Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks

RAVITEJA BOMMIREDDY

cs23b2011@iiitd

Paper Overview & Motivation

Are Language Models Truly Reasoning?

- Recent LMs like GPT-4 have shown impressive results across many NLP and reasoning benchmarks.
- However, it is unclear whether this success is due to actual reasoning or just memorizing patterns from training data.
- This paper introduces a new evaluation approach using **counterfactual task variants** to test how well LMs generalize reasoning abilities.
- Goal: Distinguish between genuine reasoning and task-s pattern matching.

What Are Counterfactual Tasks?

Testing Reasoning with Counterfactuals

- A task's core logic is preserved, but the world assumptions change (e.g., base-9 instead of base-10 for arithmetic).
- These changes are *reasonable* but *less likely to appear* during training.
- If models truly reason, they should adapt to these minor shifts in setup.
- Counterfactual tasks help isolate reasoning from simple recall.

Methodology & Setup

Designing the Evaluation Framework

- 11 diverse task types covering programming, logic, music, arithmetic, drawing, syntax, and more.
- Each task has two versions: one under default world assumptions, one under a modified (counterfactual) world.
- Evaluated four major LMs: GPT-4, GPT-3.5, Claude, and PaLM-2.
- Used both zero-shot and zero-shot chain-of-thought (CoT) prompting strategies.

Examples of Counterfactual Tasks

Task Samples Across Domains

- **Math:** Perform 2-digit addition in non-standard number bases (8, 9, 11, 16).
- **Code Execution:** Use 1-based indexing instead of the default 0-based in Python.
- **Spatial Reasoning:** Identify object positions in rotated or flipped coordinate systems.
- **Language Syntax:** Identify subject and verb in sentences with shuffled word orders (e.g., VSO, OSV).
- **Music & Drawing:** Retrieve notes in shifted keys, or generate rotated sketches of objects.

Main Results & Observations

Performance Gaps Reveal Weaknesses

- Language models consistently underperform on counterfactual tasks compared to default versions.
- This drop occurs *even when* the model understands the new rules (high CCC scores).
- Chain-of-thought prompting improves performance slightly, but does not eliminate the gap.
- Suggests reliance on memorized procedures tied to default conditions, not generalizable reasoning.

Limitations & Insights

Not Perfect, But Still Revealing

- CCCs aren’t perfect: some failure may still be due to misunderstanding.
- Some counterfactual tasks may be inherently harder (e.g., base-11 math).
- Possible overestimation: counterfactuals might still occur in training data, just rarely.
- Nonetheless, the strong, consistent performance drop is a red flag.

Final Take & Future Directions

My Thoughts & Research Potential

- Smart evaluation strategy — avoids data contamination issues.
- Demonstrates why we can’t equate task performance with real reasoning.
- Opens up space for developing **robust LMs that adapt to unseen worlds**
- Suggests training on more diverse “worlds” or adding grounded, causal modeling.

Any Questions feel free to mail me : cs23b2011@iiitdm.ac.in

THANK YOU