Qeios

Peer Review

Review of: "ARWKV: Pretraining Is Not What We Need – An RNN-Attention-Based Language Model Born From Transformer"

Sen Fang[1]

1. North Carolina State University, United States

The paper "ARWKV: Distilling Qwen 2.5 into RWKV-7" presents an interesting but ultimately underperforming approach to distilling Transformer models into RNN architectures. While the three-stage distillation process shows conceptual promise, the technical explanations lack clarity, and the performance results are disappointing. Most critically, comparative analysis reveals that DeepSeek-R1's distillation approach achieves dramatically superior results with fewer parameters – the DeepSeek-R1 model, at just 1.5B parameters, significantly outperforms ARWKV's 7B parameter models across all benchmarks, scoring 90.8% on MMLU (compared to ARWKV's best of 64.77%) and 71.5% on GPQA Diamond (versus ARWKV's 45.5-51.1%). This substantial performance gap undermines the paper's contribution and practical relevance, especially since the authors themselves acknowledge their future aim to "replicate the reasoning capabilities demonstrated by deepseek-R1" – a tacit admission that their current approach falls well short of the state-of-the-art in model distillation techniques.

Declarations

Potential competing interests: No potential competing interests to declare.

doi.org/10.32388/BN9K68