Qeios

Peer Review

Review of: "“Let’s Argue Both Sides”: Argument Generation Can Force Small Models to Utilize Previously Inaccessible Reasoning Capabilities"

Julio C. Rangel[1]

1. Data Knowledge Organization Unit, RIKEN, Wako, Japan

This paper presents **Argument Generation** (AG), a technique intended to help Large Language Models (LLMs) tap into their latent reasoning abilities. The authors compare AG against both zero-shot and chain-of-thought (CoT) prompting across several datasets. They find that AG tends to improve performance, particularly in smaller models, especially in reasoning-based selection among multiple candidate answers. The paper is conceptually sound and motivated, but it would be stronger with clearer definitions, more transparent methodology (especially around implicit assumptions and argument strength), and broader comparison.

Merits:

Innovative Prompting Technique:

AG provides a fresh approach to improving reasoning capabilities. Instead of a single reasoning path, the model generates multiple arguments and then ranks them, which can reveal previously underutilized reasoning.

Applicability to Smaller Models:

The finding that AG particularly helps smaller models is valuable. It suggests that prompt engineering might serve as a compensatory mechanism when parameter counts and training resources are limited.

Comments:

Defining “Deep Reasoning”:

The paper frequently refers to “deep reasoning” without an operational definition. Is it about multi-hop inference, logical consistency, factual verification, or something else? A clearer definition would help frame the problem.

Implicit Assumptions:

The concept of implicit assumptions is appealing, but how exactly are these assumptions derived from the prompt and candidate answers? Are these assumptions meant to represent necessary logical preconditions, or just plausible supporting facts? A worked example that goes through the full algorithm, from input question to assumption generation to final ranking, would make it more convincing.

doi.org/10.32388/1E2ZQN

Techniques to Compare With:

Multi-Agent Debate; these methods also encourage multiple viewpoints or iterative refinement of reasoning steps. Showing how AG's single-pass approach compares against other self-reflective paradigms could highlight its efficiency and where it stands in enhancing LLM reasoning capabilities.

Argument Strength and Ranking Criteria:

The mechanism behind how AG evaluates the "strength" of arguments remains opaque. Is strength judged by internal coherence, alignment with factual knowledge, logical deduction, or a combination thereof? More detail on the ranking criteria could increase trust in the method.

Declarations

Potential competing interests: No potential competing interests to declare.