FMDM @ NeurIPS 2023

GPT-4 Doesn't Know It's Wrong: An Analysis of Iterative Prompting for a Reasoning Problem

Kaya Stechly[2], Matthew Marquez[1], and Subbarao Kambhampati[1]

khstechl@asu.edu
@rao2z
https://github.com/kstechly/gptcolor

Experiment Configuration

Motivation

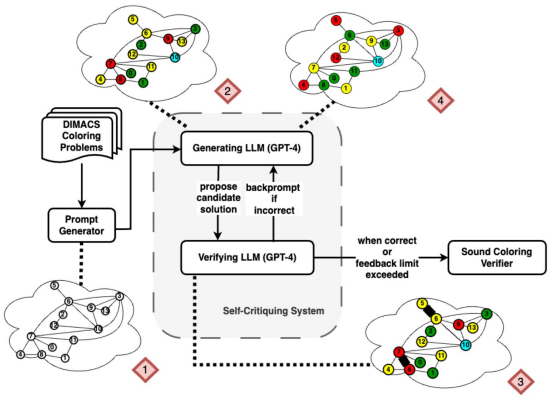
- Much recent work has focused on improving performance by augmenting LLMs with additional tools and feedback as parts of larger systems.
- We build on this with two goals:
  1. to **challenge** claims about reasoning performance, especially when self-critiquing.
  2. to move towards **simplifying** LLM-modulo architectures while maintaining improvements.
- Verification is typically easier than generation. We question **whether this is true** for LLMs.

Domain: Graph Coloring

- Given a planar graph, the problem of finding a minimal vertex labeling where no edge connects vertices with the same label is a **canonical reasoning problem**.
- Solutions are hard to generate, but **critiquing is intuitively trivial**: just list the violated constraints.

Backprompt Architecture

As we are interested in simplification, we keep the overall architecture **deliberately basic** and carefully analyze which aspects were important.



Examining Verification

1. Given some feedback, **how sensitive** are GPT-4's new generations to it? Is it oversensitive to incorrect feedback?
2. Are GPT's verification abilities any better than its solution capabilities in this domain? When they fail, **what kinds of mistakes** do they make?

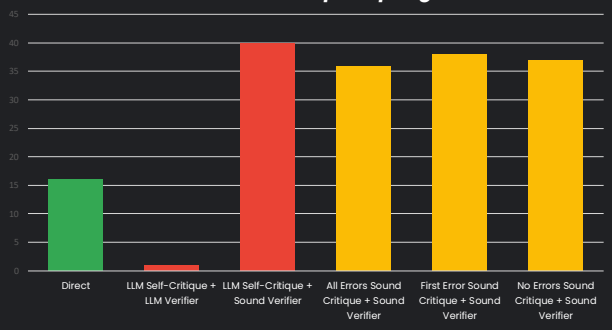
[1] SCAI, Arizona State University
[2] Linguistics, Arizona State University

This research is supported in part by ONR grants N00014-18-1-2442, N00014-18-1-2840, N00014-19-1-2119 and N00014-23-1-2409, and a JP Morgan AI Faculty Research grant.

Results and Analysis

Critique Doesn't Matter, Verification Does

Performance Over Backprompting Schemes



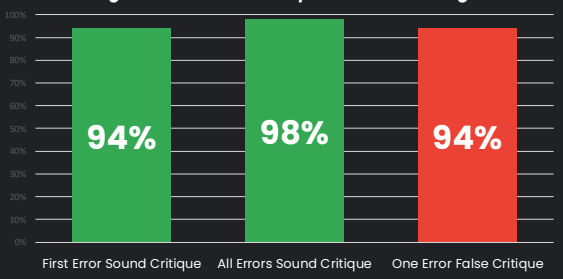
More Samples Are All You Need

	15 Rounds of Critique *(Sound Verifier)*	n=15 Completions
Avg. No. Prompts	~11.9	15
Avg. Token Cost	~17828	~5207
Avg. % Correct	37%	40%

That is to say, the same results can be achieved without giving any critique, just asking for fifteen completions to an unchanged query and evaluating them—a substantially lower token cost.

Oversensitivity to Local Feedback

% Edges Mentioned in Critique that were Changed



Mistakes In Verification

Over Various Colorings