Do Think Tags Really Help LLMs Plan? A Critical Evaluation of ReAct-Style Prompting

Mudit Verma*, Siddhant Bhambri*, Subbarao Kambhampati
School of Computing & AI, Arizona State University
{muditverma, siddhantbhambri, rao} @ asu.edu

ReAct-style prompting for Decision-Making

- ***RQ1: Location of `think'***
  - *Does the agent performance depend on interleaving reasoning trace with action execution?*
- ***RQ2: Content of `think'***
  - *How does the nature of the reasoning trace or guidance information affect the performance of LLM Agents?*
- ***RQ2: Example-Query Similarity***
  - *How does the similarity between the example (problem, solution) and the query (problem, ?), which are present in the prompt, affect LLM Agent performance?*

ReAct

Interact with a household to solve a task. Here are two examples.
<EXAMPLE 1 of same task as QUERY>
You are in the middle of a room ... <Task Description> Your task is to : put some spraybottle on toilet.

Act 1: think : To solve the task, I need to find and take a spraybottle, then put it on toilet.
Obs 1: OK.
Act 2: think : First, I need to find a spraybottle. A spraybottle is more likely to appear in cabinet (1-4), countertop (1), toilet (1), sinkbasin (1-2), garbagecan (1). I can check one by one, starting with cabinet 1.
Obs 2: OK.
Act 3: go to cabinet 1.
Obs 3: On the cabinet 1, you can see cloth 1, a soapboar 1, a soapbottle 1.
...
Act 6: think : Now I find a spraybottle (2), next I need to to take it.
Obs 6: OK.
...
<EXAMPLE 2 of same task as QUERY>
...
<QUERY>
Here is the task. You are in the middle of the room ... Your task is to put some soapbottle in toilet.

Interleaved Think tag
Reasoning Trace / Guidance
Example Task / Query Task

Prompt Variations: RQ1 and RQ2

(RQ1 a) Exemplar-CoT

<EXAMPLE 1>
You are in the middle of a room ... <Task Description>
Your task is to : put some spraybottle on toilet.

Act 1: think : To solve the task, I need to find and take a spraybottle, then put it on toilet. First, I need to find a spraybottle. A spraybottle is more likely to appear in cabinet (1-4), countertop (1), toilet (1), sinkbasin (1-2), garbagecan (1). I can check one by one, starting with cabinet 1.
~~Now~~ Once I find a spraybottle ~~(2)~~, next I need to to take it.
~~Now~~ Once I take a spraybottle (2), next I need to put it in/on toilet (1).
Obs 1: OK.
Act 2: go to cabinet 1.
Obs 2: On the cabinet 1, you can see cloth 1, a soapboar 1, a soapbottle 1.
...

(RQ1 b) Anonymized Exemplar-CoT

<EXAMPLE 1>
You are in the middle of a room ... <Task Description>
Your task is to : put some spraybottle on toilet.

Act 1: think : To solve the task, I need to find and take ~~a spraybottle~~ the object, then put it on ~~toilet~~ the desired location. First, I need to find a ~~spraybottle~~ the object. ~~A spraybottle~~ The object is more likely to appear in ~~cabinet (1-4), countertop (1),toilet (1), sinkbasin (1-2), garbagecan (1).~~ one of the different locations. I can check one by one, starting with ~~cabinet 1~~ the first location.
~~Now~~ Once I find ~~a spraybottle(2)~~ the object, next I need to take it.
~~Now~~ Once I take ~~a spraybottle (2)~~ the object, next I need to put it in/on ~~toilet (1)~~ the desired location.
Obs 1: OK.
Act 2: go to cabinet 1.
Obs 2: On the cabinet 1, you can see cloth 1, a soapboar 1, a soapbottle 1.
...

(RQ2 a) Failure

...
Act 3: open cabinet 2
Obs 3: You open the cabinet 2. The cabinet 2 is open. In it, you see a candle 1, and a spraybottle 2.
~~**Act 4:** think : Now I find a spraybottle 2. Next, I need to take it.~~
Act 4: put spraybottle 2 in/on toilet.
Obs 4: Nothing happens.
...

(RQ2 b) Failure + Explanation

...
Act 3: open cabinet 2
Obs 3: You open the cabinet 2. The cabinet 2 is open. In it, you see a candle 1, and a spraybottle 2.
~~**Act 4:** think : Now I find a spraybottle 2. Next, I need to take it.~~
Act 4: put spraybottle 2 in/on toilet.
Obs 4: Nothing happens.
Act 5: think : Nothing happens because I do not have spraybottle 2.
...

(RQ2 c) Ordering

...
Act 3: open cabinet 2
Obs 3: You open the cabinet 2. The cabinet 2 is open. In it, you see a candle 1, and a spraybottle 2.
~~**Act 4:** think : Now I find a spraybottle 2. Next, I need to take it.~~
Act 4: think : Next, I need to take the spraybottle 2. Now I find a spraybottle 2.
...

(RQ2 d) Placebo Guidance

...
Act 3: open cabinet 2
Obs 3: You open the cabinet 2. The cabinet 2 is open. In it, you see a candle 1, and a spraybottle 2.
~~**Act 4:** think : Now I find a spraybottle 2. Next, I need to take it.~~
Act 4: think : Take a deep breadth and work on this problem step by step.
...

Prompt Variations: RQ3

- **Variation 1: Synonyms - (Domain)** For this variation, we replace the object and location names in the example prompts with their synonyms. For example, ~~spraybottle~~ → aerosolbottle, ~~cabinet~~ → cupboard, and ~~microwave~~ → oven.
- **Variation 2: Problem Instance-level - Instance** We inject instance-level changes to the examples provided in the prompts. We change the goal location in exemplar problem to ensure that it does not match with any of the goal locations in query problem.
- **Variation 3: Problem Level - Both, One, All** All AlfWorld tasks share a large portion of actions (such as exploring cabinets and locations, picking objects etc.). Motivated by this, *One* uses one exemplar of an arbitrarily picked task and the other exemplar of the same task as the query. *Both* uses both exemplars from an arbitrarily picked task. Finally, *All* uses a total of six exemplars (this is the only variation where we provide more than the standard two examples as in ReAct) corresponding to each task.
- **Variation 4: Exploration Strategy - Optimal** In this variation, we provide exemplars which serendipitously take the optimal actions (as if the environment were fully observable) and therefore the example plan is the shortest possible.

Results on Prompt Variations: RQ1, RQ2 & RQ3

Average Success % of LLM for RQ1 and RQ2 on six AlfWorld tasks.

Model / Prompt	Act	ReAct	RQ1		RQ2			
			CoT	Anon. CoT	Placebo	Order	Failure	Explanation
GPT-3.5-Turbo	34.3	27.6	46.6	41	30	28.3	43.3	41.6
GPT-3.5-Instruct	44	50.7	61.9	50.7	41	42.5	47	44.7
GPT-4-0314 (Old)	-	23.3	43.3	33.3	36.6	30	50	36.6
GPT-4-0613 (Latest)	70.0	26.7	40.0	26.6	36.6	30	60	36.6
Claude-Opus	43.3	56.6	50	46.6	30	50	53.3	30

Average Success % of LLM for RQ1 and RQ2 on WebShop tasks.

Model / Prompt	Act	ReAct	RQ1		RQ2		
			CoT	Anon. CoT	Placebo	Failure	Explanation
GPT-3.5-Turbo	1.12	1.04	2.20	1.88	1.52	3.48	3.48
GPT-3.5-Instruct	7.24	7.16	7.52	6.12	7.40	7.20	7.24
GPT-4-0613 (Latest)	8	4	8	8	6	8	8
GPT-4o	4.64	2.24	4.68	4.52	4.08	4.68	4.68
Claude-Opus	4	4	4	2	4	2	4
LLAMA-3.1-8B	1.44	3.16	3.28	3.92	2.04	1.20	2.16

Average Success % of LLM for RQ3 on six AlfWorld tasks. OC: Out of context limit

Model / Prompt	Act	ReAct	RQ3					
			Domain	Instance	Optimal	All	One	Both
GPT-3.5-Turbo	34.3	27.6	1.6	30	20.1	32	28.3	1.6
GPT-3.5-Instruct	44	50.7	47.6	42.5	39.5	OC	17.9	5.2
GPT-4-0314 (Old)	–	23.3	13.3	23.3	50	23.3	16.6	0
GPT-4-0613 (Latest)	70.0	26.7	10.0	20.0	53.3	23.3	20	3.3
Claude-Opus	43.3	56.6	50	46.6	43.3	50	60	6.6

Concluding Remarks

- **Pitfalls of ReAct-style prompting:** With variations on the placement (RQ1) and content (RQ2) of the think tag, we eliminate it as the primary cause of any improvement. Furthermore, slight variations in exemplar tasks (RQ3) lead to a stark decline in success rate, clearly indicating the dependence of performance on the highly curated instance-specific examples by domain experts.
- **Lack of generalization:** Our results on both domains that the Act baseline performs much better than ReAct for several LLMs, which questions on the compatibility of ReAct to newer-age LLMs. ReAct performs worse with newer models as compared to the results they report on PaLM, which is currently decommissioned.
- **Key considerations:** Our findings caution against an uncritical adoption of ReAct-style frameworks for their putative abilities to enhance performance in domains requiring planning. To conclude, we believe that it will be helpful for practitioners and future works to take these results into account, particularly when designing prompts for text-based decision-making problems, and benefit from avoiding putting any efforts into constructing reasoning traces but rather selecting the right examples for subsequent problems.

Acknowledgement

This research is supported in part by ONR grant N0001423-1-2409, and gifts from Qualcomm, J.P. Morgan and Amazon.

https://tinyurl.com/InvestigatingReAct