Paper Replication with Anthropic - Tree of Thought

Hari Narayan, 2576394

March 2024

1 Experiment Replication with Anthropic Results

LLM settings were set as per the paper for all experiments.

1.1 Game of 24

Notes on the experimental setup:

- Only the first 30 input samples out of 100 were considered due to time constraints.
- Minimal prompt adaptations to boost structured output predictions for Claude model.

Prompting Strategy	GPT-4	Claude- 2.1
IO Prompt	3.3%	6.6%
CoT Prompt	3.3%	0%
CoT-SC (k=100)	0%	0%
ToT (b=5)	76.6%	50%
IO (Best of 100)	33.3%	10%
CoT (Best of 100)	63.3%	40%

Table 1: **Average Accuracy Scores** for Game of 24 Task: GPT-4 results are adjusted from the paper (not replicated) - Score 1 for correct solution else 0

1.2 Creative Writing

Notes on the experimental setup:

- All inputs as per paper were considered for evaluation.

- Minimal prompt adaptations to boost structured output predictions for Claude models.
- Claude-2.1 model is used as the final output evaluator (To note that the evaluator is not the same across models)

Prompting Strategy	GPT-4	Claude- 2.1
IO Prompt	6.91%	3.95%
CoT Prompt	6.93%	4.62%
ToT	7.56%	3.97%

Table 2: **Average LLM Evaluator Scores** (out of 10) for Creative Writing Task

1.3 Mini-Crossword

Notes on the experimental setup:

- All inputs as per paper were considered for evaluation.
- Minimal prompt adaptations to boost structured output predictions for Claude models.

2 Remarks

1. Within the ToT framework, there is no central entity that is aware of the process/ tree search in its entirety. For example, in the Game of 24 task, the thought generation process is basically a random process, uninformed of the task as well as the evaluations made previously. Empirical evidence of this leading to accuracy loss are plenty.
    - **Possibility for agentic implementation of ToT framework**
2. Effort is involved in framing any task into ToT framework. ToT modelling complexity in the paper:

	GPT-4 (%)			Claude-2.1(%)		
Method	Letter	Word	Game	Letter	Word	Game
IO	38.7	14	0	48.4	19	0
CoT	40.6	15.6	1	44.7	17.9	0
ToT	78	68	20	77.8	50	0

Table 3: **Accuracy Scores** (Normalised number of correct letters/ words and 0/1 score for the entire game) for Mini-Crosswords Task

- Creative Writing < Game of 24 < Mini-Crosswords

3. Is it better to populate the context with the wrong examples or is it better to clear the context so that effects similar to independent thoughts powering performance boost in the line of work related to self-consistency can be produced?

4. Call to look into theoretical studies concerning in-context learning to drive intuition.

3 Evaluation Uncertainties

3.1 Instruction Following

Even though we report the test performance in the context of a particular dataset, the evaluation is much more involved. These are automated evaluations, hence inherently these scores also evaluate the **Instruction following** capabilities of the model. A model that is better at instruction following (lets say structured output prediction) has an advantage when it comes to such automated evaluations.

For example, in the task Game Of 24,
expression = output.strip().split(")[-1].lower().replace('answer: ', ").split('=')[0]
'r': int(sympy.simplify(expression) == 24)
Here the assumptions include:

- That the final algebraic solution response will be in a new line.
- "expression" will not have any other english alphabets other than "answer:".

3.2 Task Comprehension

On a similar note is the implicit evaluation of the **Task Comprehension** capabilities of the model. This might not be an issue with more objective tasks such as Game of 24 or Mini-crosswords but subjective tasks like creative tasks pose some issues.

Here the issue is that with the given input prompt, has the model understood what exactly is the task?

- Observation: Better framing of the task sometimes lead to better results than prompting techniques.