先端科学技術研究科　修士論文要旨

所属研究室 (主指導教員)	数理情報学 (池田　和司　(教授))		
学籍番号	2411179	提出日	令和 8年 1月 18日
学生氏名	陳　俊豪		
論文題目	Inducing Robust Medical Reasoning in Small Language Models via Consistency-Aware Reinforcement Learning		

要旨

The development of complex reasoning capabilities in Large Language Model (LLM) has traditionally been associated with massive parameter scaling, often framed as ″emergent abilities″ that manifest discontinuously. However, recent scholarship suggests that this phenomenon may be an artifact of evaluation metrics rather than an inherent property of scale, implying that robust reasoning can be induced in Small Language Model (SLM) through refined learning objectives. In the medical domain, where privacy and resource constraints necessitate efficient on-device models, achieving high-fidelity reasoning in SLM remains a critical challenge. Standard Supervised Fine-Tuning (SFT) often fails to prevent unfaithful reasoning, where models generate correct answers through flawed or hallucinated logic.

In this work, we propose a constraint-driven reinforcement learning framework to enforce reasoning coherence in a 3-billion-parameter model (Qwen2.5-3B). We introduce Output-to-Thought Consistency, a programmatic reward function that penalizes discrepancies between the generated Chain-of-Thought (CoT) and the final conclusion. Leveraging Group Relative Policy Optimization (GRPO), a critic-free algorithm that stabilizes training via group-based advantage estimation, we optimize the model to internalize this consistency constraint without human-annotated process supervision.

Our experiments on the MedQA (USMLE) and PubMedQA benchmarks demonstrate that this targeted intervention yields significant performance gains, surpassing both SFT baselines and larger open-source models (e.g., MedGemma-4B). We observe that the consistency reward induces a stabilization of reasoning length and reward variance, indicative of the model acquiring a structured inference strategy rather than a sudden phase transition. These results suggest that advanced reasoning capabilities in SLM are not emergent mysteries but engineering achievements of rigorous alignment with structural consistency rewards.