Distilling LLMs into Small Language Models for Medical Reasoning

Efficient AI
Congcong Xu, Shoaib Ahmed, and Anil Telaprolu

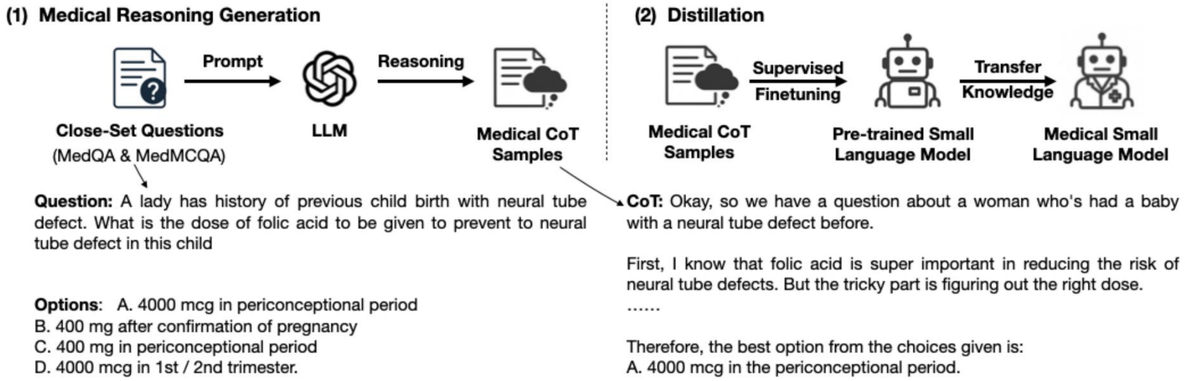
Background

- Large frontier LLMs show strong reasoning but are too costly to deploy widely while small language models provide significant benefits in cost-effectiveness, deployment efficiency, and specialized performance (Wang et al., 2024). Hospitals and medical schools need smaller, privacy-friendly models that can explain their answers.
- However, current small open-weight models have domain shifts especially on meical domain (Guo et al., 2025). They still struggle on hard medical QA benchmarks like MedQA and MedMCQA (Kim et al., 2025)
- Recent works like DeepSeek-R1 (Guo et al., 2025), s1 (Muennighoff et al., 2025), show reasoning can be distilled from LLM to small models effectively.

Aim

- We investigate distilling medical reasoning into a compact LLM to improve accuracy, reasoning ability, and efficiency for medical question answering.

Methodology



- ✓ We build a distillation pipeline where a large “teacher” LLM generates chain-of-thought explanations for medical QA questions.
- ✓ These teacher traces form a curated medical reasoning dataset combining MedQA and MedMCQA with question, step-by-step reasoning, and final answer.
- ✓ A compact open-weight “student” model is fine-tuned on this dataset using supervised learning to imitate the teacher’s answers and reasoning style.
- ✓ Reinforcement learning is then used to further shape the student’s reasoning, rewarding correct, concise, and clinically safe explanations.
- ✓ We evaluate the distilled model on standard medical QA benchmarks, targeting near-teacher accuracy with much lower computational cost.

Results

Model	MedQA	MedMCQA	Avg.
DeepSeek-67B-Chat	57.1	51.7	54.4
DeepSeek-R1-Distill-Llama-8B	20.3	27.1	23.7
Medical-CoT-Distill-Llama-8B	**35.3**	**38.8**	**37.1**

- ✓ Our distilled small language model illustrates better overall medical reasoning and accuracy than the base Llama-8B model.
- ✓ Medical reasoning ability can effectively transfered to small language model by disitllation.
- ✓ Small language models have strong domain-specific accuracy and customization

Future Works

- More complex distillation methods can be designed to effectly transfer the medical reasoning ability.
- Reinforcement learning can be carefully implied into the pipeline.

STEVENS INSTITUTE OF TECHNOLOGY

[1] Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., ... & Tan, Y. (2025). Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. Nature, 645(8081), 633-638.
[2] Muennighoff, N., Yang, Z., Shi, W., Li, X. L., Fei-Fei, L., Hajishirzi, H., ... & Hashimoto, T. B. (2025, November). s1: Simple test-time scaling. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 20286-20332).
[3] Wang, F., Zhang, Z., Zhang, X., Wu, Z., Mo, T., Lu, Q., ... & Wang, S. (2024). A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. ACM Transactions on Intelligent Systems and Technology.
[4] Kim, H., Hwang, H., Lee, J., Park, S., Kim, D., Lee, T., ... & Kang, J. (2025). Small language models learn enhanced reasoning skills from medical textbooks. NPJ digital medicine, 8(1), 240.