One-Pass to Reason: Token Duplication and Block-Sparse Mask for Efficient Fine-Tuning on Multi-Turn Reasoning

Anonymous ACL submission

Abstract

Fine-tuning Large Language Models(LLMs) on multi-turn reasoning datasets requires N (number of turns) separate forward passes per conversation due to reasoning token visibility constraints, as reasoning tokens for a turn are discarded in subsequent turns. We propose duplicating response tokens along with a custom attention mask to enable single-pass processing of entire conversations. We prove our method produces identical losses to the N-pass approach while reducing time complexity from $O(N^3)$ to $O(N^2)$ and maintaining the same memory complexity for a transformer-based model. Our approach achieves significant training speedup while preserving accuracy. Our implementation is available online[1].

1 Introduction

Recent progress in LLMs has sparked a shift from models that directly generate final responses to those that perform explicit intermediate reasoning before generating responses (referred to as reasoning models). Open-source reasoning models, such as DeepSeek-R1 (Guo et al., 2025), demonstrate high performance on several benchmarks. However, these existing reasoning models were trained primarily on single-turn reasoning data.

While numerous studies have investigated fine-tuning LLMs for multi-turn dialogues to improve coherence, context awareness, tool-calling (Wang et al., 2025; Rebedea et al., 2024), these approaches assume non-reasoning dialogues.

Training LLMs for multi-turn reasoning conversations presents novel challenges in managing token visibility. Following industry-standard practices for multi-turn conversations (OpenAI, 2024; Anthropic, 2025), reasoning models generate internal reasoning tokens, produce a response, and then discard the reasoning tokens from the context in subsequent turns. This creates two fundamental constraints that cannot be addressed with standard multi-turn optimization techniques: (1) **Visibility Constraints**: Reasoning tokens must be visible during generation but hidden from subsequent conversation turns, requiring conditional visibility that static attention masks cannot satisfy. (2) **Position ID Discrepancy**: Response tokens follow reasoning tokens during generation but directly follow human messages in a later context, creating positional misalignment.

While prior works have explored masking techniques and position ID assignments to control information flow and enable selective attention within sequences for various pre-training objectives or efficiency gains (Wang and Hegde, 2024; Du et al., 2022; Raffel et al., 2020), none address the specific challenges of multi-turn reasoning conversations where reasoning tokens must be conditionally visible across turns.

This paper addresses these challenges with two primary contributions. (1) We present a theoretical framework featuring a block-sparse visibility mask and strategic position ID assignment scheme that enables processing an entire multi-turn reasoning conversation in a single forward pass while maintaining training correctness (Theorem 2.1). (2) Due to the absence of a publicly available multi-turn reasoning dataset (to the best of our knowledge), we create and release a novel dataset, $\text{MathChat}_{\text{sync}}$Reasoning, in which each assistant message is augmented with synthetically generated reasoning. (3) We provide comprehensive empirical validation for the proposed framework on Qwen3 models.

Notation. We use $\mathcal{D}$ to denote a multi-turn reasoning dataset where each conversation $c \in \mathcal{D}$ consists of alternating human messages h_i and assistant messages a_i such that $c = (h_i, a_i)_{i=1}^{N}$ for N turns. Each assistant message a_i comprises thinking tokens t_i and response tokens r_i. We denote

[1] https://anonymous.4open.science/r/one-pass-to-reason-F6CB/

$\mathcal{H}_{<i} = (h_j, r_j)_{j=1}^{i-1}$ as conversation history before turn i. For token sequence x, s_x, and e_x represent starting and ending position IDs. The notation $x \rightarrow \mathcal{A}(\cdot)$ indicates sequences that x attends to, and $\mathcal{L}(\cdot)$ denotes language modeling loss (detailed in Appendix A.1).

2 Single Pass Fine-tuning on Multi-Turn Reasoning

In this section, we highlight the challenges associated with fine-tuning language models on multi-turn reasoning datasets. We present an optimized approach to process an entire conversation in a single forward pass. In multi-turn reasoning data, response tokens r_i must attend to reasoning tokens t_i during the generation of a_i. However, these reasoning tokens must not be visible during subsequent generation of assistant messages $a_{j>i}$. As a result, it is not possible to construct a single static attention mask that supports both conditions in a conversation within a single forward pass—a capability that is often feasible with non-reasoning datasets.

2.1 N-Pass Approach

A straightforward solution is to perform a separate forward pass for every turn $(\mathcal{H}_{<i}, h_i, a_i)$ of a given conversation c. While functionally correct, this approach is computationally inefficient: a conversation with N assistant turns results in N separate training examples. Consequently, the effective size of the dataset increases from $|\mathcal{D}|$ to $|\mathcal{D}| \times N$, inflating training time proportionally. Fig. 2 in Appendix B shows attention masks for N-Pass Approach. Fig. 2(a) shows causal attention mask at the time of generation of ith turn response tokens, and Fig. 2(b) shows causal attention mask for ith turn response tokens when they are part of context during $j > i$ turns.

2.2 1-Pass Approach

The primary challenge in applying a single forward pass during training due to discrepancy in the attention behavior of r_i can be illustrated as follows[2]:

$$r_i \rightarrow \begin{cases} \mathcal{A}(\mathcal{H}_{<i}, h_i, t_i) & \text{generation} \\ \mathcal{A}(\mathcal{H}_{<i}, h_i) & \text{context} \end{cases}$$

We can resolve this issue through the following steps:

[2]For ease of understanding, we omit the detail that each token within a token sequence also attends to all its preceding tokens, which must be encoded in the attention mask.

Duplicating response tokens of each assistant message. We duplicate the response tokens of each assistant message so that one sequence (r_i^{out}) is used during generation and attends to its associated reasoning tokens. In contrast, the other sequence (r_i^{in}) is used only as context and does not attend to reasoning tokens.

Custom Attention Mask. Duplication of response tokens makes it possible to have a single attention mask that satisfies visibility constraints. We define a custom masking strategy for each type of token sequence $(h_i, t_i, r_i^{\text{in}}, r_i^{\text{out}})$, ensuring that each token only attends to the appropriate subsequence:

$$h_i \rightarrow \mathcal{A}(\mathcal{H}_{<i}^{in}) \qquad r_i^{in} \rightarrow \mathcal{A}(\mathcal{H}_{<i}^{in}, h_i)$$
$$t_i \rightarrow \mathcal{A}(\mathcal{H}_{<i}^{in}, h_i) \qquad r_i^{out} \rightarrow \mathcal{A}(\mathcal{H}_{<i}^{in}, h_i, t_i)$$

Assigning Consistent Position IDs. After duplication of response tokens, we need to assign consistent position IDs to tokens to maintain the correct relative positions—as if multiple forward passes were performed for each turn in the conversation. If they are assigned sequentially, or the duplicated assistant response tokens share the same position IDs, it will lead to incorrect relative positions. We need a strategic way of assigning position IDs. The following assignment of the first position ID for each token sequence ensures the relative positions are correct and equivalent to N-Pass approach[3]:

$$s_{t_i} = s_{r_i^{in}} = e_{h_i} + 1$$
$$s_{r_i^{out}} = e_{t_i} + 1 \qquad s_{h_{i+1}} = e_{r_i^{in}} + 1$$

Label Mask. Duplication of the response tokens also raises the question of which tokens should be included in the loss calculation. The following label mask outlines the inclusion criteria for each token type:

$$h_i \leftarrow 0 \qquad t_i \leftarrow 1 \qquad r_i^{in} \leftarrow 0 \qquad r_i^{out} \leftarrow 1$$

Fig. 3 in Appendix B shows custom attention mask for ith turn in the 1-Pass Approach. It combines masks for generation and context from the N-Pass Approach into a single mask with position IDs and a label mask consistent with N-Pass Approach.

Theorem 2.1. *Consider a language model with output distributions determined solely by attention*

[3]Position IDs are assigned sequentially based on the order of tokens within each sequence.

patterns, positional encodings, and input representation. For any conversation c as input to the model, the sum of the N-Pass language modeling losses is equivalent to the 1-Pass loss:

$$\mathcal{L}^{\text{1-Pass}}(c) = \sum_{i=1}^{N} \mathcal{L}_i^{\text{N-Pass}}(\mathcal{H}_{<i}, h_i, a_i)$$

Proof is in the Appendix C.1

2.3 Complexity Analysis

We compare the computational complexity of our 1-Pass method against N-Pass approach for transformer-based models with hidden dimension d (Vaswani et al., 2017). Table 1 summarizes the time and memory complexities for a conversation c, where ℓ denotes its characteristic turn length.

	N-Pass	1-Pass
$T(c)$	$O(N^3\ell^2 d)$	$O(N^2\ell^2 d)$
$M(c)$	$O(N^2\ell^2)$	$O(N^2\ell^2)$

Table 1: Time and Memory Complexity for N-Pass and 1-Pass Approach.

The 1-Pass approach yields an asymptotic time complexity improvement of one order in N, offering significant speedups at scale. While it introduces a higher constant memory overhead due to token replication, both methods share the same asymptotic memory complexity. Full derivations are provided in Appendix D.

2.4 Efficient Mask Generation

While our custom attention mask (illustrated in Figure 3) enables single-pass training, generating it involves computing complex visibility patterns across token types and conversation turns. At scale, this computation could become non-trivial, particularly for longer conversations or larger batch sizes. To ensure this remains efficient, we develop an optimized mask generation algorithm that performs all operations on GPU using vectorized tensor operations. Additionally, we simplify the boolean logic for visibility constraints using Karnaugh map reduction, minimizing the number of logical operations required. We provide the complete algorithm in Appendix E.2 for practitioners seeking to implement our method efficiently.

3 Experiments

We evaluate our single-pass fine-tuning on Qwen-3 models (4B, 8B, 32B) with QLoRA (Dettmers et al., 2023). All experiments were run on a 8×H100 instance (CUDA 12.8, PyTorch 2.7.0), with our method implemented in LLaMA-Factory (Zheng et al., 2024) and benchmarked against multi-pass baselines. See Appendix E.3 for experimental setup.

3.1 Dataset Creation

Addressing lack of a public multi-turn dataset with explicit per-turn reasoning, we introduce $\text{MathChat}_{\text{sync}}$Reasoning[4], derived from $\text{MathChat}_{\text{sync}}$ (Liang et al., 2024)[5]. Assistant turns are augmented with explicit reasoning generated using gpt-4.1-mini, conditioned on dialogue history and current assistant response. Refer to Appendix E.1 for more details. All our experiments are conducted on this dataset.

3.2 Experimental Setup

We use FlashAttention2 (FA2) (Dao, 2024) and FlexAttention (Dong et al., 2024) backends. Our 1-Pass method requires a custom attention mask, thus using FlexAttention, as FA2 lacks support for passing custom attention mask; FA2's speed motivates reporting baselines on both for fair comparison. We compare our **1-Pass method** (with response token duplication) against a standard **N-Pass baseline** (requiring N forward passes). Both are evaluated with and without sequence packing[6] (Krell et al., 2022). When packing is enabled, we use llama-factory's `neat_packing` implementation: FA2 baselines rely on position IDs to separate packed sequences (Kundu et al., 2024), while our 1-pass method combines the contamination-free packing mask with our custom attention mask via logical AND.

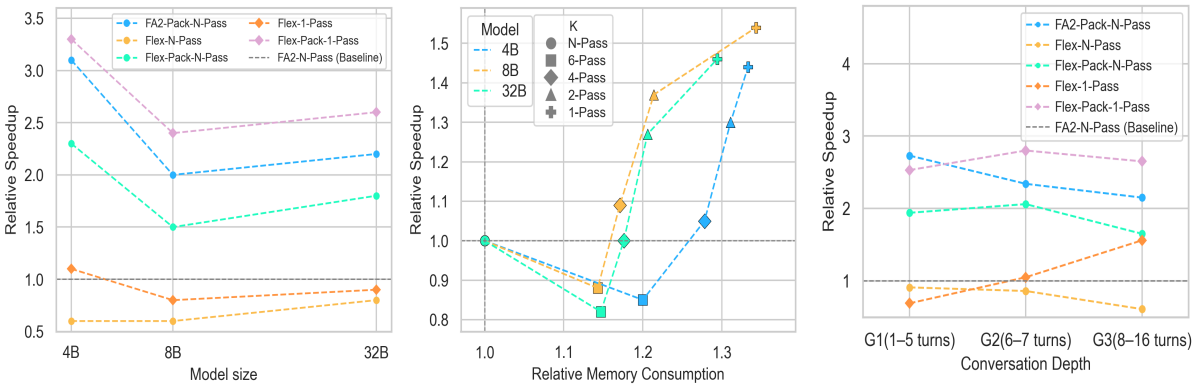
3.3 Results:

Training Speedup. Figure 1a shows training speedups. Our 1-Pass method with packing (Flex-Pack-1-Pass) is 1.05×, 1.21×, and 1.22× faster than FA2-N-Pass baseline with packing (FA2-Pack-N-Pass) on 4B, 8B, and 32B models, respectively. Despite FlexAttention's inherent slowness versus FA2, our method's single-pass efficiency compensates. Compared to N-Pass FlexAttention with packing (Flex-Pack-N-Pass), our Flex-Pack-1-Pass

[4]The dataset is provided in the supplementary material.

[5]LLaMA-Factory and MathChatSync are distributed under Apache-2.0 license, which allows modification and redistribution. Our updates are aligned with the terms of use.

[6]We set the cutoff length to the maximum number of tokens in any datapoint in the dataset for all our experiments.



(a) Speed Analysis across model sizes (relative to FA2-N-Pass)

(b) K-pass Analysis (relative to N-pass) Flex-Pack configuration

(c) Conversation Depth Analysis on Qwen3-8B (relative to FA2-N-Pass)

Figure 1: Training-time experiments

yields $1.44\times$, $1.54\times$, and $1.46\times$ speedups for 4B, 8B, and 32B models, respectively. Without packing, our 1-pass method (Flex-1-Pass) lags FA2-N-Pass baseline for 8B and 32B models. We hypothesize that this is because response-token duplication widens the length disparity between conversations, making the method more sensitive to the absence of packing than the N-Pass baseline. Across all experiments, the 1-Pass variants consume roughly 33% more GPU memory than their N-Pass counterparts.

K-Pass Trade-offs. The 1-Pass and N-Pass approaches represent two extremes: processing the entire conversation in a single pass or in as many passes as there are turns. We therefore also investigate intermediate settings, processing each conversation in K passes. Concretely, we split every dialogue into K contiguous chunks and apply our single-pass mask only to the current chunk, duplicating response tokens and computing loss exclusively for that portion (see Appendix E.4.1 for full details). Figure 1b reveals a speed-memory trade-off for K∈1,2,4,6,N. Our 1-Pass method maximizes speed with ∼33% more memory (vs. N-Pass). K=2 offers a balance (1.30×–1.37× speedups, ∼20% extra memory). Gains diminish for $K > 4$ because, beyond $K = 4$, the extra time incurred by the longer sequences created through token duplication outweighs the savings from processing a few turns together.

Conversation Scalability. The dataset contains conversations with depths from 1 to 16 turns. To analyse the effect of depth, we partition it into three groups: G1 (1–5 turns), G2 (6–7 turns), and G3 (8–16 turns)[7]. Figure 1c shows our Flex-Pack-1-Pass speedups (vs. FA2-Pack-N-Pass) grow with conversation depth ($0.93\times$, $1.19\times$, $1.23\times$ for G1, G2, G3 respectively). A similar trend appears when comparing our method without packing (Flex-1-Pass) to the FA2-N-Pass baseline: speedups of $0.69\times$, $1.05\times$, and $1.56\times$ for G1, G2, and G3, respectively. This supports the theoretical complexity reduction from $O(N^3)$ to $O(N^2)$, as efficiency gains become more pronounced with depth.

These results confirm single-pass training yields significant computational savings, aligning with theoretical advantages, making multi-turn reasoning fine-tuning practical at scale. Please refer Appendix E.4 for comprehensive results of the experiments conducted.

4 Conclusion

We presented an optimized 1-Pass training method for multi-turn reasoning that reduces time complexity from $O(N^3)$ to $O(N^2)$ via strategic token duplication and custom attention mask. Our theoretical analysis confirms loss equivalence with the N-Pass method, enabling efficient training for longer conversations. As multi-turn reasoning becomes central to complex AI tasks, our method offers a scalable and broadly applicable solution. Future work includes exploring adaptive strategies to balance memory-efficiency trade-offs. Additionally, we aim to benchmark performance on latest back-ends such as FlashAttention3 (Shah et al., 2024) and port our masking logic to these faster implementations.

[7]This uneven distribution stems from the $\text{MathChat}_{\text{sync}}$ dataset, which is heavily skewed toward 5–7 turn conversations, a bias that propagates to our reasoning corpus.

Limitations

While our single-pass approach significantly accelerates training for multi-turn reasoning, it also introduces certain constraints. First, duplicating response tokens leads to higher memory usage, which can pose challenges for GPUs with limited capacity. Second, our current implementation supports only FlexAttention, whereas FlashAttention—known for its superior speed—still requires porting our masking logic to make full use of it. Finally, although we have benchmarked N-pass performance on FlashAttention 2, assessing it on the newest iteration, FlashAttention 3, remains future work.

References

Anthropic. 2025. Anthropic extended thinking. https://docs.anthropic.com/en/docs/build-with-claude/extended-thinking. Accessed: 2025-04-17.

Tri Dao. 2024. Flashattention-2: Faster attention with better parallelism and work partitioning. In *Proceedings of the 12th International Conference on Learning Representations*. Version 1.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA.

Juechu Dong, Boyuan Feng, Driss Guessous, Yanbo Liang, and Horace He. 2024. Flex attention: A programming model for generating optimized attention kernels. *arXiv preprint arXiv:2412.05496*. Version 1.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*. Version 1.

Mario Michael Krell, Matej Kosec, Sergio P. Perez, and Andrew Fitzgibbon. 2022. Efficient sequence packing without cross-contamination: Accelerating large language models without impacting performance. *arXiv preprint arXiv:2107.02027*. Version 2.

Achintya Kundu, Rhui Dih Lee, Laura Wynter, Raghu Kiran Ganti, and Mayank Mishra. 2024. Enhancing training efficiency using packing with flash attention. *arXiv preprint arXiv:2407.09105*. Version 6.

Zhenwen Liang, Dian Yu, Wenhao Yu, Wenlin Yao, Zhihan Zhang, Xiangliang Zhang, and Dong Yu. 2024. Mathchat: Benchmarking mathematical reasoning and instruction following in multi-turn interactions. *arXiv preprint arXiv:2405.19444*. Version 1.

OpenAI. 2024. Openai reasoning. https://platform.openai.com/docs/guides/reasoning. Accessed: 2025-04-17.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI. Accessed: 10-11-2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Traian Rebedea, Makesh Sreedhar, Shaona Ghosh, Jiaqi Zeng, and Christopher Parisien. 2024. CantTalkAboutThis: Aligning language models to stay on topic in dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12232–12252, Miami, Florida, USA. Association for Computational Linguistics.

Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. 2024. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. In *Advances in Neural Information Processing Systems*. Version 2.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Franklin Wang and Sumanth Hegde. 2024. Accelerating direct preference optimization with prefix sharing. *arXiv preprint arXiv:2410.20305*. Version 2.

Zezhong Wang, Xingshan Zeng, Weiwen Liu, Liangyou Li, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2025. ToolFlow: Boosting LLM tool-calling through natural and coherent dialogue synthesis. In *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)*, pages 4246–4263, Albuquerque, New Mexico. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

A Background

A.1 Language Modeling Loss

For a token sequence ($\mathcal{H}_{<i}$, h_i, a_i), the language modeling loss (Radford et al., 2018) for assistant message a_i can be expressed as:

$$\mathcal{L}(\mathcal{H}_{<i}, h_i, a_i) = -log(P_\Theta(a_i|(\mathcal{H}_{<i}, h_i)) \quad (1)$$

where language model is parameterized by Θ.

B Masking Schemes for N-Pass and 1-Pass Approaches

Figures 2 and 3 complement the main discussion in Section 2 by illustrating, in greater detail, how attention masks are structured under our N-Pass and 1-Pass training paradigms.

Figure 2 shows the standard causal masks for the N-Pass approach—first, when generating the ith turn's response tokens, and second, when these tokens become part of the context in subsequent turns ($j > i$).

Figure 3 then demonstrates how we incorporate token duplication and a custom visibility mask to enable a single forward pass, while ensuring that each token attends only to the appropriate subsequence. These visualizations clarify the differences between the two methods and underscore why the 1-Pass approach requires careful token and mask design to achieve loss equivalence with the N-Pass baseline.

C Proofs

C.1 Proof for Theorem 2.1

We establish the equivalence by demonstrating that both approaches yield identical probability distributions over sequences, which directly implies equal language modeling losses.

The proof proceeds in three parts: we show that (1) position encodings are equivalent, (2) attention patterns are identical, and (3) the resulting loss functions are mathematically equivalent.

Part I: Position Encoding Equivalence. Consider the position ID assignments for turn i as defined in Section 2.2. In the 1-Pass approach, output tokens receive positions:

$$s_{t_i} = e_{h_i} + 1$$
$$s_{r_i^{out}} = e_{t_i} + 1$$

while input tokens from previous turns $j < i$ receive:

$$s_{r_j^{in}} = e_{h_j} + 1$$
$$s_{h_{j+1}} = e_{r_j^{in}} + 1$$

This assignment ensures that tokens maintain the same relative positional relationships as in the N-Pass approach, where each turn processes tokens sequentially within separate forward passes.

Part II: Attention Pattern Preservation. The custom attention mask defined in Section 2.2 ensures causal dependencies are preserved. For turn i, the attention patterns are:

Output tokens:

$$t_i \to \mathcal{A}\left(\mathcal{H}_{<i}^{in}, h_i\right)$$
$$r_i^{out} \to \mathcal{A}\left(\mathcal{H}_{<i}^{in}, h_i, t_i\right)$$

Input tokens from previous turns $j < i$:

$$h_j \to \mathcal{A}\left(\mathcal{H}_{<j}^{in}\right)$$
$$r_j^{in} \to \mathcal{A}\left(\mathcal{H}_{<j}^{in}, h_j\right)$$

These patterns exactly replicate the causal attention available in the N-Pass approach.

Part III: Loss Function Equivalence. The language modeling loss for turn i in the N-Pass approach is:

$$\mathcal{L}_i^{N\text{-}Pass}(\mathcal{H}_{<i}, h_i, a_i) = -\log P_\theta\left(t_i, r_i \mid \mathcal{H}_{<i}, h_i\right) \quad (2)$$

By the autoregressive factorization:

$$\mathcal{L}_i^{N\text{-}Pass}(\mathcal{H}_{<i}, h_i, a_i) = -\log P_\theta\left(t_i \mid \mathcal{H}_{<i}, h_i\right) - \log P_\theta\left(r_i \mid \mathcal{H}_{<i}, h_i, t_i\right) \quad (3)$$

The total loss across all turns is:

$$\mathcal{L}^{N\text{-}Pass}(c) = \sum_{i=1}^{N} \mathcal{L}_i^{N\text{-}Pass}(\mathcal{H}_{<i}, h_i, a_i) \quad (4)$$

For the 1-Pass approach, the loss is computed as:

$$\mathcal{L}^{1\text{-}Pass}(c) = -\sum_{i=1}^{N} \left[\log P_\theta\left(t_i \mid \mathcal{H}_{<i}^{in}, h_i\right) + \log P_\theta\left(r_i^{out} \mid \mathcal{H}_{<i}^{in}, h_i, t_i\right)\right] \quad (5)$$

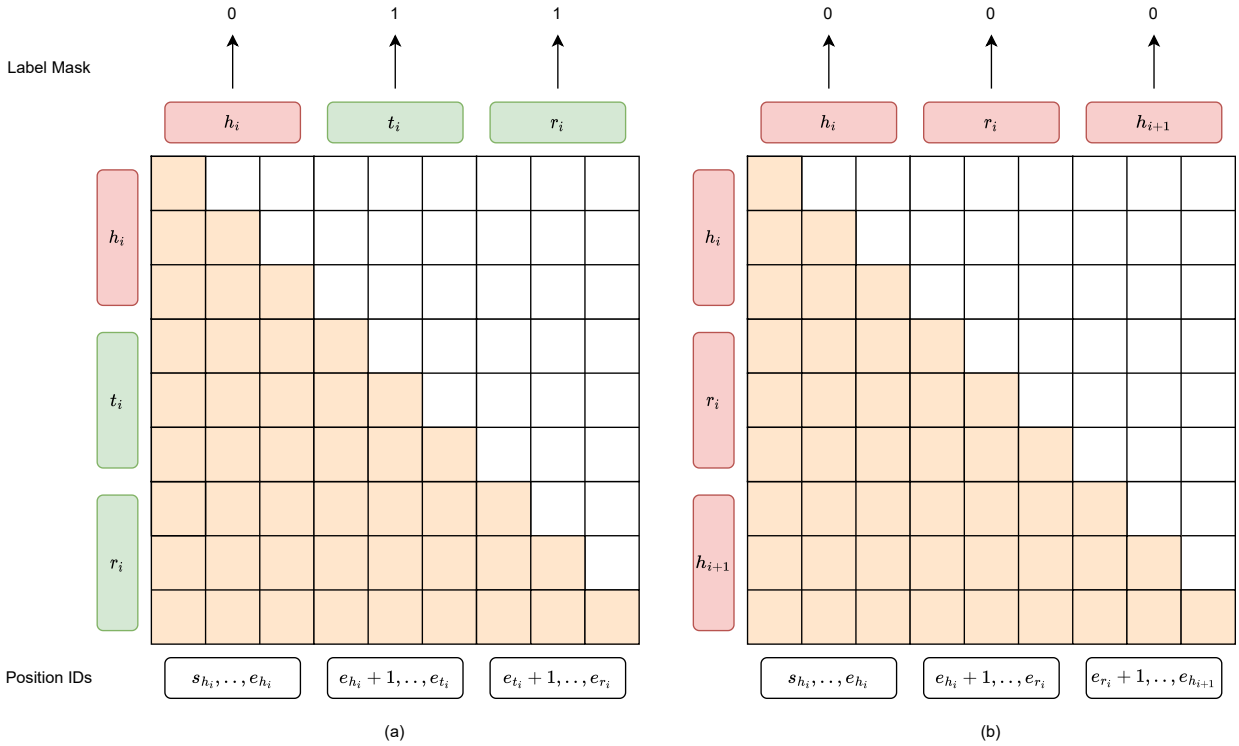


Figure 2: Causal Attention Masks for N-Pass Approach ▢ represents non-zero attention. (a) Attention Mask for generation of response tokens. (b) Attention Mask when response tokens are in context.

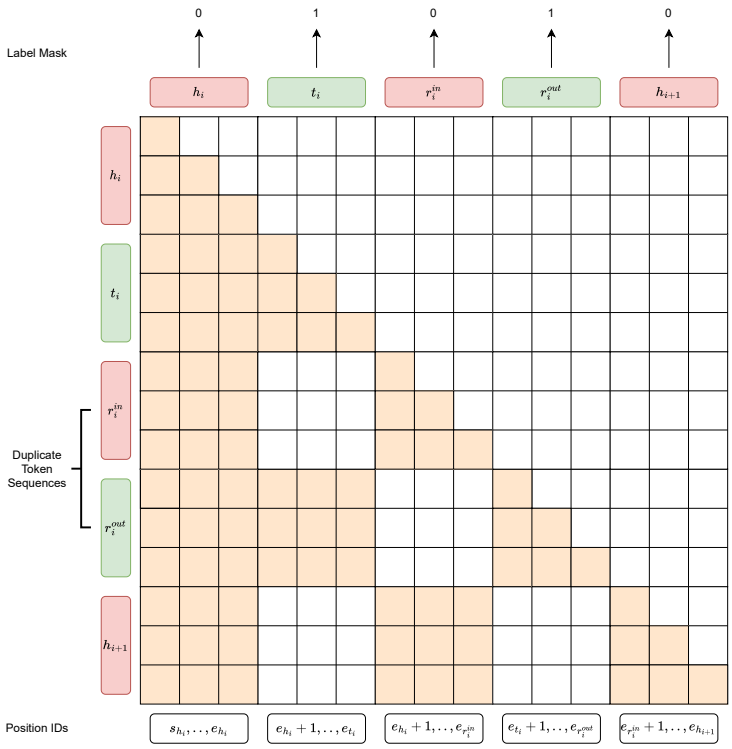


Figure 3: Custom Attention Mask for 1-Pass Approach. ▢ represents non-zero attention.

Key insight: Since $r_j = r_j^{in} = r_j^{out}$ (identical content in different positions) and the position encodings and attention patterns are equivalent as established in Parts I and II, the internal representations are identical. Therefore:

$$P_\theta\left(t_i \mid \mathcal{H}_{<i}, h_i\right) = P_\theta\left(t_i \mid \mathcal{H}_{<i}^{in}, h_i\right) \quad (6)$$

$$P_\theta\left(r_i \mid \mathcal{H}_{<i}, h_i, t_i\right) = P_\theta\left(r_i^{out} \mid \mathcal{H}_{<i}^{in}, h_i, t_i\right) \quad (7)$$

Combining equations (4), (5), (6), and (7):

$$\mathcal{L}^{N\text{-}Pass}(c) = \mathcal{L}^{1\text{-}Pass}(c) \quad (8)$$

D Complexity Analysis

D.1 Input Length

D.1.1 N-Pass Approach

In the N-Pass approach, each turn i is processed in a separate forward pass. The input to the model at turn i is:

$$\mathcal{H}_{<i}, h_i, t_i, r_i$$

because human and assistant response tokens from previous turns remain in the conversation history, while earlier reasoning tokens are discarded.

Let $L_{N\text{-}Pass}$ denote the maximum input length possible for the N-Pass approach for a conversation c. It can be defined by:

$$L_{N\text{-}Pass} = \sum_{i=1}^{N}(|h_i| + |r_i|) + max_{i=1}^{N}|t_i|, \quad (9)$$

which is sum of all the human messages and response tokens for entire conversation and maximum length of thinking tokens across turns. To simplify further, assume:

$$|h_i|, |t_i|, |r_i| \in O(\ell).$$

where ℓ denote the characteristic turn component length, defined as $\ell = P_{95}(|h_i|, |t_i|, |r_i| : i \in [1, N], c \in \mathcal{D})$, where P_{95} is the 95th percentile operator. Then:

$$L_{N\text{-}Pass} \in O\big((2N+1)\ell\big) = O(N\ell). \quad (10)$$

D.1.2 1-Pass Approach

Our 1-Pass approach processes the entire conversation c in a single forward pass. The input length L_{1-Pass} can be calculated as:

$$L_{1\text{-}Pass} = \sum_{i=1}^{N}\big(|h_i| + |t_i| + 2|r_i|\big) \in O\big(4N\ell\big) = O\big(N\ell\big). \quad (11)$$

D.2 Time Complexity Analysis

For a transformer with hidden dimension d and context length n, each layer requires $O(n^2 d)$ operations when $n \gg d$ (Vaswani et al., 2017).

N-Pass Approach: Under the N-Pass approach, each of the N turns requires a forward pass, each operating on $O(L_{N\text{-}Pass}) = O(N\ell)$ tokens. Thus, for conversation c:

$$T_{N\text{-}Pass}(c) \in O\big(N \times (N\ell)^2 d\big) = O\big(N^3\ell^2 d\big). \quad (12)$$

1-Pass Approach: In the 1-Pass approach, all the conversation tokens are given as input at once, thus operating on $L_{1\text{-}Pass}$ tokens yielding a cost of:

$$T_{1\text{-}Pass}(c) \in O\big((4N\ell)^2 d\big) = O\big(N^2\ell^2 d\big). \quad (13)$$

This represents a factor of N improvement in asymptotic complexity, with substantial gains for large N.

D.3 Memory Complexity Analysis

A transformer layer with input context length n has memory complexity $O(n^2)$ assuming $n \gg d$.

N-Pass Approach: Peak Memory requirement for N-Pass approach is at $L_{N\text{-}Pass}$ input. Thus for conversation c:

$$\mathrm{M}_{N\text{-}Pass}(c) \in O\big((2N{+}1)^2\ell^2\big) = O\big(N^2\ell^2\big). \quad (14)$$

1-Pass Approach: Memory requirement for 1-Pass approach can be given by:

$$\mathrm{M}_{1\text{-}Pass}(c) \in O\big((4N)^2\ell^2\big) = O\big(N^2\ell^2\big). \quad (15)$$

Though 1-Pass incurs a higher constant factor due to response token replication, both approaches exhibit identical asymptotic memory complexity.

E Experiments

E.1 Dataset Creation

To enable supervised training with explicit step-by-step reasoning, we construct and release MathChat$_{\text{sync}}$Reasoning along with its generation script. The dataset is obtained by augmenting the original MathChat$_{\text{sync}}$ corpus (Liang et al., 2024) with a synthetically-generated rationale for every assistant turn. The procedure comprises three stages.

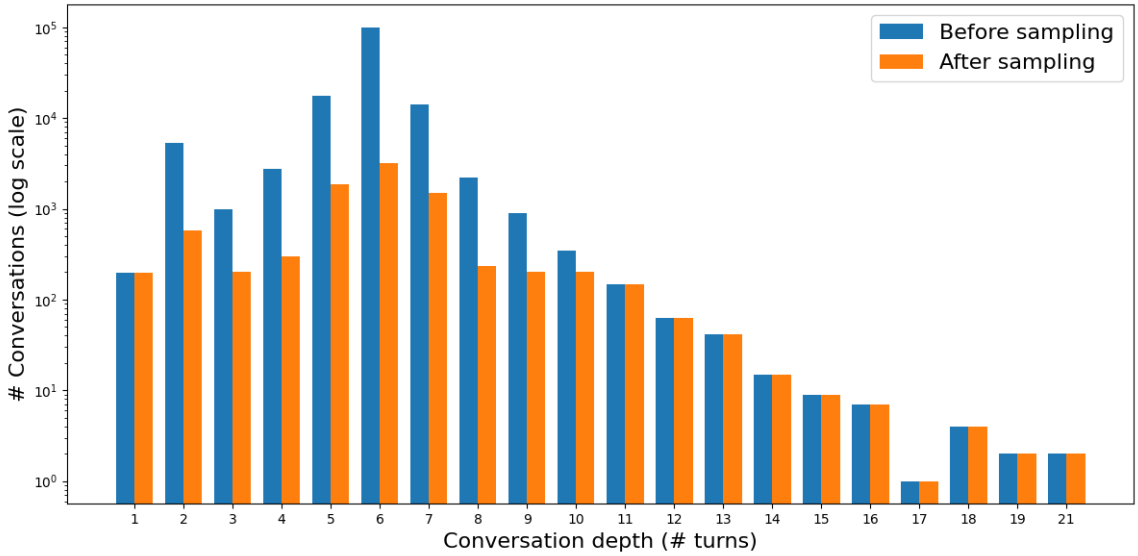


Figure 4: Dataset depth distribution: before vs. after sampling

1. Source corpus. MathChat$_{\text{sync}}$ is a synthetic, dialogue-based mathematics tutoring dataset containing 144,978 conversations with alternating human and assistant messages but no reasoning traces.

2. Depth-balanced sampling. Conversation depth in MathChat$_{\text{sync}}$ is highly skewed toward six-turn dialogues (69 % of all conversations; see Figure 4). To mitigate this bias, we first downsample depth-6 dialogues from 100,443 to 30,000 instances. From the resulting pool we draw a stratified sample of 8,000 conversations.

- For each depth d, we calculate the proportion of the pool that depth represents.
- We allocate to that depth the corresponding proportion of the 8,000-conversation budget, rounding up to the nearest whole conversation.
- If the resulting number is below 200, we raise it to (i) 200 or (ii) the total number of conversations available at that depth, whichever is smaller. This guarantees broad coverage across conversation depths.

The final split contains 8,797 conversations. Figure 4 compares the depth distribution before and after sampling.

3. Reasoning augmentation. For every assistant turn we generate an intermediate reasoning string using gpt-4.1-mini. The model is provided with (i) the dialogue history up to the current human utterance and (ii) the assistant's reply, and is instructed to output only the hidden rationale that could have produced that reply. These rationales are concatenated to the original conversations to form MathChat$_{\text{sync}}$Reasoning.

E.2 Efficient mask generation

Algorithm 1 is an efficient algorithm for generating the custom attention mask required by our 1-Pass training method. The algorithm leverages vectorized GPU operations to compute visibility patterns without explicit loops.

Implementation Notes:

• All operations are performed on GPU using PyTorch's vectorized tensor operations

• Role IDs: 0 = padding, 1 = human, 2 = thinking, 3 = response (first copy), 4 = response (second copy)

• The boolean expression in Step 3 is optimized using Karnaugh map reduction to minimize logical operations

• The algorithm avoids explicit loops by leveraging broadcasting and logical operations

• For CPU tensors, we temporarily move computation to GPU before returning results to the original device

E.3 Experimental Setup

All training runs are initiated using llamafactory-cli in SFT mode. We apply QLoRA with 4-bit NF4 quantization, using a LoRA rank of 32 and a scaling factor of $\alpha = 64$. Training is performed for three epochs with bfloat16 (bf16) precision.

Algorithm 1 Efficient Custom Attention Mask Generation

Require: Role IDs tensor $\mathbf{R} \in \{0, 1, 2, 3, 4\}^{B \times L}$ where B is batch size, L is sequence length
Ensure: 4D attention mask $\mathbf{M} \in \mathbb{R}^{B \times 1 \times L \times L}$

// Step 1: Compute turn IDs via cumulative sum
$\mathbf{R}_{\text{shift}} \leftarrow \text{roll}(\mathbf{R}, \text{shift} = 1, \text{dim} = 1)$
$\mathbf{R}_{\text{shift}}[:, 0] \leftarrow 0$
$\text{turn_increment} \leftarrow (\mathbf{R} \neq 0) \wedge (\mathbf{R} = 1) \wedge (\mathbf{R}_{\text{shift}} \neq 1)$
$\mathbf{T} \leftarrow \text{cumsum}(\text{turn_increment}, \text{dim} = 1)$
$\mathbf{T}[\mathbf{R} = 0] \leftarrow 0$ {Zero out padding positions}

// Step 2: Create base causal non-padding mask
$\mathbf{i} \leftarrow [0, 1, \ldots, L-1]$
$\text{non_pad} \leftarrow (\mathbf{R} \neq 0)$
$\mathbf{M}_{\text{base}} \leftarrow (\mathbf{i}[:, \text{None}] \geq \mathbf{i}[\text{None}, :]) \wedge \text{non_pad}[:, :, \text{None}] \wedge \text{non_pad}[:, \text{None}, :]$

// Step 3: Apply role-specific visibility constraints (K-map optimized)
$\text{turn_equal} \leftarrow (\mathbf{T}[:, :, \text{None}] = \mathbf{T}[:, \text{None}, :])$
$\mathbf{R}_i \leftarrow \mathbf{R}[:, :, \text{None}]; \mathbf{R}_j \leftarrow \mathbf{R}[:, \text{None}, :]$
$\mathbf{M}_{\text{final}} \leftarrow \mathbf{M}_{\text{base}} \wedge \big[(\mathbf{R}_j = 1) \vee (\mathbf{R}_j = 4 \wedge \text{turn_equal})$
$\vee (\mathbf{R}_j = 3 \wedge \mathbf{R}_i \neq 4) \vee (\mathbf{R}_j = 3 \wedge \neg\text{turn_equal})$
$\vee (\mathbf{R}_j = 2 \wedge \text{turn_equal} \wedge \mathbf{R}_i \neq 3)\big]$

// Step 4: Convert to 4D attention weights
$\mathbf{M} \leftarrow \text{where}(\mathbf{M}_{\text{final}}.\text{unsqueeze}(1), 0, -\infty)$
return $\mathbf{M}$

We enable the Liger kernel for improved efficiency. Each GPU processes a batch size of 2, with gradient accumulation over 4 steps. This setup yields an effective batch size of 64 across the 8-GPU node.

E.4 Comprehensive Results

We report the complete numerical results that support the figures in Section 3 in Tables 2, 3 and 4. We report two metrics for every configuration:

- **Throughput** ("samples per sec.") — the average number of *full conversations* processed per second.
- **Peak GPU memory** — the peak memory recorded during training.

E.4.1 Implementing K-Pass Processing

To obtain the results in Table 3 we extend our Optimised 1-Pass scheme to an intermediate K-*Pass* schedule. Assume a conversation contains N assistant turns $(h_1, t_1, r_1), \ldots, (h_N, t_N, r_N)$.

(a) **Chunking the dialog.** We partition the conversation into K contiguous chunks, each containing $\lceil N/K \rceil$ turns (the last chunk may be shorter).

(b) **Selective token duplication.** Within the *current* chunk we apply the same response-token duplication as in Section 2.2: $r_i^{\text{in}}, r_i^{\text{out}}$. All earlier chunks act purely as context and therefore retain their original, non-duplicated responses. This progressively lowers the number of duplicated tokens as K increases, which is the main source of the memory savings reported in Table 3.

(c) **Attention and position IDs.** The custom attention mask and position-ID assignment described in Section 2.2 are applied *only* to the duplicated tokens of the active chunk. Context tokens keep the standard causal mask.

(d) **Loss computation.** The label mask is set to `1` for t_i and r_i^{out} *inside* the active chunk and `0` elsewhere, so each pass trains only on the new turns while reusing earlier content as fixed context.

Conceptually, the K-Pass schedule interpolates between the extremes:

Model Size	Run Setting	Samples per sec.	Peak Memory(GB)	Relative Speedup	Relative Peak Memory
4B	FA2-N-Pass(Baseline)	1.985	9	1.0	1.00
	FA2-Pack-N-Pass	6.241	9	3.1	1.00
	Flex Atten-N-Pass	1.286	9	0.6	1.00
	Flex Atten+Packing-N-Pass	4.550	9	2.3	1.00
	Flex-1-Pass	2.107	12	1.1	1.33
	Flex-Pack-1-Pass	6.552	12	3.3	1.33
8B	FA2-N-Pass(Baseline)	2.307	14	1.0	1.00
	FA2-Pack-N-Pass	4.522	14	2.0	1.00
	Flex-N-Pass	1.365	14	0.6	1.00
	Flex-Packing-N-Pass	3.561	14	1.5	1.00
	Flex-1-Pass	1.736	18.8	0.8	1.34
	Flex-Pack-1-Pass	5.484	18.8	2.4	1.34
32B	FA2-N-Pass(Baseline)	0.601	34	1.0	1.00
	FA2-Pack-N-Pass	1.299	34	2.2	1.00
	Flex-N-Pass	0.465	34	0.8	1.00
	Flex-Packing-N-Pass	1.078	34	1.8	1.00
	Flex-1-Pass	0.521	44	0.9	1.29
	Flex-Pack-1-Pass	1.578	44	2.6	1.29

Table 2: **Throughput and peak memory across execution strategies.** FA2 = FlashAttention 2; Flex = FlexAttention. Pack denotes dynamic sequence-packing; "1-Pass" is our proposed approach. Relative columns are computed with respect to the corresponding FA2–N-Pass baseline.

Model Size	K	Samples per sec.	Peak Memory(GB)	Relative Speedup	Relative Peak Memory
4B	N-Pass(baseline)	4.55	9	1.00	1.00
	6-Pass	3.89	10.8	0.85	1.20
	4-Pass	4.76	11.5	1.05	1.28
	2-Pass	5.91	11.8	1.30	1.31
	1-Pass	6.55	12	1.44	1.33
8B	N-Pass(baseline)	3.56	14	1.00	1.00
	6-Pass	3.13	16	0.88	1.14
	4-Pass	3.87	16.4	1.09	1.17
	2-Pass	4.87	17	1.37	1.21
	1-Pass	5.48	18.8	1.54	1.34
32B	N-Pass(baseline)	1.08	34	1.00	1.00
	6-Pass	0.88	39	0.82	1.15
	4-Pass	1.08	40	1.00	1.18
	2-Pass	1.37	41	1.27	1.21
	1-Pass	1.58	44	1.46	1.29

Table 3: **Speed–memory trade-off as a function of K.** Each dialogue is split into K equal-length chunks that are processed sequentially in a *single* forward/backward pass. $K = N$ corresponds to the per-turn baseline, while $K = 1$ is our single-pass method. All experiments use the FlexAttention backend with sequence packing (Flex-Pack), the configuration that achieved the best overall speed in our primary evaluation.

	Run Setting	Samples per sec.	Peak Memory(GB)	Relative Speedup	Relative Peak Memory
Group 1	FA2-N-Pass(Baseline)	2.54	14	1.00	1
	FA2-Pack-N-Pass	6.93	14	2.73	1
	Flex-N-Pass	2.32	14	0.91	1
	Flex-Packing-N-Pass	4.94	14	1.94	1
	Flex-1-Pass	1.74	18.8	0.69	1.34
	Flex-Pack-1-Pass	6.43	18.8	2.53	1.34
Group 2	FA2-N-Pass(Baseline)	1.02	14	1	1
	FA2-Pack-N-Pass	2.39	14	2.34	1
	Flex-N-Pass	0.87	14	0.86	1
	Flex-Packing-N-Pass	2.10	14	2.06	1
	Flex-1-Pass	1.07	18.8	1.05	1.34
	Flex-Pack-1-Pass	2.86	18.8	2.80	1.34
Group 3	FA2-N-Pass(Baseline)	1.06	14	1	1
	FA2-Pack-N-Pass	2.28	14	2.15	1
	Flex-N-Pass	0.65	14	0.61	1
	Flex-Packing-N-Pass	1.75	14	1.65	1
	Flex-1-Pass	1.66	18.8	1.56	1.34
	Flex-Pack-1-Pass	2.81	18.8	2.65	1.34

Table 4: **Impact of conversation depth (Qwen-3 8B).** Group 1 (1–5 turns), Group 2 (6–7 turns), and Group 3 (8–16 turns). Our 1-Pass approach gains more speed as depth increases, in line with the theoretical $O(N^2)$ vs. $O(N^3)$ complexity gap.

- $K = N$ reproduces the per-turn baseline (no response duplication, minimal memory, maximal passes);
- $K = 1$ is our 1-Pass method (maximum duplication, single pass, fastest).