

First, our streaming dataloader enables pipeline overlapping among RL tasks, which greatly reduces the end-to-end execution time of post-training. Second, in off-policy scenarios, we design a delayed parameter update mechanism that well balances algorithmic convergence and training efficiency. By allowing a continuous one-step asynchronization between *actor rollout* and *actor update*, this approach effectively eliminates the bubbles by minimizing warm-up and cool-down phases. Finally, fine-grained execution time optimization is achieved through proactive resource planning and dynamic load-balancing strategies, fully unlocking the scalability of task-separated frameworks in LLM post-training.

4.1 Streaming Pipeline Overlapping across RL Tasks

Leveraging TransferQueue, AsyncFlow not only simplifies the dataflow management but also enhances training throughput by enabling pipeline overlapping across RL tasks. In this paradigm, all the training and inference instances only need to interact with the streaming dataloader, which dynamically schedules and redirects the finest-grained data samples across tasks. This architecture inherently supports pipeline overlapping, which is demonstrated in Fig.7. Different from the design in existing frameworks [5, 38], we can easily extend this overlapping strategy to any RL algorithms with

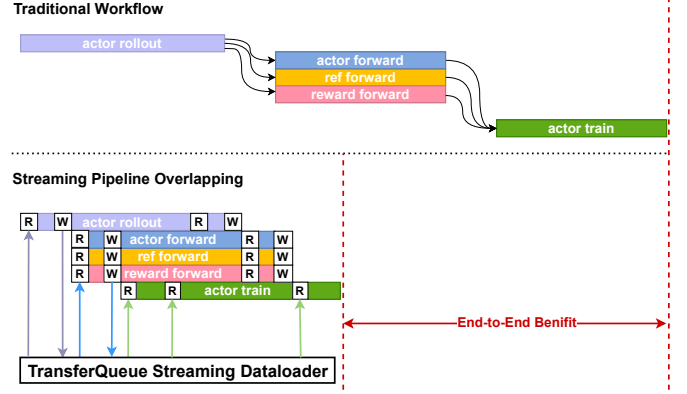


Figure 7. Streaming pipeline overlapping.

varying tasks, eliminating the need to manually reschedule the data streams.

4.2 Asynchronous Off-Policy Bubble Reduction

4.2.1 Basic asynchronous RL algorithm. Traditional RL algorithms for LLM post-training predominantly adopt a synchronous on-policy paradigm, where *actor rollout* and *actor update* operate on identical parameter states. While this ensures algorithmic convergence, its computational inefficiency—stemming from strict synchronization requirements—severely limits scalability for large-scale post-training. Recent advancements, however, have challenged this paradigm: optimization techniques such as partial rollout [29] and streaming rollout [24] practically relax the on-policy assumption by allowing response data generated by older models used in the update stage. These innovations create opportunities to explore asynchronous, off-policy frameworks, which promise significant gains in training efficiency by pipeline overlapping while maintaining convergence guarantees.

As illustrated in Fig.8(a), on-policy algorithms enforce strict synchronization between *actor rollout* and *actor update* stages, which leads to warm-up and cool-down bubbles across iterations. A straightforward mitigation strategy involves enlarging the global batch size of *actor rollout* stage, transitioning to an asynchronous off-policy scenario where *actor update* utilizes responses generated by stale parameters, as illustrated in Fig.8(b). This adaption extends the stable phase of the RL pipeline, thereby reducing the proportion of pipeline bubbles in warm-up and cool-down phases. However, algorithmic convergence constraints impose strict limits on the allowable version differences between *actor rollout* and *actor update* stages. Empirical studies demonstrate that one-step asynchronization between *actor rollout* and *actor update* will not lead to significant performance or convergence degradation [13, 38]. While the performance drops logarithmically with the increasing version differences beyond this threshold. This trade-off highlights the critical

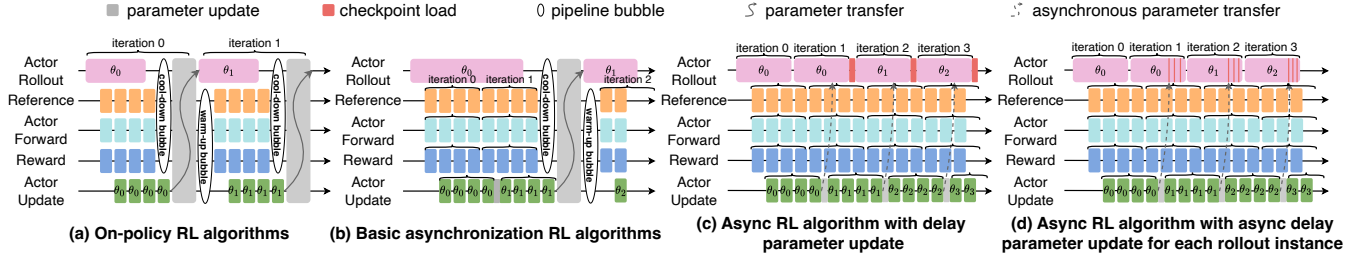


Figure 8. Illustration of the asynchronous off-policy RL workflow. Leveraging the proposed delay parameter update mechanism in (c), AsyncFlow can extend the steady phase of the RL pipeline to further reduce pipeline bubbles. By sequentially enabling parameter updates for rollout instances in (d), we can achieve a sub-step asynchrony.

need for mechanisms that reconcile training efficiency with algorithmic stability in large-scale LLM post-training.

4.2.2 Delayed parameter update mechanism. To solve the above dilemma, we present a delayed parameter update mechanism that further eliminates pipeline bubbles by decoupling the model weight of *actor rollout* from *actor update*. As illustrated in Fig.8(c), this mechanism defers parameter update by one step, enabling continuous rollout during the transition period across iterations. Specifically, the actor rollout worker does not immediately halt generation when the actor update is completed. Instead, it continues generating responses using the old model weights while asynchronously writing the received new parameters to the host memory. The new parameters will load to the Ascend Neural Network Processing Units (NPU) when the current generation iteration is completed, reducing the exposed synchronization overhead to a relatively fast host-to-device (H2D) transmission. By allowing a continuous one-step asynchronization, it enables the stable phase to be nearly infinitely extendable by eliminating warm-up and cool-down phases. While conceptually similar to StreamRL [38], our approach further enhances scalability by overlapping all RL tasks within the training engine through the centralized dataflow management of TransferQueue. In summary, this design effectively reduces the warm-up and cool-down pipeline bubbles in RL post-training workflows, fully exploiting the architectural potential of task-separated frameworks.

Furthermore, we propose a novel parameter updating mechanism that enables a sub-step asynchronous algorithm workflow, as illustrated in Fig.8(d). To efficiently supply data for downstream tasks, we usually allocate abundant hardware resources to the *actor rollout* task. This setup allows rollout instances to perform parameter updates sequentially, while the remaining instances continue fulfilling the data requirements of downstream training tasks. Consequently, in each iteration, we can leverage the most recently updated parameters to generate part of data, achieving sub-step asynchrony. Additionally, this design also minimizes the overhead of checkpoint loading, thereby enhancing system efficiency.

We leave the implementation details of this mechanism for an important future work.

4.2.3 Parameter update overlapping. The parameter update module is mainly composed of a *WeightSender* deployed on the training cluster and a *WeightReceiver* on the inference cluster. To minimize the weight transmission time in the RL training pipeline, the sender and receiver are designed to support both the synchronous and asynchronous modes.

In synchronous mode, *actor rollout* is blocked by the parameter update process. To shorten the exposed transmission time, we leverage high-bandwidth HCCL links across Ascend NPUs to transmit the model weights. To further reduce the end-to-end time, we further develop an asynchronous parameter update process that can fully overlap with computation tasks for asynchronous RL algorithms. Specifically, model weights from the training engine are offloaded to the host device and asynchronously transmitted to the inference engine over the host network, decoupling computational workloads from the weight synchronization process. This design ensures that the parameter update process neither stalls nor interferes with ongoing computational tasks, maintaining a continuous RL workflow.

4.3 Task Resource Planning

To achieve the ideal workflow illustrated in Fig.8(c), we build a graph-based resource planning module that accurately searches the optimal configuration under given resource constraints. Through simulating the computational and communication time under varying configurations, we can acquire the resource allocation setting and hyper-parameters that minimize the end-to-end time of the whole RL workflow.

Given the large search space for LLM post-training, we adopt a hybrid cost model combining analytical-based method and profiling-based method. The analytical-based method estimates the execution time leveraging the hardware specifications and theoretical computation and communication volumes, offering a fast evaluation that can quickly narrow down the search space. In contrast, the profiling-based

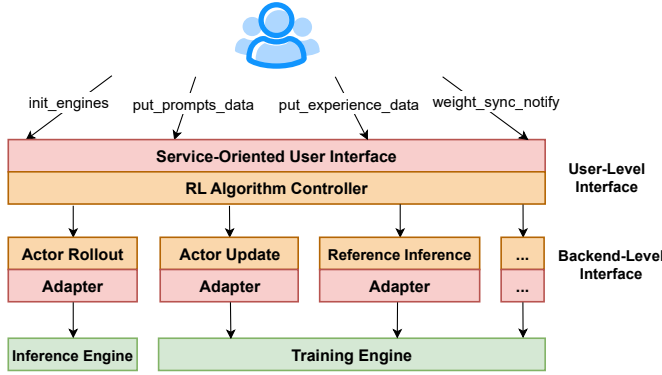


Figure 9. Architecture of service-oriented user interface.

method provides block-level performance by actually running training and inference tasks. It provides an accurate evaluation with the cost of a much higher time consumption. Leveraging the hybrid cost model, we ensure a good balance between efficiency and accuracy, making it well-suited for optimizing the complex RL workflows for LLM post-training.

5 Service-Oriented User Interface

To deliver an enhanced user experience, RL frameworks should provide a high-level abstraction layer that orchestrates all the algorithm-specific tasks. For academic research, this design offers a unified entry point for algorithm development, enabling rapid experimentation through standardized APIs. For industrial scenarios, it is critical to support various training and inference backends while maintaining architectural flexibility. Therefore, maintaining a proper user interface is of critical importance.

In AsyncFlow, we implement a hierarchical service-oriented interface as shown in Fig.9. Specifically, the user-level interface encapsulates RL algorithm logics for end-users, exposing a set of key APIs to control the post-training workflow. Besides, the backend-level interface offers a modular abstraction of RL tasks, decoupling the algorithm logic from execution engines through several backend adapters. The above design achieves a clear separation of concerns: algorithm researchers can interact with high-level APIs for algorithm designs, while infrastructure engineers can focus on low-level implementation to achieve a higher efficiency. The synergy of these interfaces well-balance academic flexibility with industrial scalability, ensuring AsyncFlow serves both rapid algorithm iteration and production-grade deployment demands.

5.1 User-Level Interface

AsyncFlow provides an RL algorithm controller within the *Trainer* class, serving as the centralized entry point for the main training workflow. This abstraction seamlessly organizes critical RL tasks, such as *generate_sequences* and

```

1 # Base adapter
2 class RLAdapter:
3     pass
4
5 class MindSpeedAdapter(RLAdapter):
6     def __init__(self, forward_backward_func,
7                 model, batches, forward_step):
8         self.forward_backward_func =
9             forward_backward_func
10        self.forward_step = forward_step
11        self.model = model
12        self.batches = batches
13        ...
14
15 # Abstraction for RL task: compute_log_prob
16 def compute_log_prob(self):
17     ...
18     losses_reduced = self.
19     forward_backward_func(
20         forward_step_func=self.forward_step,
21         data_iterator=iter(self.batches),
22         model=self.model,
23         micro_batch_size=self.micro_batch_size,
24         forward_only=True,
25         collect_non_loss_data=True,
26     )
27     ...
28
29 class VLLMAdapter(RLAdapter):
30     pass

```

Code 2. Backend-level interface.

update. Researchers can easily modify the core RL algorithms through the *Trainer* class. To facilitate workflow automation in industrial scenarios, we implement several key APIs for initiating the full post-training task with minimal configuration, such as:

- *init_engines*: Initialize the training and inference engines.
- *put_prompts_data*: Load the prompt dataset to the post-training system.
- *put_experience_data* and *get_experience_data*: Coordinate the generated experience data between training and inference engines.
- *weight_sync_notify*: Notify training and inference engine to update the model weights.

The above user-level interface offers intuitive control over the post-training workflow, thereby streamlining system utilization.

5.2 Backend-Level Interface

Considering the varying training and inference backend engines, we design a low-level abstraction of RL tasks through the *Adapter* class, which is demonstrated in Code. 2

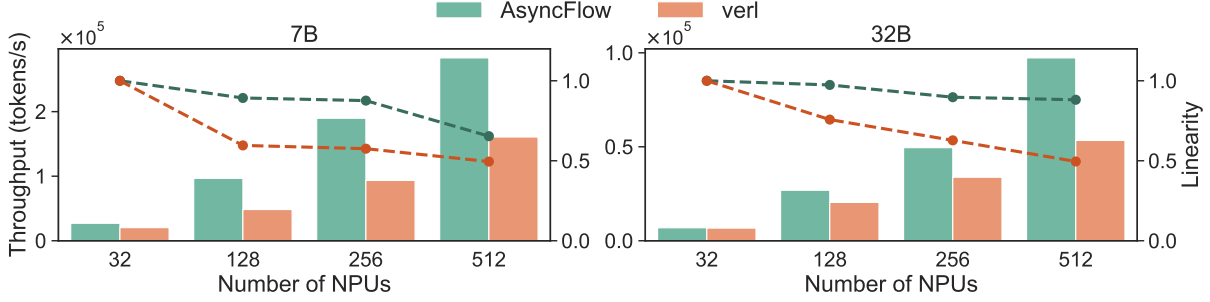


Figure 10. End-to-end throughput and scalability analysis across varying cluster and model sizes. We report the average throughput across 10-20 iterations to reduce measurement fluctuations at the beginning.

This layer ensures seamless integration of diverse training/inference backends (e.g., FSDP, DeepSpeed, vLLM) while maintaining compatibility with custom-designed frameworks, allowing engineers to optimize hardware utilization or switch systems without disrupting workflow integrity.

6 Evaluation

6.1 Experiment Setup

Models. During the evaluation of AsyncFlow framework, we choose Qwen2.5 series of models [34], ranging from 7B to 32B parameters. Qwen2.5 is widely used for both academic research and industrial workflows, serving as a common base model for RL post-training evaluation [38].

RL algorithms. We implement Group Relative Policy Optimization (GRPO) [20] as a representative RL algorithm for evaluation. GRPO eliminates the need for a separate critic model by leveraging multiple responses per prompt to estimate preference signals. This design greatly improves post-training efficiency, as demonstrated by its successful application in DeepSeek-R1 [4]. Support for PPO (Proximal Policy Optimization) [23] is currently under development.

Datasets. We adopt the DeepScaleR dataset [10] for RL post-training. It contains more than 40 thousand mathematics question-answer-solution pairs that carefully compiled from AIME (American Invitational Mathematics Examination), AMC (American Mathematics Competition), etc. Leveraging this dataset, DeepScaleR-1.5B-Preview has surpassed the performance of OpenAI o1-Preview [16] in several benchmarks, demonstrating the effectiveness of the dataset. In recently developed RL post-training frameworks [2], DeepScaleR acts as a common dataset choice.

Hardware configuration and parallelism strategies. We use large-scale Ascend NPU clusters to evaluate the proposed RL framework. Each node has 16 NPUs, with system memory of 2880 GB.

Software versions. We use Ascend Extension for PyTorch 7.0.0 (PyTorch-2.5.1) and Ascend Compute Architecture for Neural Networks (CANN) 8.1.RC1 to serve as the basic software platform for NPU support. In AsyncFlow, we

use vLLM-Ascend 0.7.3 as the inference backend, along with MindSpeed as the training backend.

Baselines. We compare the proposed AsyncFlow with verl [27], which is the state-of-the-art task-collocated RL framework that leveraging an efficient 3D-HybridEngine to reduce the resharding overhead. Besides the high training efficiency, its combination of single-controller and multi-controller paradigm also greatly simplifies the software development. During the experiment, we choose Pytorch FSDP as the training backend and vLLM-ascend 0.7.3 [8] as the inference backend. To run on Ascend NPU platform, we made necessary adaptations to the verl (version dated April 7, 2025, with commit d13434f).

6.2 Overall Performance Analysis

We conduct extensive experiments on clusters scaling from 32 to 1024 NPUs, evaluating the performance of AsyncFlow on Qwen2.5-7B and Qwen2.5-32B models. As shown in Fig.10, AsyncFlow consistently outperforms the verl baseline across all configurations, achieving an average throughput gain of 1.59×. Notably, AsyncFlow demonstrates a significant performance improvement in large-scale clusters, where the peak performance achieves 2.03× for 7B model on 256 NPUs. For 512 NPUs, AsyncFlow still achieves 1.76× and 1.82× of throughput over verl. While even at 32 NPUs, AsyncFlow maintains a 33.4% higher throughput for 7B model, demonstrating robust adaptability to resource-constrained environments. This phenomenon is in line with previous studies, where task-separated frameworks demonstrate a superior potential in large-scale scenarios [27, 38]. Critically, AsyncFlow sustains a high scaling efficiency, maintaining a linearity of 0.65 and 0.88 when cluster size expands 16×—a breakthrough that paves the way for efficient training of LLM reasoning agents at industrial scales.

6.3 Ablation Studies

To validate the effectiveness of the proposed methods, we conduct a series of ablation studies summarized in Table.1.

Baseline. We establish a baseline scenario representing a conventional task-separated RL framework by disabling

Table 1. Performance improvement breakdown for 7B model running on 512 NPUs.

No.	Setting	# Normalized Throughput
①	Baseline	1
②	w/TransferQueue	2.01
③	② + w/Asyn.Opt	2.74

all proposed optimizations. In this scenario, each RL task is allocated to separated hardware resources, and only one task executes at any given time. This sequential workflow is illustrated at the top of Fig.7.

TransferQueue. As a core feature of AsyncFlow, TransferQueue enables fine-grained overlapping across RL tasks. Compared to the baseline, integrating TransferQueue yields a 2.01 \times throughput.

Asynchronous workflow optimization. To minimize the idle periods across iterations, we implement an optimized asynchronous workflow. It includes the delayed parameter update mechanism, overlapping techniques, and task resource allocation strategies. The optimization further improves throughput by 36.3% over the TransferQueue-enabled baseline.

6.4 Optimized Workflow of AsyncFlow

To rigorously evaluate the efficacy of the proposed asynchronous off-policy workflow, we present the empirical execution timeline of distributed training and inference instances in a Gantt chart (Fig. 11). It reveals that RL tasks achieve substantial parallelism under optimized dataflow scheduling, with minimal inter-task idle times. This observation empirically validates that task-separated RL frameworks can strike a good balance between resource utilization and scalability, enabling efficient post-training for large-scale settings.

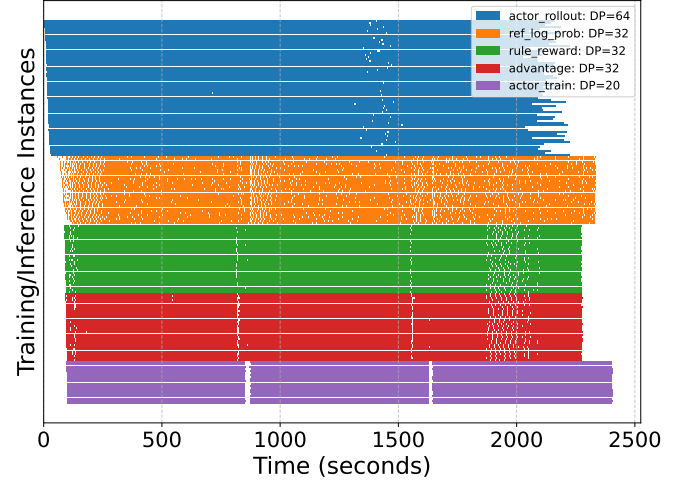
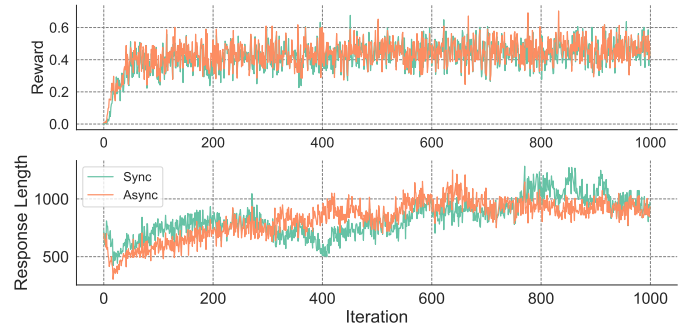
6.5 Stability of Asynchronous RL Algorithm

To evaluate whether the proposed asynchronous RL workflow impacts model performance, we measure the average reward and response length given the same clock-time constraint, which are demonstrated in Fig.12. The experiments are conducted on a 7B model deployed across 16 NPUs, with the asynchronous workflow alternately enabled and disabled. Results demonstrate negligible differences in reward scores, and the variance of response lengths demonstrates a convergence trend.

7 Related Work

7.1 LLM Post-Training Frameworks

As post-training phases have become indispensable for deploying state-of-the-art large language models to refine their capabilities, numerous frameworks have emerged to efficiently manage and optimize this process. In this study, we

**Figure 11.** AsyncFlow workflow of each training and inference instances. We showcase the 32B model with 512 NPUs, with iteration 0-3.**Figure 12.** Reward and average response length comparison for the proposed asynchronous RL workflow and vanilla synchronous RL workflow.

classify post-training frameworks into two paradigms: **task-collocated** and **task-separated**, based on how reinforcement learning tasks map to physical devices.

Task-collocated frameworks execute both *actor rollout* and *actor update* on the same set of devices. Examples include TRL [32], DeepSpeed-Chat [35], NeMo-Aligner [26], RLH-Fuse [39], and verl [27]. These frameworks achieve high resource utilization for small-scale post-training, since all the devices are working on the same computation task at any given time.

Task-separated frameworks, in contrast, decouple *actor rollout* and *actor update* into distinct hardware resources. Recent high-quality frameworks such as OpenRLHF [5], k1.5 [29], Seed1.5-Thinking [24], StreamRL [38], and AReaL [2] demonstrate strong competitiveness in large-scale post-training scenarios. This paradigm aligns with the divergent computational demands of LLM inference (memory-intensive)

and training (compute-intensive), enabling better hardware utilization through specialized resource allocation.

7.2 RLHF Algorithms

Among RLHF algorithms, PPO [23] serves as the foundational algorithm, orchestrating four core components: the actor model, the reference model, the reward model, and the critic model. This multi-component architecture introduces intrinsic computational complexity, posing challenges to scalability and training efficiency. To mitigate these limitations, recent variants such as GRPO [20] and Decoupled Clip and Dynamic sAmpling Policy Optimization (DAPO) [36] simplify the workflow by eliminating specific components. Among these variants, GRPO removes the critic model by approximating advantages through group-relative comparisons between policy and reference outputs, while DAPO further eliminates the reference model via a dynamic reference strategy that leverages historical policy checkpoints as adaptive baselines. These simplifications aim to reduce computational overhead but inherently trade off certain functionalities, such as the stability guarantees provided by explicit value function estimation in PPO.

7.3 LLM Post-Training System Optimization

Considering the end-to-end time composition of the RL post-training workflow, the *actor rollout* phase has drawn significant research attention due to its autoregressive inference nature, which fundamentally distinguishes it from training tasks. RLHFuse [39] introduces inter-stage fusion strategies that enable concurrent execution of *actor rollout* with downstream tasks through an automated migration mechanism. StreamRL [38] further enhances scheduling flexibility by incorporating a response length predictor. Through categorizing prompts into several response length levels, it dynamically dispatches them to different instances, balancing throughput maximization with skewness due to long-tail responses. Similarly, the partial rollout technology is introduced in k1.5 [29], which truncates long responses to enable pipelining of downstream tasks without waiting for the full generation. Notably, the rollout phase in post-training systems lacks strict service-level objectives (SLOs) compared to inference serving, enabling broader scheduling strategies that maximize the system throughput.

8 Conclusion

In this work, we propose a task-separated reinforcement learning framework designed to deliver a modular and scalable post-training capability, especially in large-scale clusters.

To achieve this goal, we first introduce TransferQueue, a streaming dataloader that provides centralized data management with distributed storage capabilities. Through dynamically routing the complex data dependencies across training

and inference instances, it overcomes bottlenecks imposed by static data dependency graphs that are pre-defined before starting the post-training task. Furthermore, we develop an asynchronous workflow based on a producer-consumer abstraction, which balances training efficiency and convergence stability by allowing one-step parameter asynchronization between *actor rollout* and *actor update*. By overlapping parameter updates with computation tasks, this design minimizes hardware idling across iterations while maintaining algorithmic integrity. For usability design, we implement a set of service-oriented interfaces that provide two-level abstraction of both algorithm workflow and the backend engines, which can effectively bridge the gap between theoretical research and industrial deployment. Extensive experiments demonstrate that AsyncFlow achieves significant throughput improvements over state-of-the-art baselines, where we highlight that our task-separated framework exhibits superior linearity when scaling to larger clusters.

For future work, we identify critical opportunities in rollout system design. Under the constraint that the generated responses sufficiently support downstream tasks, we can stagger the timing of parameter updating for each inference instance to eliminate synchronization barriers during response generation. This approach not only reduces the exposed transition time, but also enables sub-step asynchronization workflow where the staleness threshold between *actor rollout* and *actor update* falls below one training step.

In summary, our study sheds light on the design of task-separated RL frameworks, which demonstrate superior scalability and efficiency in large-scale post-training scenarios.

Acknowledgments

We would like to express our sincere gratitude to Jianping Sun, Chao Bai, Long Chen, Benzhe Ning, Xushi Li, Qianqian Cheng, Yayun Ji, Yongwen Li, Wenqin Tan, Yebin Zhang, Peihan Liu, Haoran Dong, Cunle Qian, Xi Chen, Wei Zhou, Yonghao Song, Cangsu Hu, Yinlei Sun, Liangjun Feng, Jiangben Wang and Bo Wang for their invaluable contributions.

References

- [1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.
- [2] Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiahu Wang, et al. 2025. AReal: A Large-Scale Asynchronous Reinforcement Learning System for Language Reasoning. *arXiv preprint arXiv:2505.24298* (2025).
- [3] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
- [4] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025.

- Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).
- [5] Jian Hu, Xibin Wu, Zilin Zhu, Weixun Wang, Dehao Zhang, Yu Cao, et al. 2024. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143* (2024).
 - [6] Huawei. 2025. MindSpeed-RL. <https://gitee.com/ascend/MindSpeed-RL>.
 - [7] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).
 - [8] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*. 611–626.
 - [9] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).
 - [10] Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Erran Li, Raluca Ada Popa, and Ion Stoica. 2025. DeepScaleR: Surpassing O1-Preview with a 1.5B Model by Scaling RL. <https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005bed8ca303013a4e2>. Notion Blog.
 - [11] Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I Jordan, et al. 2018. Ray: A distributed framework for emerging AI applications. In *13th USENIX symposium on operating systems design and implementation (OSDI 18)*. 561–577.
 - [12] Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. 2023. Scaling data-constrained language models. *Advances in Neural Information Processing Systems* 36 (2023), 50358–50376.
 - [13] Michael Noukhovitch, Shengyi Huang, Sophie Xhonneux, Arian Hosseini, Rishabh Agarwal, and Aaron Courville. 2024. Asynchronous RLHF: Faster and More Efficient Off-Policy RL for Language Models. *arXiv preprint arXiv:2410.18252* (2024).
 - [14] NVIDIA. 2024. Data Flywheel. <https://www.nvidia.com/en-us/glossary/data-flywheel/>.
 - [15] NVIDIA. 2024. TensorRT-LLM. <https://nvidia.github.io/TensorRT-LLM/>.
 - [16] OpenAI. 2024. Learning to reason with LLMs. <https://openai.com/index/learning-to-reason-with-llms/>.
 - [17] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.
 - [18] Ruoyu Qin, Zheming Li, Weiran He, Jiale Cui, Feng Ren, Mingxing Zhang, Yongwei Wu, Weimin Zheng, and Xinran Xu. 2025. Mooncake: Trading More Storage for Less Computation — A KVCache-centric Architecture for Serving LLM Chatbot. In *23rd USENIX Conference on File and Storage Technologies (FAST 25)*. USENIX Association, Santa Clara, CA, 155–170. <https://www.usenix.org/conference/fast25/presentation/qin>
 - [19] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).
 - [20] Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. 2024. Group robust preference optimization in reward-free rlhf. *Advances in Neural Information Processing Systems* 37 (2024), 37100–37137.
 - [21] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 3505–3506.
 - [22] Redis. 2025. Redis. <https://redis.io/>.
 - [23] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
 - [24] ByteDance Seed, Yufeng Yuan, Yu Yue, Mingxuan Wang, Xiaochen Zuo, Jiaze Chen, Lin Yan, Wenyan Xu, Chi Zhang, Xin Liu, et al. 2025. Seed1.5-Thinking: Advancing Superb Reasoning Models with Reinforcement Learning. *arXiv preprint arXiv:2504.13914* (2025).
 - [25] SemiAnalysis. 2024. Scaling Laws – O1 Pro Architecture, Reasoning Training Infrastructure, Orion and Claude 3.5 Opus “Failures”. <https://semianalysis.com/2024/12/11/scaling-laws-o1-pro-architecture-reasoning-training-infrastructure-orion-and-claude-3-5-opus-failures/>.
 - [26] Gerald Shen, Zhilin Wang, Olivier Delalleau, Jiaqi Zeng, Yi Dong, Daniel Egert, Shengyang Sun, Jimmy Zhang, Sahil Jain, Ali Taghibakhshi, et al. 2024. Nemo-Aligner: Scalable toolkit for efficient model alignment. *arXiv preprint arXiv:2405.01481* (2024).
 - [27] Guangming Sheng, Chi Zhang, Zilinfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv:2409.19256* (2024).
 - [28] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-LM: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053* (2019).
 - [29] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599* (2025).
 - [30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
 - [31] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. Position: Will we run out of data? Limits of LLM scaling based on human-generated data. In *Forty-first International Conference on Machine Learning*.
 - [32] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformer Reinforcement Learning. <https://github.com/huggingface/trl>.
 - [33] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652* (2021).
 - [34] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115* (2024).
 - [35] Zhewei Yao, Reza Yazdani Aminabadi, Olatunji Ruwase, Samyam Rajbhandari, Xiaoxia Wu, Ammar Ahmad Awan, Jeff Rasley, Minjia Zhang, Conglong Li, Connor Holmes, et al. 2023. DeepSpeed-Chat: Easy, fast and affordable rlhf training of chatgpt-like models at all scales. *arXiv preprint arXiv:2308.01320* (2023).
 - [36] Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. 2025. DAPO: An open-source llm reinforcement learning system at scale.

- arXiv preprint arXiv:2503.14476* (2025).
- [37] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* 1, 2 (2023).
- [38] Yinmin Zhong, Zili Zhang, Xiaoni Song, Hanpeng Hu, Chao Jin, Bingyang Wu, Nuo Chen, Yukun Chen, Yu Zhou, Changyi Wan, et al. 2025. StreamRL: Scalable, Heterogeneous, and Elastic RL for LLMs with Disaggregated Stream Generation. *arXiv preprint arXiv:2504.15930* (2025).
- [39] Yinmin Zhong, Zili Zhang, Bingyang Wu, Shengyu Liu, Yukun Chen, Changyi Wan, Hanpeng Hu, Lei Xia, Ranchen Ming, Yibo Zhu, and Xin Jin. 2025. Optimizing RLHF Training for Large Language Models with Stage Fusion. arXiv:2409.13221 [cs.LG] <https://arxiv.org/abs/2409.13221>