Curriculum-Guided Reinforcement Learning for Improving Long-Context Logical Reasoning in Foundation Models

Ankit Das
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
ankitd@oregonstate.edu

Larry Rhodes
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
larry.rhodes316@ku.edu

Abstract

Foundation models have demonstrated remarkable capabilities in natural language understanding and generation, yet their performance on long-context logical reasoning tasks remains inconsistent and fragile. Traditional reinforcement learning approaches for fine-tuning these models often rely on uniform reward signals that do not account for the hierarchical nature of reasoning chains. This paper proposes and examines a curriculum-guided reinforcement learning framework designed to systematically improve long-context logical reasoning in foundation models. The framework structures training progression through increasingly complex reasoning episodes, where the curriculum is defined by context length, logical depth, and inter-sentence dependency distance. We analyze the architectural trade-offs inherent in scaling such curricula, including the computational overhead of maintaining long-range attention, the risk of catastrophic forgetting during curriculum transitions, and the need for dynamic reward shaping. Infrastructure considerations for distributed training of large models under curriculum constraints are evaluated, with emphasis on memory-efficient checkpointing and asynchronous policy updates. Robustness and fairness implications are discussed in terms of reward bias amplification, distributional shift in reasoning strategies, and the potential for curriculum design to inadvertently reinforce narrow reasoning patterns. Policy and governance challenges related to deploying curriculum-trained models in sensitive domains such as legal reasoning and medical diagnosis are explored. Empirical case illustrations from recent experiments on multi-hop question answering and mathematical proof generation provide grounding for the conceptual arguments. The paper concludes with a forward-looking perspective on integrating meta-learning principles into curriculum design and on aligning reasoning curricula with human cognitive development trajectories.

Keywords

curriculum learning, reinforcement learning, foundation models, long-context reasoning, reward shaping, distributed training, fairness, robustness.

1. Introduction

The rapid advancement of foundation models has transformed the landscape of natural language processing, enabling systems to perform tasks that range from summarization to code generation with unprecedented fluency. However, a critical limitation persists in the domain of long-context logical reasoning, where models must retain, integrate, and manipulate information across extended sequences of text. Such reasoning is fundamental to applications in legal document analysis, scientific literature synthesis, and multi-turn dialogue systems, yet current models frequently exhibit hallucinations, contradictory conclusions, or simple failures of transitive inference when the context exceeds a few thousand tokens. The problem is compounded by the fact that standard fine-tuning protocols, including reinforcement learning from human feedback, tend to optimize for local coherence rather than global logical consistency. There is a growing need for training paradigms that explicitly structure the learning process to build robust reasoning capabilities over long horizons.

Curriculum-guided reinforcement learning offers a promising avenue for addressing this challenge. Inspired by educational theories that advocate for progressive skill acquisition, curriculum learning organizes training data or tasks in an order of increasing difficulty. When combined with reinforcement learning, the curriculum can shape the exploration and exploitation dynamics of the agent, guiding it through a sequence of reasoning problems that gradually require longer contexts and deeper logical chains. This paper investigates how such a curriculum can be designed, implemented, and deployed for foundation models, focusing on the system-level implications rather than on algorithmic specifics. We examine the architectural modifications needed to support curriculum-based training, the trade-offs between computational cost and reasoning fidelity, and the broader societal consequences of deploying models trained under such regimes.

The structure of the paper is as follows. Section 2 reviews related work in curriculum learning, reinforcement learning for language models, and long-context reasoning. Section 3 presents the proposed framework, detailing the components of a curriculum-guided reinforcement learning system. Section 4 discusses architectural considerations and fundamental trade-offs. Section 5 addresses infrastructure and deployment challenges. Section 6 explores robustness, fairness, and policy implications. Section 7 concludes with future directions.

2. Background and Related Work

Curriculum learning has a rich history in machine learning, originating from the idea that models benefit from being exposed to easier examples before harder ones [1]. Early work demonstrated improved convergence and generalization in neural networks when training data were ordered by difficulty [2]. In the context of reinforcement learning, curriculum approaches have been applied to robotic control tasks, where the agent learns to manipulate objects of increasing complexity or in environments with escalating noise [3]. More recently, researchers have explored automatic curriculum generation, where the training distribution is dynamically adjusted based on the agent's current performance [4]. Despite these advances, the application of curriculum learning to large language models, particularly for long-context reasoning, remains underexplored.

Reinforcement learning has become a standard technique for aligning language models with human preferences, as exemplified by reinforcement learning from human feedback [5]. In this paradigm, a reward model trained on human judgments provides scalar feedback to the policy model, which is updated using proximal policy optimization or similar algorithms [6]. However, these methods typically treat each response as a monolithic action, failing to differentiate between local correctness and global logical structure. Some recent efforts have

introduced hierarchical reinforcement learning for reasoning, where high-level plans guide low-level token generation [7]. The required reference by Dou et al. [13] introduces a plan-then-action framework that provides high-level planning guidance to reinforcement learning for LLM reasoning, which aligns closely with the curriculum approach described here.

Long-context reasoning has been addressed through architectural innovations such as sparse attention mechanisms, memory-augmented transformers, and recurrence within transformer layers [8]. For example, the Transformer-XL and Longformer models extend context windows through segment-level recurrence or sliding windows, but they still struggle with tasks that require precise logical dependencies across widely separated tokens [9]. Retrieval-augmented generation offers an alternative by offloading long-range memory to external knowledge bases, but this introduces latency and retrieval quality dependencies [10]. The curriculum approach does not replace these architectural solutions but complements them by guiding the learning process itself.

The intersection of curriculum learning and reinforcement learning for reasoning in language models has been explored in a few recent works. Some studies have used difficulty scoring based on answer consistency or entropy to reweight training examples [11]. Others have proposed progressive task decomposition, where a complex reasoning problem is broken into subproblems that are learned sequentially [12]. The required work [13] presents a method where a high-level planner generates action sequences that guide a lower-level policy, effectively creating a two-stage curriculum. Our framework builds on these ideas by generalizing the curriculum to multiple dimensions: context length, logical depth, and inter-sentence dependency distance. We also explicitly consider the system-level implications that are often omitted in algorithmic studies.

3. Curriculum-Guided Reinforcement Learning Framework

The proposed framework consists of three interconnected components: a reasoning environment, a curriculum scheduler, and a policy learning module. The reasoning environment is defined by a set of long-context reasoning tasks drawn from domains such as multi-hop question answering, mathematical proof verification, and legal argument analysis. Each task instance includes a context passage of variable length, a query, and a ground-truth answer or reasoning chain. The environment provides a reward signal that measures not only the final answer correctness but also the logical coherence of intermediate steps, as assessed by a trained verifier model or rule-based constraints.

The curriculum scheduler determines the sequence of task instances presented to the agent during training. It operates along three axes of difficulty: context length, logical depth, and inter-sentence dependency distance. Context length refers to the number of tokens in the input passage, which is increased gradually from short passages (e.g., 500 tokens) to long ones (e.g., 16,000 tokens). Logical depth denotes the number of inferential steps required to derive the answer, such as the number of hops in a multi-hop question. Dependency distance captures the average distance between premise statements and their subsequent use in the reasoning chain, measured in tokens or sentences. The scheduler can be hand-designed or learned via meta-reinforcement learning. In either case, the progression is adaptive: if the agent achieves a certain success rate at the current difficulty level, the scheduler advances to the next level; if performance degrades, it may revert to easier instances.

The policy learning module uses a reinforcement learning algorithm, such as proximal policy optimization, to update the foundation model's parameters based on the rewards received. A

key innovation is the incorporation of reward shaping that reflects curriculum progression. Early in training, rewards emphasize local correctness and short-term reasoning steps, while later stages reward global coherence and long-range consistency. This shaping can be implemented through an auxiliary reward that decays with the number of steps taken or that penalizes contradictions found in the reasoning chain. To prevent catastrophic forgetting when transitioning between curriculum levels, we employ experience replay with a buffer that maintains a mixture of past and current tasks, weighted by recency and difficulty.

The interplay between the curriculum scheduler and the policy update requires careful orchestration. The scheduler must decide when to increase difficulty without overwhelming the agent, while the policy must maintain exploration in the face of shifting reward landscapes. This resembles the trade-off between exploitation and exploration in multi-armed bandits, but extended to a continuum of tasks. We advocate for the use of a separate meta-controller that learns the curriculum schedule itself, based on observed learning progress, thereby reducing manual design effort.

4. Architectural Considerations and Trade-offs

Implementing a curriculum-guided reinforcement learning system for long-context reasoning imposes substantial demands on the underlying model architecture. The most immediate challenge is the quadratic scaling of attention with context length. While sparse attention mechanisms alleviate this to some extent, they often introduce information loss that can harm logical reasoning, especially when dependencies cross the boundaries of sparse attention windows. A curriculum that progressively increases context length must therefore be coupled with an attention architecture that maintains fidelity across all distances. One approach is to use a hybrid architecture that combines sliding window attention with global memory tokens that capture long-range dependencies. The memory tokens can be learned jointly during curriculum training, with their capacity expanding as the context grows.

Another architectural trade-off concerns the representation of reasoning steps. Many foundation models generate answers autoregressively, producing tokens one by one without explicit representation of intermediate reasoning states. Curriculum-guided reinforcement learning benefits from architectures that produce structured outputs, such as chain-of-thought sequences or plan-action pairs as in [13]. However, forcing the model to output explicit reasoning steps increases sequence length and computational cost. There is a fundamental tension between the desire for transparency of reasoning and the efficiency of inference. The curriculum can mitigate this by initially requiring explicit reasoning traces only for short contexts and gradually eliminating the explicitness requirement as the model internalizes reasoning patterns.

Memory management during training is another critical architectural consideration. Curriculum approaches often involve backpropagation through entire episodes, which may span thousands of tokens. Gradient computation over such long sequences is memory-intensive, even with gradient checkpointing and model parallelism. A practical architecture must support distributed training where the curriculum scheduler runs on a controller node and multiple worker nodes execute policy rollouts. The workers must synchronize their experience buffers and gradient updates asynchronously to avoid straggler effects. This introduces communication overhead that can negate the benefits of parallelism if not carefully managed. Techniques such as population-based training, where multiple agents explore different curriculum levels simultaneously and share successful strategies, can help balance exploration and exploitation.

The choice of reinforcement learning algorithm also interacts with the curriculum design. Proximal policy optimization is popular for its stability, but its clipped objective may hinder adaptation when the curriculum introduces tasks with very different reward distributions. Algorithms that maintain a trust region, such as trust region policy optimization, may be more robust but are computationally expensive. An alternative is to use advantage-weighted regression, which can be combined with off-policy corrections to reuse data across curriculum levels. The trade-off between sample efficiency and stability must be evaluated empirically across different curriculum structures.

5. Infrastructure and Deployment Challenges

Deploying a curriculum-trained foundation model for long-context reasoning in production environments raises infrastructure concerns that go beyond those of standard fine-tuned models. The curriculum scheduler itself becomes a critical component of the training pipeline, requiring its own compute resources for monitoring agent performance and updating difficulty parameters. In a large-scale training setup spanning hundreds of GPUs, the scheduler must operate in a distributed fashion, potentially using reinforcement learning at the meta-level to adjust the curriculum in real time. This adds complexity to the training infrastructure, which must be fault-tolerant and scalable.

Inference deployment also presents unique challenges. A model trained under a curriculum that progressively increases context length may perform optimally only within the range of context lengths seen during training. If deployed in a scenario where context length far exceeds the maximum training length, performance may drop abruptly. Therefore, infrastructure must include mechanisms for truncation, retrieval augmentation, or context window management that gracefully degrade rather than fail. For real-time applications such as conversational agents or legal document review, latency constraints may necessitate reduced context representations. The curriculum could be adapted to include latency-aware reward shaping, penalizing excessively long reasoning chains, but this introduces a new axis of difficulty.

Model serving infrastructure must support dynamic batching for variable-length inputs. Since long-context reasoning tasks often have highly variable token counts, static batch sizes lead to inefficiencies. Adaptive batching algorithms that group sequences of similar length can mitigate this, but they require careful tuning to avoid introducing unfairness, where longer sequences receive poorer service. Fairness concerns also arise in the distribution of compute resources among different reasoning tasks. A deployment that prioritizes short-context queries over long-context ones may systematically disadvantage users whose queries require extensive background context, potentially reflecting socioeconomic or cultural biases in query length.

Sustainability is another dimension. Training foundation models with curriculum-guided reinforcement learning is energy-intensive because of the need for multiple gradient updates across curriculum stages, the overhead of reward model inference, and the communication costs of distributed scheduling. Researchers and practitioners must consider carbon footprint reduction strategies, such as using curriculum stages that allow early stopping when model performance saturates, or employing once-in-a-lifetime training runs that amortize the cost over multiple downstream tasks. We argue that curriculum design should incorporate energy budgets as part of the reward function, incentivizing models that achieve correct reasoning with minimal computation.

6. Robustness, Fairness, and Policy Implications

The robustness of curriculum-trained models to distributional shift is a central concern. A curriculum that orders tasks by difficulty might inadvertently create a training distribution that is overly structured, causing the model to become brittle when faced with real-world reasoning problems that do not follow the same progression. For example, a legal reasoning task may require simultaneously handling long context, deep logical depth, and high dependency distance, which the curriculum never presented together. To improve robustness, the curriculum should include a final phase that mixes tasks of all difficulty levels in random order, effectively performing a deconfounding step.

Fairness issues can arise from the reward shaping mechanism. If the reward model is trained on human feedback that reflects cultural or educational biases, the curriculum may amplify these biases by rewarding reasoning patterns typical of a particular demographic. For instance, a curriculum that prioritizes logical depth over evidential reasoning may disadvantage individuals from educational systems that emphasize empirical examples over formal deduction. The design of the reward function must be transparent and subject to auditing. Furthermore, the curriculum scheduler itself could be biased if it uses performance metrics that are not equally applicable across different reasoning styles. We propose that fairness audits be integrated into the curriculum validation pipeline, testing whether the trained model exhibits consistent reasoning quality across subgroups defined by language, cultural background, or domain expertise.

Policy implications are significant, especially for high-stakes applications. A foundation model trained with curriculum-guided reinforcement learning for long-context reasoning could be deployed in judicial support systems, where it analyzes case law and recommends verdicts. The regulatory frameworks governing such systems must address questions of liability when the model's reasoning is flawed. Unlike traditional software, the model's reasoning is opaque, and the curriculum trajectory may encode unintended heuristics. Policymakers could require that curriculum designs be documented and made publicly available, along with performance on standardized fairness benchmarks. The European Union's Artificial Intelligence Act, for example, classifies certain AI systems as high-risk and mandates conformity assessments. Curriculum-trained models would likely fall under this category, necessitating independent audits.

Another policy dimension is the potential for dual use. A curriculum that improves long-context reasoning could be applied to generate misleading but logically coherent disinformation, or to analyze classified documents. Governance frameworks should include safeguards such as differential privacy during training to prevent extraction of sensitive reasoning patterns, and usage monitoring to detect deployment in unauthorized domains. International collaboration on standards for curriculum-based training could help mitigate risks while fostering beneficial applications in education, science, and medicine.

7. Conclusion

Curriculum-guided reinforcement learning offers a systematic method for improving long-context logical reasoning in foundation models by structuring training progression along axes of difficulty. This paper has examined the framework at a system level, highlighting the architectural trade-offs between attention efficiency and reasoning fidelity, the infrastructure demands of distributed curriculum scheduling, and the robustness and fairness considerations that arise from reward shaping and curriculum design. The integration of high-level planning

guidance, as exemplified by the work of Dou et al. [13], demonstrates the potential of hierarchical approaches within curricula. However, substantial challenges remain, including the need for adaptive curriculum generation that responds to agent learning dynamics, the computational cost of long-context training, and the ethical governance of models deployed in critical reasoning tasks.

Future research should explore the incorporation of meta-learning principles that allow the curriculum scheduler to learn transferable strategies across different reasoning domains. Additionally, aligning curriculum design with human cognitive development trajectories, such as the progression from concrete to abstract reasoning, could produce models that are more interpretable and trustworthy. Ultimately, the success of curriculum-guided reinforcement learning for long-context reasoning will depend not only on algorithmic advances but also on careful consideration of the socio-technical systems in which these models are embedded.

References

1. Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. Proceedings of the 26th Annual International Conference on Machine Learning, 41–48.
2. Weinshall, D., Cohen, G., & Amir, D. (2018). Curriculum learning by transfer learning: Theory and experiments with deep networks. Proceedings of the 35th International Conference on Machine Learning, 5238–5246.
3. Florensa, C., Held, D., Wulfmeier, M., Zhang, M., & Abbeel, P. (2017). Reverse curriculum generation for reinforcement learning. Proceedings of the 1st Annual Conference on Robot Learning, 482–495.
4. Portelas, R., Colas, C., Hofmann, K., & Oudeyer, P. Y. (2020). Teacher algorithms for curriculum learning of deep RL in continuously parameterized environments. Proceedings of the 2020 Conference on Robot Learning, 83–94.
5. Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30.
6. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.
7. Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., ... & Callan, J. (2023). Active retrieval augmented generation. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 7969–7992.
8. Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.
9. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-XL: Attentive language models beyond a fixed-length context. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2978–2988.
10. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

11. Kumar, A., Ma, T., & Liang, P. (2022). Self-training with selective reweighting improves robustness to distribution shift. Advances in Neural Information Processing Systems, 35, 14122–14136.

12. Xu, Y., Song, G., Qiu, Z., & Sun, M. (2022). Progressive reasoning for complex question answering. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 5423–5435.

13. Dou, Z., Zhao, Q., Wan, Z., Zhang, D., Wang, W., Raiyan, T., ... & Biswas, S. (2025). Plan Then Action: High-Level Planning Guidance Reinforcement Learning for LLM Reasoning. arXiv preprint arXiv:2510.01833.

14. Badrinath, A., & Kochenderfer, M. J. (2023). Reward shaping for safe reinforcement learning. Journal of Artificial Intelligence Research, 78, 571–608.

15. Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1).

16. Han, R., Chen, C., & Sun, X. (2024). Fairness-aware curriculum learning for language models. Proceedings of the 2024 Conference on Fairness, Accountability, and Transparency, 211–225.

17. Rajeswar, S., Courville, A., & Bengio, Y. (2020). Meta-learning for curriculum generation in reinforcement learning. arXiv preprint arXiv:2007.00324.

18. Narasimhan, K., & Barzilay, R. (2021). Learning compositional reasoning via structured curriculum. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 2340–2353.

19. Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. Advances in Neural Information Processing Systems, 32.

20. Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. Harvard Data Science Review, 4(1).