How do LLMs Solve Multi-step Reasoning? An Algorithmic Evaluation

Oliver Eberle
Technische Universität Berlin, Berlin, Germany

Thomas McGee
University of California, Los Angeles, Los Angeles, California, United States

Hamza Giaffar
University of California, San Diego, San Diego, California, United States

Taylor Webb
Microsoft Research, New York, New York, United States

Ida Momennejad
Microsoft Research, New York, New York, United States

Abstract

What algorithms do LLMs actually learn and use to solve problems? Studies addressing this question are sparse, as research priorities are focused on improving performance through scale. Here we introduce a framework for systematic research into the algorithms that LLMs learn and use (AlgEval). Toward this goal, we conducted a graph navigation study that typically requires multi-step search, and evaluated whether Llama-3.1-8B uses classic search algorithms. We formed top-down hypotheses about candidate algorithms (e.g., breadth first, BFS, or depth first search, DFS), and tested these hypotheses via circuit-level analysis of attention patterns and hidden states or representations. We found that 1) Extracting possible sequences processed by the model's layer-by-layer representations did not support either BFS or DFS. 2) Attention patterns showed a cascading shift toward the correct path as the prompt was processed. 3) Projecting node-token representations across layers to a manifold revealed gradual separation of the goal from its competitor in representation space. Overall, our results don't support the idea that the model relies on forming or using an accurate map of the environment, and instead of a step by step search, it seems to rely on more policy-dependent shifts. Future work can connect these findings to failure modes in multi-step reasoning. A rigorous, algorithmic evaluation of how LLMs solve tasks offers an alternative to resource-intensive scaling, potentially enabling more sample-efficient training, performance, and novel architectures.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*