ICaRus: Identical Cache Reuse for Efficient Multi Model Inference

Anonymous authors
Paper under double-blind review

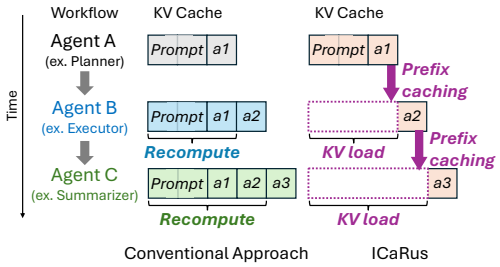
Abstract

Multi model inference, where multiple task-specialized models collaborate to solve complex real-world problems, has recently emerged as a prominent paradigm, particularly in the development of agentic AI systems. However, in such scenarios, each model must maintain its own Key-Value (KV) cache for the identical prompt, leading to substantial memory consumption. This explosive growth of KV caches forces LLM serving systems to evict previously stored caches, which in turn introduces significant recomputation overhead whenever the evicted caches are required again. Moreover, prefix caching is inherently infeasible across different models, forcing each model to recompute KV cache for the identical prompt, which leads to signficant overhead. To alleviate these issues, we propose **I**dentical **Ca**che **Reus**e (**ICaRus**), a novel architecture that allows multiple models to share identical KV caches across all layers. ICaRus is based on the key observation that a decoder-only Transformer can be conceptually decomposed into a logical encoder, which generates KV caches, and a logical decoder, which predicts output tokens from the KV caches. ICaRus fine-tunes only the logical decoder while freezing the logical encoder, enabling multiple models to share an identical KV cache. This eliminates cache memory explosion and unexpected evictions while also allowing cross-model reuse of KV caches for new input tokens, thereby removing redundant recomputation in multi model inference achieving both efficiency and scalability. Moreover, by incorporating lightweight adapters such as LoRA, ICaRus parallelizes KV cache generation and next-token prediction during decoding. ICaRus achieves comparable accuracy to task-specific fine-tuned model across a diverse set of tasks, while allowing multiple specialized models to fully share KV caches. ICaRus achieves up to $11.1\times$ lower P95 latency and $3.8\times$ higher throughput in agentic workflow with 8 different models, compared to conventional multi model system.

1 Introduction

Large Language Models (LLMs) have shown strong performance across domains (Zhao et al., 2024; Dubey et al., 2024; Comanici et al., 2025; Yang et al., 2025); however, a single model struggles with complex tasks that demand multi step reasoning and domain-specific expertise (Tang et al., 2020; Yao et al., 2023; Sun et al., 2024). Recently, the emerging paradigm of multi model inference addresses this limitation by orchestrating task-specialized models, achieving higher accuracy and problem-solving ability than a general-purpose model (Fu et al., 2023; Du et al., 2024; Shen et al., 2024; Subramaniam et al., 2025). However, this paradigm introduces severe challenges in managing the Key-Value (KV) cache: each model maintains its own cache even for identical prefixes, causing memory consumption to grow rapidly with the number of models. Once GPU memory is saturated by KV cache, serving systems (Kwon et al., 2023; Zheng et al., 2024) must evict caches, which triggers redundant recomputation and significantly degrades throughput. Furthermore, because KV caches are model-specific, prefix caching (Kwon et al., 2023; Zheng et al., 2024) cannot be applied across different models, which forces identical prompts to rebuild KV caches independently and thereby increases latency.

Previous KV cache optimization techniques, such as pruning (Zhang et al., 2023), quantization (Hooper et al., 2024; Yang et al., 2024), and inter-layer sharing (Qiao et al., 2024), reduce cache size while minimizing accuracy degradation. Unlike traditional LRU-based prefix caching, KVFlow



(a) KV Cache management strategies in agent workflow using multi model

	Single Model	Multi Model	
		Conventional	ICaRus (Ours)
Training Method	Prompting	Fine-tuning	**Fine-tuning** (only logical Decoder)
Task Performance	Weak	Strong	**Strong**
KV Sharing	Inherent	Unsupported	**Supported**
KV Memory Usage	Low	High	**Low**
# Prefill Recomputation	Low	High	**Low**

(b) Comparison of ICaRus and conventional approaches

Figure 1: Comparison of KV cache management strategies and effectiveness in multi model scenarios between conventional approaches and ICaRus.

(Pan et al., 2025) schedules KV cache eviction and prefetching based on anticipated agent workflow, reducing recomputation overhead. However, these methods focus only on single model cache management, leaving unresolved the challenges of cache explosion and the lack of KV cache sharing of prefix in multi model settings. DroidSpeak (Liu et al., 2024b) addressed multi model KV cache management by sharing non-sensitive layer caches between a base model and its fine-tuned variants, thereby reducing recomputation cost. However, this approach has inherent limitations, as caches from sensitive layers remain unshared and must still be recomputed.

To address these issues, we propose **I**dentical **Ca**che **Reus**e (**ICaRus**), a novel architecture that enables multiple models to share and reuse the same KV cache across all layers. The core idea of ICaRus originates from conceptually decomposing a decoder-only Transformer into two parts: a logical encoder, which is responsible for generating KV cache, and a logical decoder, which predicts the next token from the cache. We freeze the logical encoder of pretrained LLM (i.e. base model) and fine-tune only the logical decoder for each specific task using lightweight adapters such as LoRA (Hu et al., 2022). Since all specialized models share the identical logical encoder, the KV cache generated for an identical prompt is likewise identical, enabling direct sharing without redundant memory usage as shown in Fig. 1(a). This prevents GPU memory from rapidly saturating due to KV cache growth, avoiding costly recomputation caused by cache eviction. Moreover, shared KV caches enable prefix caching across models, eliminating redundant computation for identical prompts and further improving efficiency as depicted in Fig. 1(b). In addition, ICaRus leverages the adapter architecture to generate the KV cache for the next step in parallel with the next-token computation during the decode phase. We evaluate ICaRus across diverse tasks including mathematics, coding, and knowledge understanding on a wide range of model families and scales (LLaMA-3.1-8B, Qwen3-1.7B/8B/14B). The results demonstrate that ICaRus achieves accuracy comparable to task-specific fine-tuned models, even though ICaRus-tuned models are able to share KV caches across tasks. Furthermore, when integrated into the vLLM serving system and evaluated in various multi agent scenarios including ReAct (Yao et al., 2023) and Reflexion (Shinn et al., 2023), ICaRus delivers as much as a $11.1\times$ reduction in 95th-percentile (P95) latency and a $3.8\times$ throughput gain compared to conventional multi model systems.

In summary, the main contributions of this work are as follows:

- We propose ICaRus, the first architecture that enables multiple decoder-only Transformers to fully share KV caches across all layers, providing a principled solution to inefficiencies in conventional multi model serving approach.
- We demonstrate that ICaRus achieves accuracy comparable to task-specific fine-tuning across diverse tasks (mathematics, coding, and knowledge understanding) and model architectures (LLaMA-3.1-8B, Qwen3-1.7B/8B/14B).
- We confirm that ICaRus significantly improves efficiency in multi agent workflows, achieving up to $11.1\times$ reduction in P95 latency and $3.8\times$ improvement in throughput compared to conventional multi model system.

2 BACKGROUND & MOTIVATION

Key-Value Cache in LLM Serving Systems. During autoregressive inference, decoder-only Transformers generate tokens sequentially, where each new token depends on all previously generated tokens. Computing self-attention naïvely for every step requires recomputation over the entire sequence, incurring a per-token complexity of $\mathcal{O}(n^2)$ where n is the sequence length. To avoid this quadratic overhead, modern LLM serving systems cache the key and value representations of previously processed tokens (Vaswani et al., 2017). By reusing these cached states, each new decoding step only attends to the most recent token, reducing the per-token attention complexity to $\mathcal{O}(n)$ and thereby significantly lowering computational cost. However, the size of KV caches grows linearly with both sequence length and model depth, imposing substantial memory pressure on GPU-based serving systems (Kwon et al., 2023; Zheng et al., 2024). Consequently, memory-efficient cache management has emerged as a critical challenge for scalable LLM deployment.

Prefix Caching in LLM Serving Systems. Prefix caching is a widely adopted optimization that reuses the KV cache corresponding to a fixed prefix across multiple queries sharing the same initial context (Kwon et al., 2023; Zheng et al., 2024). This technique is particularly effective in scenarios such as retrieval-augmented generation (RAG) (Lewis et al., 2020) and instruction-tuned applications (Chung et al., 2024; Ouyang et al., 2022), where prompts often contain long but invariant components like system prompts, task-specific templates, or retrieved documents. By reusing the cached key-value states of these repeated prefixes, serving systems can avoid redundant computation during the prefill phase, effectively reducing the computational complexity from $\mathcal{O}(n^2)$ to $\mathcal{O}(mn)$, where n denotes the sequence length and m denotes the variable suffix length with $m \ll n$, thereby improving both throughput and latency. Moreover, prefix caching is highly beneficial in multi-turn conversational settings, where a large dialogue history is preserved across turns and only the most recent user utterance changes; by caching the KV states of the shared history and computing attention only for newly appended tokens, serving systems can efficiently support interactive dialogues without recomputing the entire context at every turn (Kim et al., 2025).

Agentic AI Workflow and Multi Model Inference. Agentic AI and workflow-based reasoning have given rise to complex pipelines in which models are orchestrated to perform specialized roles. For instance, ReAct (Yao et al., 2023) alternates between *Thought* $\rightarrow$ *Act* $\rightarrow$ *Observation*, Reflexion (Shinn et al., 2023) incorporates self-evaluation loops, LATS (Zhou et al., 2024) explores reasoning through parallel branch expansion, and LLMCompiler (Kim et al., 2024) constructs a DAG to schedule overlapping tool and model calls. When executed within a single model, such workflows can leverage prefix caching to avoid redundant computation, thereby reducing effective memory usage, lowering P95 latency, and improving throughput (Kim et al., 2025). However, in multi model settings where task-specialized models collaborate within a single pipeline, each model must maintain its own KV cache even for identical prefixes. Such KV cache duplication leads to memory usage that grows linearly with the number of active models; once GPU capacity is saturated, this growth inevitably triggers cache eviction, which in turn forces recomputation of evicted prefixes. Moreover, since prefix caching typically operates only within individual models, identical prefixes must be recomputed separately across models, leading to redundant prefill computation that inflates both latency and energy consumption. These limitations underscore the need for new architectures that support cross-model KV sharing and prefill de-duplication in multi model inference.

3 DESIGN OF ICARUS

3.1 DECODER-ONLY TRANSFORMER AS LOGICAL ENCODER AND DECODER

We first present a mathematical formulation of the decoder-only Transformer, which predicts the next token conditioned on the current token context. Specifically, we abstract x_i, k_i, and v_i as the i-th token, its key representation, and its value representation, respectively, and denote the decoder-only Transformer by F. In this case, the next-token generation from the current token context in a decoder-only Transformer can be expressed as $x_{i+1} = F(x_1, x_2, \ldots, x_i)$. To generate the next token x_{i+1}, the model requires two types of information: the current token x_i and the accumulated key–value pairs. We denote the key set and value set up to step i as $K_{1:i} = \{k_1, k_2, \ldots, k_i\}, V_{1:i} = \{v_1, v_2, \ldots, v_i\}$. More concretely, in the attention operation, the query derived from x_i is generated

anew at each step, whereas the keys and values are continuously appended to the cache and reused across subsequent decoding steps. In other words, the query does not persist beyond its step, but the KV pairs accumulate and form the long-term memory. This dependency can be expressed as

$$x_{i+1} = F(x_1, x_2, \ldots, x_i) = F\big(x_i,\, K_{1:i}, V_{1:i}\big). \tag{1}$$

Eq.1 indicates that a decoder-only Transformer predicts the next token conditioned on the current token x_i and the KV cache constructed up to this point. More generally, the generation process can be decomposed into two conceptual stages: (1) constructing the key set K_i and the value set V_i from the input sequence $x_{1:i} = \{x_1, x_2 \ldots, x_i\}$, and (2) decoding the next token x_{i+1} based on the current token x_i together with the accumulated sets (K_i, V_i). Formally, this can be expressed as

$$K_{1:i}, V_{1:i} = E(x_{1:i}), \tag{2}$$

$$x_{i+1} = D\big(x_i,\, K_{1:i}, V_{1:i}\big), \tag{3}$$

where E denotes the logical encoder that transforms the input sequence into its key and value representations, thereby constructing the KV cache, and D denotes the logical decoder that consumes the current token and the KV set to generate the next token. Importantly, a decoder-only Transformer can be interpreted as the special case where the parameters of the logical encoder and logical decoder are identical.

3.2 ICaRus: Identical Cache Reuse across LLMs

As described in Section 3.1, a decoder-only model can be decomposed into a logical encoder, which generates key–value pairs from a given token, and a logical decoder, which predicts the next token using the current token and the accumulated KV cache, as shown in Eqs. 2–3. From this perspective, task-specific fine-tuning can be viewed as jointly training both the logical encoder and the logical decoder to specialize in a given task. While such task-tuned models achieve strong task-specific capabilities, each maintains its own logical encoder thereby preventing KV cache sharing even when prompts are identical across models.

Building on this insight, we propose the ICaRus architecture which fine-tunes only the logical decoder of a decoder-only Transformer as below.

$$K_{1:i}, V_{1:i} = E_t(x_{1:i}) = E(x_{1:i}), \tag{4}$$

$$x^t_{i+1} = D_t\big(x_i,\, K_{1:i}, V_{1:i}\big), \tag{5}$$

Here, t and D_t denote a specific task and the logical decoder fine-tuned for that task, respectively. Specifically, the logical encoder (E) and the logical decoder (D) are initialized with the parameters of the base model, a pretrained decoder-only Transformer. The task-specific logical decoder D_t in Eq. 5 is then trained, starting from the base decoder D, to predict the next token x_{i+1} under two

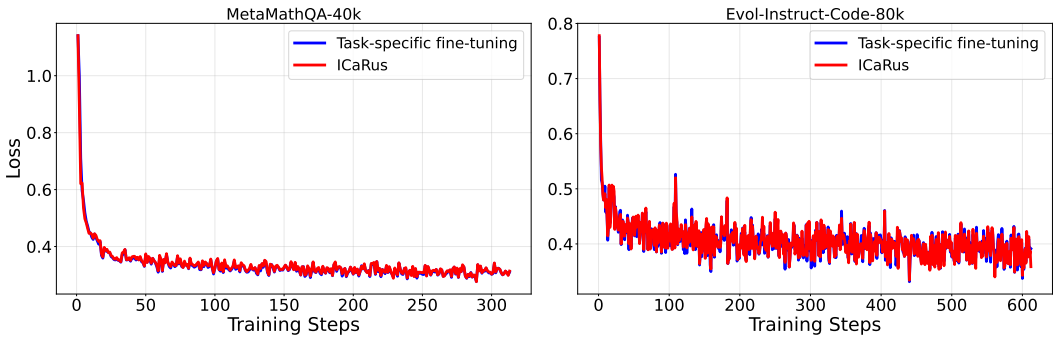


Figure 2: Training loss curves of task-specific fine-tuning and ICaRus, both applied with LoRA on LLaMA-3.1-8B, trained on the MetaMathQA-40k and Evol-Instruct-Code-80k dataset.

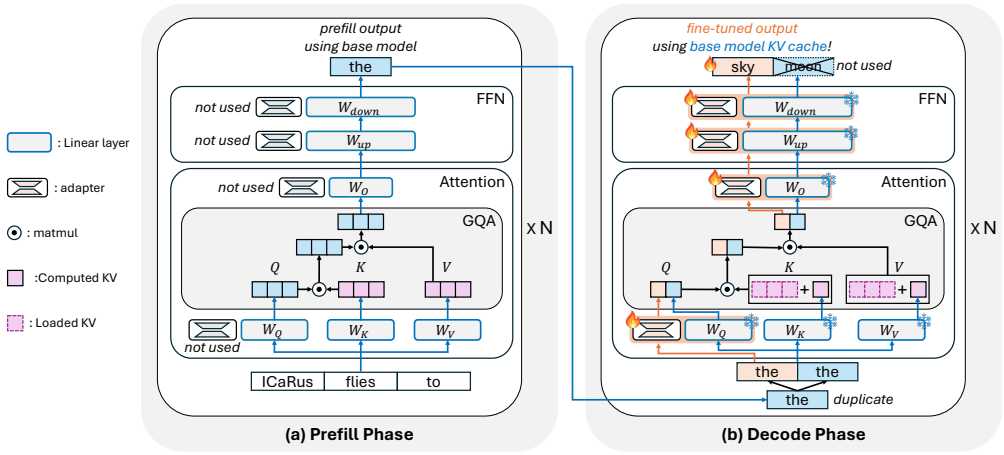


Figure 3: Overview of the ICaRus architecture. The base model, a pretrained decoder-only Transformer, serves as the logical encoder, while the adapter-tuned model (consisting of the base model and a tunable adapter) serves as the logical decoder. The blue and orange lines indicate computations performed by the base model and the adapter-tuned model, respectively.

objectives: (1) specializing in the target task, and (2) leveraging the KV cache generated by the frozen logical encoder in Eq. 4. As a result, multiple task-specific logical decoders (e.g., D_{math}, D_{coding}, $D_{\text{reasoning}}$) can share a single logical encoder (i.e., $E_{\text{math}} \equiv E_{\text{coding}} \equiv E_{\text{reasoning}} \equiv E$), which is identical to the base model, thereby enabling all models to reuse the identical KV cache generated by the shared encoder, as illustrated in Fig. 1.

During training, the input data are duplicated and provided to both the logical encoder and the logical decoder. The logical encoder generates the corresponding key–value representations, while the logical decoder computes attention over these representations with its final output used to compute the training loss for gradient updates. The logical encoder is kept frozen during training to ensure cache sharing across tasks. We confirmed the validity of the proposed ICaRus architecture by analyzing the convergence curves of LLaMA-3.1-8B during training on the MetaMathQA-40k (Yu et al., 2023) and Evol-Instruct-80k (Roshdieh, 2023) datasets as shown in Fig. 2. The results show that our ICaRus achieves stable loss convergence comparable to conventional task-specific fine-tuning. This training paradigm is consistent with prior effective fine-tuning that train only a subset of parameters within a single model (Hu et al., 2022; Liu et al., 2024a; Jiang et al., 2024; Woo et al., 2025).

3.3 Optimizing ICaRus for Multi Model Inference

In Section 3.2, we introduced the concept and training methodology of ICaRus. In this section, we explain how ICaRus operates in multi model inference scenarios and discuss its key optimization strategies. During the prefill phase, ICaRus uses only the logical encoder, which encodes the input prompt into a KV cache and produces the next token. In the subsequent decode phase, ICaRus duplicates the current token (x_i) and performs two operations: (1) encoding x_i into a key–value pair (k_i, v_i) through the logical encoder, and (2) predicting the task-specific output token (x_{i+1}) through the logical decoder by using the duplicated x_i together with the accumulated KV cache ($\{k_1, \ldots, k_i\}, \{v_1, \ldots, v_i\}$), as in Eq. 5.

Sequential execution of the logical encoder and decoder may incur up to 2× latency overhead compared to a single model execution, since both weights and KV caches are accessed twice. To mitigate the problem, we insert and fine-tune only lightweight adapters within the logical decoder instead of fully fine-tuning the decoder. Consequently, the logical encoder and logical decoder share most parameters except for the adapters, enabling the shared parameters to be loaded only once and allowing the computations of the two modules to be executed in parallel as depicted in Fig. 3.

In addition, because both models attend to the identical KV cache generated by the base model, we optimize attention computation by concatenating the query representations of the logical encoder and decoder along the head dimension (Fig. 3). This enables parallel attention computation without

Table 1: Space and time complexity comparisons between single model and multi model scenarios.

Scenario	Method	Space Complexity	Time Complexity		
		Total	Prefill	Decode (per token)	
				Memory Access	Compute
Single Model	—	$\mathcal{O}(M + L_t)$	$\mathcal{O}(ML_t + L_t^2)$	$\mathcal{O}(M + L_t)$	$\mathcal{O}(M + L_t)$
Multi Model	BaseLine	$\mathcal{O}(M + NL_t)$	$\mathcal{O}(N(ML_t + L_t^2))$	$\boldsymbol{\mathcal{O}(M + L_t)}$	$\boldsymbol{\mathcal{O}(M + L_t)}$
	ICaRus	$\boldsymbol{\mathcal{O}(M + L_t)}$	$\boldsymbol{\mathcal{O}(ML_t + L_t^2)}$	$\boldsymbol{\mathcal{O}(M + L_t)}$	$\mathcal{O}(2M + 2L_t)$

redundant KV cache reads. Consequently, although the decoding phase of ICaRus appears to double the computational workload by running both the logical encoder and decoder, the system adds only negligible latency overhead. This is because parallel execution generates memory traffic (base parameters, KV caches, and lightweight adapter weights) that is almost the same as that of a single model. The detailed algorithm can be found in Appendix B,

To validate the effectiveness of ICaRus, we further analyze the time and space complexity of multi model systems built with the conventional approach (baseline) and with ICaRus, using N adapters in multi agent scenarios. Table 1 summarizes the results. We denote the input prompt length as L_i, the number of interaction turns per adapter as t, and the number of output tokens per turn as L_o, with the total sequence length $L_t = L_i + tL_o$. The base model size is represented by M. In the baseline, each model independently allocates KV memory and recomputes prefill for the same prompt, yielding space complexity $\mathcal{O}(M + NL_t)$ and prefill complexity $\mathcal{O}(N(ML_t + L_t^2))$. In contrast, ICaRus shares a single KV cache across models, reducing both to single model order, with space $\mathcal{O}(M + L_t)$ and prefill $\mathcal{O}(ML_t + L_t^2)$. The advantage grows with longer sequences from inter-model communication and with larger agent counts N.

During decoding, the baseline requires $\mathcal{O}(M + L_t)$ memory access and computation per token because each adapter-tuned model reads the model weights and its own KV cache. ICaRus computes both the logical encoder and decoder ($\mathcal{O}(2M + 2L_t)$) but parallelizes most of the computation so that the model and KV cache are read only once, restoring $\mathcal{O}(M + L_t)$. In multi-model, long-context, many-turn settings where decoding is memory-bound, memory access dominates; accordingly, ICaRus achieves decoding latency comparable to the baseline.

4 Evaluation

4.1 Experimental Setup

We evaluate ICaRus from two perspectives: (1) fine-tuning accuracy and (2) performance in multi model inference. In section 4.2, we train ICaRus using LLaMA-3.1-8B (Dubey et al., 2024) and Qwen3-1.7B/8B/14B-Base (Yang et al., 2025) without thinking on task-specific datasets, including MetaMathQA-40k for mathematics (Yu et al., 2023), Evol-Instruct-Code-80k for coding (Roshdieh, 2023), and OASST1 for instruction tuning (Köpf et al., 2023). Models are then evaluated on benchmarks aligned with each task: GSM8K (Cobbe et al., 2021) and GSM-Plus (Li et al., 2024) for mathematics, HumanEval (Chen et al., 2021) and HumanEval+ (Liu et al., 2023) for coding, and GPQA-Diamond (Rein et al., 2024) for knowledge understanding using lm-eval-harness (Biderman et al., 2024) and EvalPlus (Liu et al., 2023), measuring 0-shot accuracy. For comparison, we also consider task-specific fine-tuning, where both ICaRus and the baselines employ LoRA (Hu et al., 2022).

For multi model inference (Section 4.3), we measure latency and throughput under representative agent workflows such as ReAct (Yao et al., 2023) and Reflexion (Shinn et al., 2023). We adapt these workflows to a multi model, multi-turn request-routing setup: within a single workflow, successive requests from a multi-turn interaction are routed in a round-robin manner to different models. In this setting, the baseline is a conventional multi-LoRA system, whereas ICaRus replaces it with a cache-sharing multi agent system. To ensure a fair comparison, we integrate both systems into the vLLM serving framework and evaluate them under identical settings. More details can be found in the Appendix A.

Table 2: Comparisons of prior task-specific fine-tuning and ICaRus on various datasets.

Model	Method	Math		Coding		Knowledge
		GSM8K	GSM+	HEval	HEval+	GPQA
LLaMA3.1-8B	No-tuning	25.9	18.0	36.6	29.9	16.7
	Task-tuning	**69.7**	**48.5**	48.2	41.5	27.3
	ICaRus (Ours)	67.9	45.8	**48.2**	**43.9**	**28.8**
Qwen3-8B-Base	No-tuning	11.8	12.5	68.3	61.6	24.2
	Task-tuning	85.4	66.1	81.7	75.6	**34.3**
	ICaRus (Ours)	**87.3**	**67.5**	**86.6**	**79.9**	33.8

Table 3: Comparison of task-specific fine-tuning and ICaRus across different model sizes (Qwen3-1.7B/8B/14B-Base) trained on the MetaMathQA-40K dataset.

Model	Qwen3-1.7B-Base		Qwen3-8B-Base		Qwen3-14B-Base	
Method	Task-tuning	ICaRus	Task-tuning	ICaRus	Task-tuning	ICaRus
GSM8K	73.2	**74.0**	85.4	**87.3**	85.6	**88.8**
GSM+	53.7	**54.1**	66.1	**67.5**	66.7	**68.8**

4.2 Fine-Tuning Accuracy

Accuracy on diverse task. We first train and evaluate ICaRus alongside task-specific fine-tuning (Task-tuning) across mathematics, coding, and instruction-tuning tasks using LLaMA-3.1-8B and Qwen3-8B, as reported in Table 2. The results show that ICaRus achieves accuracy comparable to, or even surpassing, task-specific fine-tuning across all tasks. In particular, for the Qwen3-8B-Base model, ICaRus outperforms prior task-tuned models by at least 1.4% on benchmark evaluations for both mathematics and coding tasks. We expect that the superior accuracy of ICaRus stems from a generalization effect: by fine-tuning only the logical decoder while keeping the logical encoder frozen, ICaRus reduces the risk of overfitting compared to full task-specific fine-tuning.

Scaling with model size. We also examine the scalability of ICaRus with respect to model size by conducting experiments on Qwen3-1.7B/8B/14B-Base in Table 3. The results show that ICaRus consistently achieves higher accuracy compared to prior task-specific fine-tuned models, with improvements exceeding 2% on Qwen3-14B-Base, demonstrating that our method remains competitive as model capacity increases.

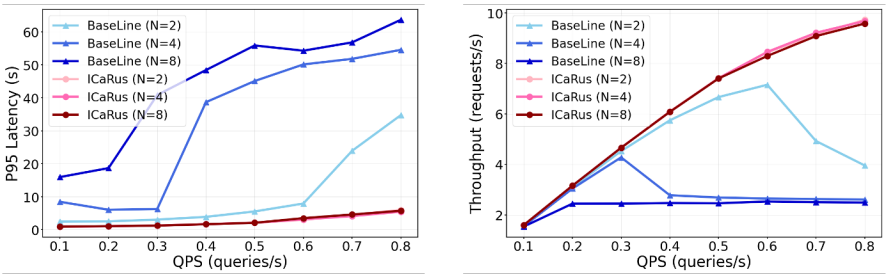
Table 4: Comparison of task-specific fine-tuning and ICaRus in both single and multi model inference scenarios.

# Model	Method	Math		Coding		Knowledge	Avg.
		GSM8K	GSM-Plus	HEval	HEval+	GPQA	
Single Model	No-tuning	25.9	18.0	36.6	29.9	16.7	25.4
	Math-tuning	**69.7**	**48.5**	42.7	36.6	20.7	43.6
	Code-tuning	22.8	17.5	**48.2**	41.5	21.7	30.3
	Instruct-tuning	24.5	16.5	44.5	39.0	27.2	30.3
Multi Model	Task-tuning	**69.7**	**48.5**	**48.2**	41.5	27.2	**47.0**
	ICaRus (Ours)	67.9	45.8	**48.2**	**43.9**	**28.8**	46.9

Multi domain orchestration results. Table 4 compares ICaRus orchestration with different task-specific fine-tuning configurations using LLaMA-3.1-8B. Each task-tuned model is fine-tuned on a single domain-specific dataset (MetaMathQA for mathematics, Evol-Instruct-Code-80K for coding, and OASST1 for instruction-tuning). The results show that while a single task-specific fine-tuned model achieves high accuracy on its target task, the model suffers from significant performance degradation on other tasks. In contrast, a multi model orchestration system composed of multiple

task-specific fine-tuned models achieves consistently high accuracy across all tasks. Our ICaRus orchestration system also attains accuracy comparable to such multi model systems, while additionally benefiting from KV cache sharing across agents, which enables orchestration at substantially lower computational cost.

4.3 Performance on Multi Model Inference



(a) P95 latency across QPS

(b) Throughput across QPS

Figure 4: P95 latency and throughput of ICaRus compared with multiple task-specific agents fine-tuned from the LLaMA-3.1-8B base model under the ReAct workflow. Here, N denotes the number of LoRA modules, which are integrated into multi model systems built using either the conventional approach or ICaRus.

P95 latency and throughput across QPS. ICaRus consistently outperforms a baseline multi model system across all load levels in both latency and throughput, as evaluated on LLaMA-3.1-8B under the ReAct workflow (Fig. 4). We measure performance as the number of queries per second (QPS) increases; latency is reported at the 95th percentile (P95).

A key advantage of ICaRus is its ability to reuse identical prefix caches across models, avoiding the redundant recomputation required in baseline systems where each model reconstructs its own cache. For example, at QPS 0.3 with 4 models, ICaRus reduces P95 latency by $5.1\times$ compared to the baseline, and this benefit becomes more pronounced as the number of models increases.

As the QPS increases, the cumulative KV cache size of baseline systems soon exceeds GPU memory capacity, triggering eviction of previously stored KV caches and their subsequent recomputation. Consequently, throughput first plateaus and then declines, with the degradation occurring earlier as the number of models increases (e.g., at 0.6 QPS for two models and 0.3 QPS for four models; Fig. 4(b). In contrast, ICaRus avoids redundant cache growth through cross-model KV sharing, allowing throughput to continue increasing even as baseline systems plateau and decline.

Consequently, when comparing maximum achievable throughput, ICaRus outperforms the baseline by $1.4\times$, $2.3\times$, and $3.8\times$ with 2-, 4-, and 8-model systems, respectively. At the QPS where baseline systems reach their peak throughput, ICaRus also achieves substantially lower P95 latency-$3.8\times$, $5.1\times$, and $11.1\times$ for 2-, 4-, and 8- models, respectively.

Performance under diverse workflows or models. We further evaluate baseline systems and ICaRus systems across different models (LLaMA-3.1-8B and Qwen3-14B-Base) and multi agent workflows (ReAct and Reflexion). Specifically, we measure P95 latency over varying QPS and the maximum throughput achieved at the optimal QPS setting, as summarized in Fig. 5.

ICaRus prevents KV cache explosion and enables cross-model prefix caching, thereby achieving lower P95 latency and higher throughput in multi agent workflows. These gains persist even for larger models like Qwen3-14B, where ICaRus achieves up to $7.4\times$ lower latency and $3.6\times$ higher throughput compared to the baseline.

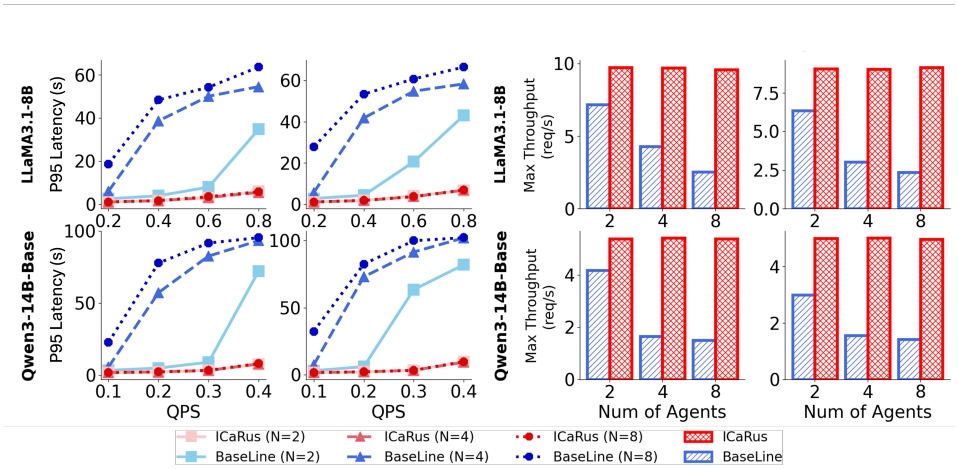


Figure 5: Comparison of P95 latency and maximum throughput across QPS for LLaMA3.1-8B and Qwen-3-14B Base under ReAct and Reflexion workflows.

5 Related Work

Multi model Inference Leveraging multiple models has been widely explored as a way to improve performance over a single model. Routing methods either select the most appropriate model or use multiple models in a cascade (Chen et al., 2024; Shnitzer et al., 2024), while ensemble approaches combine the outputs of multiple models, either at the token level (Yu et al., 2024; Huang et al., 2024) or at the reasoning step level (Park et al., 2025). Multi model approaches have also been applied in multi agent systems, where interactions among agents have been shown to enhance performance across diverse tasks (Fu et al., 2023; Sun et al., 2024; Du et al., 2024). In these systems, each agent used either a base model or fine-tuned variants obtained with methods such as LoRA or instruction tuning (Mineiro, 2024; Liu et al., 2025).

KV Cache Optimization KV cache stores the keys and values of previous tokens to avoid redundant recomputation during autoregressive generation and is traditionally used on a per-request basis (Vaswani et al., 2017). Prefix caching techniques extend the lifetime of the KV cache beyond a single request, enabling multiple turns or related requests to share the same cache (Gao et al., 2024; Gim et al., 2024). However, prefix caching alone cannot address the challenge of deploying multiple models, as KV caches cannot be shared across different models even for identical prompts, and each model generates a distinct KV cache. DroidSpeak (Liu et al., 2025) addresses this issue by reusing the KV cache of a shared foundational model for non-sensitive layers, while selectively recomputing only the sensitive layers in each agent model. This approach requires identifying sensitive layers that must be recomputed by the agent model, thereby affecting subsequent layers. On a different axis, KVFlow (Pan et al., 2025) manages KV caches by evicting and prefetching based on predetermined agentic workflows instead of an LRU policy, but it remains a single model approach with agents defined by prompts.

6 Conclusion

In this work, we presented ICaRus, a KV cache-sharing architecture for multi model inference. ICaRus addresses the memory inefficiency of conventional systems by enabling cross-model KV cache reuse, while maintaining accuracy through fine-tuning. Experiments across mathematics, coding, and instruction-following tasks confirm that ICaRus delivers accuracy on par with task-specific fine-tuned models, yet achieves significantly lower latency and higher throughput in multi agent workflows. Taken together, these results establish ICaRus as a principled approach for scalable and efficient multi model inference. Looking ahead, we expect ICaRus to extend to large-scale models, heterogeneous agent systems, and real-world deployment scenarios where scalability and efficiency are increasingly critical.

Reproducibility Statement

We formulated the concept of the logical encoder and decoder in detail, which forms the foundation of the ICaRus algorithm, in Section 3.1. Furthermore, we provided a rigorous mathematical formulation of ICaRus, along with its training procedure and convergence of the loss curve, in Section 3.2. The inference process of ICaRus and the corresponding optimization strategies are described in Section 3.3, with pseudocode provided in Appendix B. Finally, the detailed experimental setup for both training and inference is presented in Section 4.1 and Appendix A.

References

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024. URL https://arxiv.org/abs/2405.14782.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=cSimKw5p6R.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25:70:1–70:53, 2024. URL https://jmlr.org/papers/v25/23-0870.html.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit S. Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Sharon Silver, Ayzaan Wahid, Sergey Brin, Yves Raimond, Klemen Kloboves, Cindy Wang, Nitesh Bharadwaj Gundavarapu, Ilia Shumailov, Bo Wang, Mantas Pajarskas, Joe Heyward, Martin Nikoltchev, Maciej Kula, Hao Zhou, Zachary Garrett, Sushant Kafle, Sercan Arik, Ankita Goel, Mingyao Yang, Jiho Park, Koji Kojima, Parsa Mahmoudieh, Koray Kavukcuoglu, Grace Chen, Doug Fritz, Anton Bulyenov, Sudeshna Roy, Dimitris Paparas, Hadar Shemtov, Bo-Juen Chen, Robin Strudel, David Reitter, Aurko Roy, Andrey Vlasov, Changwan Ryu, Chas Leichner, Haichuan Yang, Zelda Mariet, Denis Vnukov, Tim Sohn, Amy Stuart, Wei Liang, Minmin Chen, Praynaa Rawlani, Christy Koh, JD Co-Reyes, Guangda Lai, Praseem Banzal, Dimitrios Vytiniotis, Jieru Mei, and Mu Cai. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261, 2025. doi: 10.48550/ARXIV.2507.06261. URL https://doi.org/10.48550/arXiv.2507.06261.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=zj7YuTE4t8.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. Improving language model negotiation with self-play and in-context learning from ai feedback, 2023. URL https://arxiv.org/abs/2305.10142.

Bin Gao, Zhuomin He, Puru Sharma, Qingxuan Kang, Djordje Jevdjic, Junbo Deng, Xingkun Yang, Zhou Yu, and Pengfei Zuo. {Cost-Efficient} large language model serving for multi-turn conversations with {CachedAttention}. In *2024 USENIX Annual Technical Conference (USENIX ATC 24)*, pp. 111–126, 2024.

In Gim, Guojun Chen, Seung-seob Lee, Nikhil Sarda, Anurag Khandelwal, and Lin Zhong. Prompt cache: Modular attention reuse for low-latency inference. In P. Gibbons, G. Pekhimenko, and C. De Sa (eds.), *Proceedings of Machine Learning and Systems*, volume 6, pp. 325–338, 2024.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. Kvquant: Towards 10 million context length LLM inference with KV cache quantization. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/028fcbcf85435d39a40c4d61b42c99a4-Abstract-Conference.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yichong Huang, Xiaocheng Feng, Baohang Li, Yang Xiang, Hui Wang, Ting Liu, and Bing Qin. Ensemble learning for heterogeneous large language models with deep parallel collaboration. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=7arAADUK6D.

Ting Jiang, Shaohan Huang, Shengyue Luo, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. Mora: High-rank updating for parameter-efficient fine-tuning. *CoRR*, abs/2405.12130, 2024. doi: 10.48550/ARXIV.2405.12130. URL https://doi.org/10.48550/arXiv.2405.12130.

Jiin Kim, Byeongjun Shin, Jinha Chung, and Minsoo Rhu. The cost of dynamic reasoning: Demystifying AI agents and test-time scaling from an AI infrastructure perspective. *CoRR*, abs/2506.04301, 2025. doi: 10.48550/ARXIV.2506.04301. URL https://doi.org/10.48550/arXiv.2506.04301.

Sehoon Kim, Suhong Moon, Ryan Tabrizi, Nicholas Lee, Michael W. Mahoney, Kurt Keutzer, and Amir Gholami. An LLM compiler for parallel function calling. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=uQ2FUoFjnF.

Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in neural information processing systems*, 36:47669–47681, 2023.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In Jason Flinn, Margo I. Seltzer, Peter Druschel, Antoine Kaufmann, and Jonathan Mace (eds.), *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pp. 611–626. ACM, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. Gsm-plus: A comprehensive benchmark for evaluating the robustness of llms as mathematical problem solvers. *arXiv preprint arXiv:2402.19255*, 2024.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Advances in Neural Information Processing Systems*, 36:21558–21572, 2023.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024a. URL https://openreview.net/forum?id=3d5CIRG1n2.

Yuhan Liu, Esha Choukse, Shan Lu, Junchen Jiang, and Madan Musuvathi. Droidspeak: Enhancing cross-llm communication. *CoRR*, abs/2411.02820, 2024b. doi: 10.48550/ARXIV.2411.02820. URL https://doi.org/10.48550/arXiv.2411.02820.

Yuhan Liu, Yuyang Huang, Jiayi Yao, Shaoting Feng, Zhuohan Gu, Kuntai Du, Hanchen Li, Yihua Cheng, Junchen Jiang, Shan Lu, Madan Musuvathi, and Esha Choukse. Droidspeak: Kv cache sharing for cross-llm communication and multi-llm serving, 2025. URL https://arxiv.org/abs/2411.02820.

Paul Mineiro. Online joint fine-tuning of multi-agent flows, 2024. URL https://arxiv.org/abs/2406.04516.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural*

Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Zaifeng Pan, Ajjkumar Patel, Zhengding Hu, Yipeng Shen, Yue Guan, Wan-Lu Li, Lianhui Qin, Yida Wang, and Yufei Ding. Kvflow: Efficient prefix caching for accelerating llm-based multi-agent workflows. *CoRR*, abs/2507.07400, 2025. doi: 10.48550/ARXIV.2507.07400. URL https://doi.org/10.48550/arXiv.2507.07400.

Sungjin Park, Xiao Liu, Yeyun Gong, and Edward Choi. Ensembling large language models with process reward-guided tree search for better complex 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. 10256–10277, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.515. URL https://aclanthology.org/2025.naacl-long.515/.

Aurick Qiao, Zhewei Yao, Samyam Rajbhandari, and Yuxiong He. Swiftkv: Fast prefill-optimized inference with knowledge-preserving model transformation. *CoRR*, abs/2410.03960, 2024. doi: 10.48550/ARXIV.2410.03960. URL https://doi.org/10.48550/arXiv.2410.03960.

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.

Nick Roshdieh. Evol-instruct-code-80k. https://huggingface.co/datasets/nickrosh/Evol-Instruct-Code-80k-v1, 2023. Hugging Face dataset.

Weizhou Shen, Chenliang Li, Hongzhan Chen, Ming Yan, Xiaojun Quan, Hehong Chen, Ji Zhang, and Fei Huang. Small llms are weak tool learners: A multi-llm agent. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 16658–16680. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.929. URL https://doi.org/10.18653/v1/2024.emnlp-main.929.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Tal Shnitzer, Anthony Ou, Mírian Silva, Kate Soule, Yuekai Sun, Justin Solomon, Neil Thompson, and Mikhail Yurochkin. Large language model routing with benchmark datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Zb0ajZ7vAt.

Vighnesh Subramaniam, Yilun Du, Joshua B. Tenenbaum, Antonio Torralba, Shuang Li, and Igor Mordatch. Multiagent finetuning: Self improvement with diverse reasoning chains. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=JtGPIZpOrz.

Qiushi Sun, Zhangyue Yin, Xiang Li, Zhiyong Wu, Xipeng Qiu, and Lingpeng Kong. Corex: Pushing the boundaries of complex reasoning through multi-model collaboration. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=7BCmIWVT0V.

Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. Progressive layered extraction (PLE): A novel multi-task learning (MTL) model for personalized recommendations. In Rodrygo L. T. Santos, Leandro Balby Marinho, Elizabeth M. Daly, Li Chen, Kim Falk, Noam Koenigstein, and Edleno Silva de Moura (eds.), *RecSys 2020: Fourteenth ACM Conference on Recommender Systems, Virtual Event, Brazil, September 22-26, 2020*, pp. 269–278. ACM, 2020. doi: 10.1145/3383313.3412236. URL https://doi.org/10.1145/3383313.3412236.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Sunghyeon Woo, Sol Namkung, Sunwoo Lee, Inho Jeong, Beomseok Kim, and Dongsuk Jeon. Paca: Partial connection adaptation for efficient fine-tuning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=iYkhxre0In.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

June Yong Yang, Byeongwook Kim, Jeongin Bae, Beomseok Kwon, Gunho Park, Eunho Yang, Se Jung Kwon, and Dongsoo Lee. No token left behind: Reliable KV cache compression via importance-aware mixed precision quantization. *CoRR*, abs/2402.18096, 2024. doi: 10.48550/ARXIV.2402.18096. URL https://doi.org/10.48550/arXiv.2402.18096.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.

Yao-Ching Yu, Chun Chih Kuo, Ye Ziqi, Chang Yucheng, and Yueh-Se Li. Breaking the ceiling of the LLM community by treating token generation as a classification for ensembling. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1826–1839, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.99. URL https://aclanthology.org/2024.findings-emnlp.99/.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark W. Barrett, Zhangyang Wang, and Beidi Chen. H2O: heavy-hitter oracle for efficient generative inference of large language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6ceefa7b15572587b78ecfcebb2827f8-Abstract-Conference.html.

Justin Zhao, Timothy Wang, Wael Abid, Geoffrey Angus, Arnav Garg, Jeffery Kinnison, Alex Sherstinsky, Piero Molino, Travis Addair, and Devvret Rishi. Lora land: 310 fine-tuned llms that rival gpt-4, A technical report. *CoRR*, abs/2405.00732, 2024. doi: 10.48550/ARXIV.2405.00732. URL https://doi.org/10.48550/arXiv.2405.00732.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark W. Barrett, and Ying Sheng. Sglang: Efficient execution of structured language model programs. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang

(eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/724be4472168f31ba1c9ac630f15dec8-Abstract-Conference.html.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning, acting, and planning in language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=njwv9BsGHF.

APPENDICES

A EXPERIMENTAL SETUP

A.1 TRAINING SETUP

All experiments were conducted on a single node with 8xNVIDIA A100 GPUs (80GB each). Each GPU processed a micro-batch of size 1, and we applied gradient accumulation over 16 steps, resulting in an effective batch size of 128 examples across all devices. This corresponds to approximately 131k tokens per optimization step when the maximum sequence length was 1024, and 262k tokens when it was 2048.

We trained on three datasets: MetaMathQA (40k sampled examples), Evol-Instruct (80k full set), and OASST1 (10k sampled examples). The maximum sequence length was set to 2048 for Evol-Instruct and 1024 for the others. The number of training epochs was 1 for MetaMathQA and Evol-Instruct, and 3 for OASST1.

Optimization was performed using the AdamW optimizer with default hyperparameters (β_1=0.9, β_2=0.999) and a weight decay of 0.01. We used a cosine learning rate decay schedule with a warmup ratio of 0.03, and performed a grid search over learning rates $\{1 \times 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}\}$. No additional regularization techniques (e.g., dropout or gradient clipping) were applied.

For all experiments, we applied low-rank adaptation (LoRA) with a rank of 128 and an α of 256.

A.2 MULTI MODEL INFERENCE SETUP

A.2.1 AGENT WORKFLOW SELECTION AND DESIGN

We designed our experimental setup to evaluate the scalability and performance characteristics of multi model AI agent systems under realistic workload conditions. For this study, we selected two representative agent workflows that exemplify different reasoning patterns commonly deployed in production environments:

ReAct (Yao et al., 2023): This framework synergizes chain-of-thought reasoning with external tool use through an iterative process where agents generate reasoning traces and task-specific actions in an interleaved manner. In the ReAct paradigm, agents alternate between internal reasoning (thoughts) and external actions (tool calls), with each iteration consisting of a thought-action-observation cycle. This pattern is particularly effective for tasks requiring dynamic interaction with external knowledge bases and APIs.

Reflexion (Shinn et al., 2023): This framework reinforces language agents through linguistic feedback, maintaining reflective text in an episodic memory buffer to improve decision-making across multiple trials. Unlike ReAct, Reflexion adds self-evaluation capabilities where agents generate verbal reinforcement cues to assist in self-improvement, storing these experiences in long-term memory for rapid adaptation. This approach enables agents to learn from past mistakes without requiring model fine-tuning, achieving superior performance on complex reasoning tasks.

A.2.2 MULTI MODEL ARCHITECTURE WITH LORA ADAPTERS

To simulate realistic multi-tenant agent deployments, we implemented a novel multi model inference setup where each agent instance operates with its own Low-Rank Adaptation (LoRA) adapter. This configuration mirrors production scenarios where different agents may require specialized model behaviors or domain-specific fine-tuning. Specifically, we matched the number of concurrent agents to the number of LoRA adapters, ensuring that each agent maintains its own parameter space.

This architectural choice has significant implications for system resources:

1. **Memory Overhead**: Each agent maintains its own KV cache throughout multi-turn interactions. With N concurrent agents, the memory requirement scales by a factor of N, as each agent's context must be preserved independently across conversation turns.

2. **Computational Load**: Multi-turn agent requests generate new computational burdens at each interaction step. As agents progress through reasoning chains (ReAct) or reflection cycles (Reflexion), each turn requires fresh attention computations over the accumulated context, leading to quadratic scaling in computational complexity.

A.2.3 Workload Characterization

For workload modeling, we based our input/output distributions and tool-calling patterns on empirical measurements from Kim et al. (2025), which provides comprehensive statistics on real-world agent workflow characteristics. These patterns informed our synthetic workload generation, ensuring our experiments reflect actual deployment scenarios.

A.2.4 Experimental Parameters

We conducted systematic scaling experiments with the following configuration:

Agent Scaling: We evaluated system behavior with 2, 4, and 8 concurrent agents to understand how resource contention and memory pressure evolve with increasing agent density.

Request Rate (QPS):

- For Qwen2.5 14B: Tested at 0.1, 0.2, 0.3, and 0.4 QPS
- For Llama 3.1 8B: Tested at 0.2, 0.4, 0.6, and 0.8 QPS

The different QPS ranges reflect the computational differences between model sizes, with the smaller 8B model capable of sustaining higher request rates.

Throughput Measurement: We measured actual system throughput at the 0.8 QPS configuration to empirically determine system saturation points under peak load conditions.

Batch Size and Latency Dynamics: To understand latency behavior under constrained conditions, we fixed the total request count at 128 while varying QPS. This experimental design differs from unbounded request streams where continuously arriving requests would cause monotonically increasing batch sizes and consequently unbounded growth in 95th percentile latency. Under our fixed-request protocol, we observed that 95th percentile latency initially increases with QPS but eventually saturates at a plateau, indicating the system reaches a steady-state where all requests are being processed within the available compute budget.

This saturation behavior provides critical insights into:

- The maximum sustainable batch size for each agent configuration
- The point at which additional request rate increases no longer impact tail latency
- The effective capacity limits of multi agent systems under resource constraints

A.2.5 Rationale and Implications

Our experimental design captures several critical aspects of production multi agent systems:

1. **Resource Isolation**: By assigning separate LoRA adapters to each agent, we model scenarios where agents require distinct specializations (e.g., different domains, languages, or task-specific fine-tuning).
2. **Memory Pressure**: The multiplicative effect of agent count on KV cache requirements reflects real-world memory bottlenecks in multi-tenant deployments.
3. **Workflow Diversity**: The combination of ReAct's tool-calling patterns and Reflexion's self-improvement cycles represents a broad spectrum of agent behavioral patterns, from reactive tool use to iterative refinement.
4. **Scaling Characteristics**: Our range of agent counts (2–8) and QPS values provides insights into both vertical scaling (request rate) and horizontal scaling (agent parallelism) dimensions.

This setup enables us to quantify the trade-offs between agent autonomy, system throughput, and resource utilization in modern AI agent deployments, providing actionable insights for practitioners deploying multi agent systems at scale.

B Pseudo Algorithm

B.1 Prefill Phase in ICaRus

Algorithm 1: Prefill Phase (Standard Linear Only)

Input: Prompt tokens $P \in \mathcal{V}^N$
Output: First token $y_{\text{prefill}} \in \mathcal{V}$, $\texttt{KV\_CACHE}[1 \dots L]$

$X_1 \leftarrow \text{Embed}(P) \in \mathbb{R}^{N \times d}$
for $i = 1$ **to** L **do**
  $Q_i \leftarrow \text{Linear}(X_i; W_q^i), K_i \leftarrow \text{Linear}(X_i; W_k^i), V_i \leftarrow \text{Linear}(X_i; W_v^i)$
  $Q_i, K_i \in \mathbb{R}^{N \times d_k}, V_i \in \mathbb{R}^{N \times d_v}$
  /* generate KV cache (w. the Logical Encoder) */
  $\texttt{KV\_CACHE}[i] \leftarrow (\boldsymbol{K_i}, \boldsymbol{V_i})$
  $A_i \leftarrow \text{Attention}(Q_i, K_i, V_i) \in \mathbb{R}^{N \times d_v}$
  $X_{i+1} \leftarrow \text{FFN}(\text{AttentionOutput}(A_i)) \in \mathbb{R}^{N \times d}$
$y_{\text{prefill}} \leftarrow \text{Sample}(\text{LMHead}((X_{L+1}[N]))$ // Prefill Result

B.2 DECODE PHASE IN ICARUS

Algorithm 2: ICaRus Linear

Input: $X \in \mathbb{R}^{2\times T\times d}$ // batch=2, seqlen T, hidden size d

X[0]: Input for Logical Encoder (Base model)
X[1]: Input for Logical Decoder (Base model + Adaptive model)

Output: $Y \in \mathbb{R}^{2\times T\times d}$

/* Parallel execution for Base Model and Adaptive Model */
$\boldsymbol{X}_{\text{temp}} \leftarrow \mathbf{Linear}(\boldsymbol{X})$
$X_{\text{temp}}[1] \leftarrow X_{\text{temp}}[1] + \text{AdaptiveLinear}(X_{\text{temp}}[1])$
$Y \leftarrow X_{\text{temp}}$

Algorithm 3: Decode Phase (w. ICaRus Linear)

Input: $y_{\text{prefill}} \in \mathcal{V}$, **KV_CACHE**$[1 \dots L]$

KV_CACHE: Prompt KV cache from Logical Encoder (Base Model)

Output: Generated tokens $Y = (y_{N+1}, y_{N+2}, \dots, y_{N+T})$
(where N is the prompt length, T is the number of generated tokens)

$Input\_Token \leftarrow y_{\text{prefill}}$
for $t = 1 \dots T$ **do**
  $X_1 \leftarrow \text{Embed}(Input\_Token) \in \mathbb{R}^{N\times d}$
  /* Stack hidden states for ICaRus Execution */
  $\boldsymbol{X}_1^{\text{pair}} \leftarrow \mathbf{stack\_batch}(\boldsymbol{X_1}, \boldsymbol{X_1})$ // shape: [2,1,d]
  for $i = 1$ **to** L **do**
    /* KV cache from base model for sharing */
    $\boldsymbol{K}_i^{\text{step}} \leftarrow \mathbf{Linear}(\boldsymbol{X}_i; \boldsymbol{W}_k^i), \boldsymbol{V}_i^{\text{step}} \leftarrow \mathbf{Linear}(\boldsymbol{X}_i; \boldsymbol{W}_v^i)$
    $(K_i^{\text{cache}}, V_i^{\text{cache}}) \leftarrow \text{KV\_CACHE}[i]$
    $K_i \leftarrow \text{concat\_sequence}(K_i^{\text{cache}}, K_i^{\text{step}})$
    $V_i \leftarrow \text{concat\_sequence}(V_i^{\text{cache}}, V_i^{\text{step}})$
    $\text{KV\_CACHE}[i] \leftarrow (K_i, V_i)$
    $Q_i^{\text{pair}} \leftarrow \text{ICaRusLinear}(X_i^{\text{pair}}; W_q^i, A_q^i)$ // shape: [2,1,H,d_k]
    /* Enable attention parallelism via GQA */
    $\boldsymbol{Q}_i \leftarrow \mathbf{concat\_numhead}(\boldsymbol{Q}_i^{\text{pair}}[0], \boldsymbol{Q}_i^{\text{pair}}[1])$ // shape: [1,2*H,d_k]
    $\boldsymbol{A}_i \leftarrow \mathbf{GQA}(\boldsymbol{Q}_i, \boldsymbol{k}_i, \boldsymbol{V}_i)$ // shape: [1,2*H,d_v]
    $\boldsymbol{A}_i^{\text{pair}} \leftarrow \mathbf{transpose\_and\_reshape}(\boldsymbol{A}_i)$ // shape: [2,1,H,d_v]
    $Z_i^{\text{pair}} \leftarrow \text{ICaRusLinear}(A_i^{\text{pair}}; W_o^i, A_o^i)$ // shape: [2,1,d]
    /* FFN: up → act → down (W.ICaRusLinear) */
    $F_i^{\text{pair}} \leftarrow \text{FFN}(Z_i^{\text{pair}})$ // shape: [2,1,d]
  /* use only Adaptive Result */
  $new\_token \leftarrow \text{Sample}(\text{LMHead}(F_{L+1}^{\text{pair}}[1]))$
  $Y \leftarrow \text{concat}(Y, new\_token)$
  $Input\_Token \leftarrow new\_token$