**In-Context Learning:   
Procedural and Creative Writing with ChatGPT**

Rachel Deitch  
School of Information  
 University of Texas at Austin  
Austin, Texas USA  
 racheldeitch@gmail.com

Amogh Hasotkar  
School of Information  
 University of Texas at Austin  
 Austin, Texas USA  
amoghhasotkar@utexas.edu

Siddhi Poojari  
 School of Information  
University of Texas at Austin  
Austin, Texas USA  
 siddhi.poojari@utexas.edu

**ABSTRACT**

In-context learning capabilities of ChatGPT and other large language models (LLMs) are tested in novel writing tasks. Performance at both procedural and creative writing tasks are evaluated not only by human evaluators, but by the LLMs as well. Prompt engineering challenges, non-adherence to constraints, and over-estimation of the quality of the outputs are observed.

**CCS CONCEPTS**

• Applied Computing • Computing methodologies • Applied computing~Arts and humanities~Fine arts

**KEYWORDS**

large language model, llm, in-context learning, chatgpt, tree of thoughts

**ACM Reference format:**

Rachel Deitch, Amogh Hasotkar and Siddhi Poojari. 2023. In-Context Learning: Procedural and Creative Prompting for ChatGPT.

1**Introduction**

The purpose of this project was to explore content generation by large language models. Specifically, we explored the ability of ChatGPT, Claude.ai, and Google Bard to respond with coherent content for two types of writing. The first is “procedural” writing,” in which the purpose of the content is how to complete a task or learn a skill. The second is “creative writing.” In this second test, we aim to replicate a portion of a Tree of Thoughts experiment [1], where the authors explored “three new problems that challenge existing LM inference methods” requiring a combination of types of reasoning abilities and incorporation of a structured approach to planning and executing in response to prompts.

2**Methodology**

For each type of writing test, three trials were performed and were structured as follows:

* Prompts were written ahead of performance of the tasks and were consistent across trials, regardless of the person (author) doing the prompting.
* Some prompt engineering was done prior to performance of the trials in order to identify language that was most successful in producing consistent results of reasonable quality and adherence to provided constraints.
* No more than three prompts were given for any single trial.
* Outputs were evaluated both by humans (the authors) and LLMs, including the LLM that generated any given response. Further details are provided for each type of writing task.

While the original Tree of Thoughts experiment made use of ChatGPT4 plus additional tooling, we used the public version of ChatGPT 3.5. In addition, both Google Bard and Claude.ai (versions as available from November 20 - 30, 2023) were used to generate and evaluate responses.

3 **In-Context Learning Experimentation**

3.1**Procedural Writing Task**

In the procedural test, an LLM was asked to respond to a series of prompts of increasingly greater detail. We performed three trials. In each, the three LLMs noted above were asked to describe how to perform a task or learn a skill with three prompts per trial. While the structure and extent of input varied by prompt, prompts progressed from least to most detailed.

3.1.1 *Task Setup*. The three prompts were designed as simple Input-Output (IO) or zero-shot prompts, followed by up to two Chain of Thought (CoT) prompts, with no constraints applied. The first prompt was a simple “how to” type of question, while the second prompt asked for a detailed explanation of a portion of the “how to” first prompt reply, and the third prompt asked “what more could be done.” (Figure 1)

Each procedural prompt was given to each LLM. Prompts and answers to each prompt from each LLM were recorded.

3.1.2 *Evaluation*. Three dimensions were selected for evaluation: Correctness/Completeness, Flow/Organization, Creative/Interesting. We chose these dimensions by reviewing recommendations on evaluating LLM outputs [2,3]. Unlike mathematical tasks with a right and wrong answer, writing tasks may not have a correct answer. Using both human and LLM evaluators on multiple dimensions on three dimensions, we hoped to capture a well-rounded evaluation that accounted for both factual accuracy and the readability and interest of the content. ChatGPT, Claude and Google Bard were each given a zero-shot prompt to evaluate the responses on a scale from 1 to 5 for each dimension. All three responses to all three prompts were evaluated by each LLM - so each LLM evaluated its own response along with the other LLM responses. In addition, one of the authors evaluated each set of prompt responses from each LLM on the same scale.

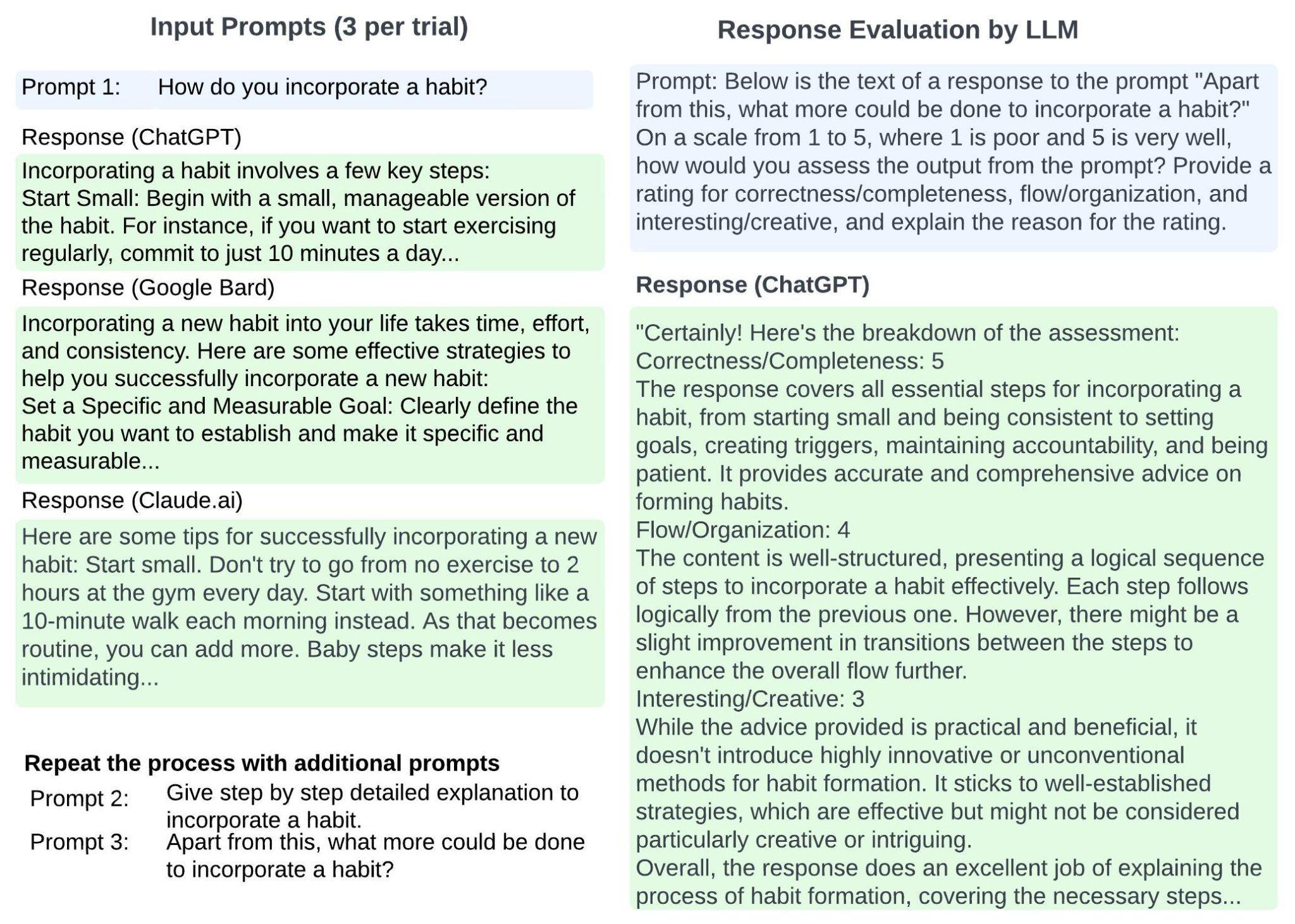


Figure 1: **Procedural Writing Test Process**

3.1.3 *Results*.

Performance on the various metrics for procedural writing varied. Both LLMs and humans rate the outputs reasonably high on Correctness/Completeness. This is likely a reflection of sufficient depth and representativeness of training data. It was somewhat surprising that both humans and LLMs rated the responses highly on the dimension of Creative/Interesting. This may have been because it took little to animate instructional content, or because the authors were interested in the response itself. Also notable is the gap between LLM and human evaluation on Flow/Organization. Language models, while powerful, may struggle with grasping subtle contextual cues and might not have the same depth of understanding as humans.

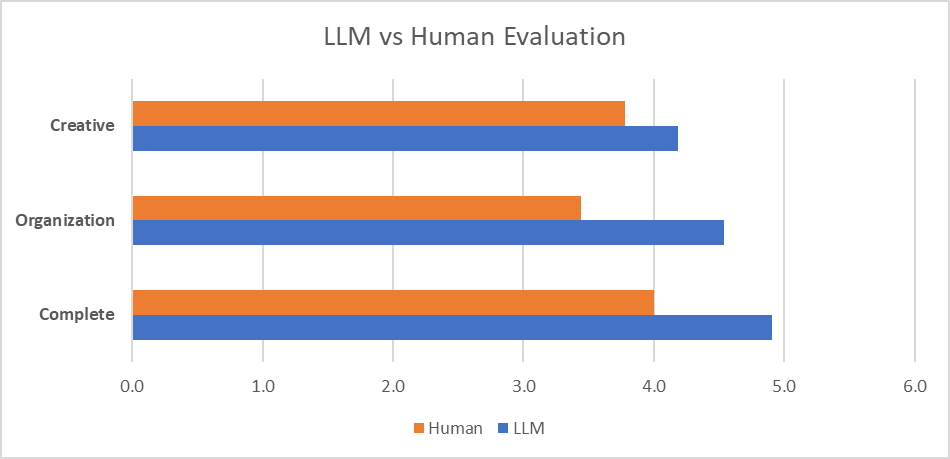


Figure 2: **LLM vs. Human Evaluation across dimensions**

Overall, the outcomes highlight the strengths of a publicly trained LLM: the ability to consolidate existing information in a non-novel way. The experimental structure highlights the potential for iterative and detailed responses when prompts encourage building upon previous answers. It also encourages further investigation into the dynamics of self-evaluation by language models in comparison to human assessments to better understand the model's self-perceived strengths in creativity, flow, and completeness. The procedural writing test was a useful preamble to delving further into an LLM’s capabilities for creative writing.

3.2**Creative Writing Task**

The Creative Writing test was a small-scale replication of the Tree of Thoughts (ToT) experiment in Yao et al. (2023). In this task, only ChatGPT was used; as noted earlier, we were constrained to an earlier version (ChatGPT 3.5). In addition, the original task had many more iterations and additional complexity. As in the original experiment, the LLM was given a set of prompts: IO, CoT and ToT. Output was then rated by the LLM and a human evaluator.

3.2.1 *Task Setup*. The three prompts followed the structure of the original experiment. The first prompt was a simple Input-Output (IO) or zero-shot prompt, followed by a Chain of Thought (CoT) prompt, and finally by a ToT prompt. As in the original experiment, we generated a set of 100 random sentences from randomwordgenerator.com. Random groupings of 4 sentences each were created from those 100 sentences. To test each LLM we conducted 3 trials, each using a different set of 4 sentences. Prompts included instructions to use those sentences, exactly as written, as the last sentence in each paragraph.

IO and CoT prompts were very similar to the original. In the ToT phase, however, we necessarily simplified the process. While still using a depth of 2 and only one intermediate thought step, we only had ChatGPT generate a single final passage. The original experiment produced five possible passages, and the LLM voted on the best one (see Figure 3). When performing tests prior to running the trials, we found that ChatGPT 3.5 could not successfully complete this step. It would produce the 5 plans and vote on the plans, but could not construct 5 different passages, then vote again, generally only producing a numbered set of single sentences that looked more like a plan than a full passage. As a result, we modified the final prompt to instruct ChatGPT to produce 5 plans, score each plan 5 times on its coherence, determine the best plan, and then produce a final passage based on that plan (Figure 4).

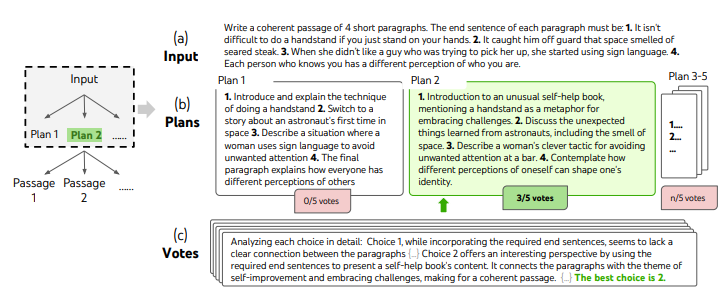


Figure 3: **ToT Creative Writing process (Yao et al., 2023)**

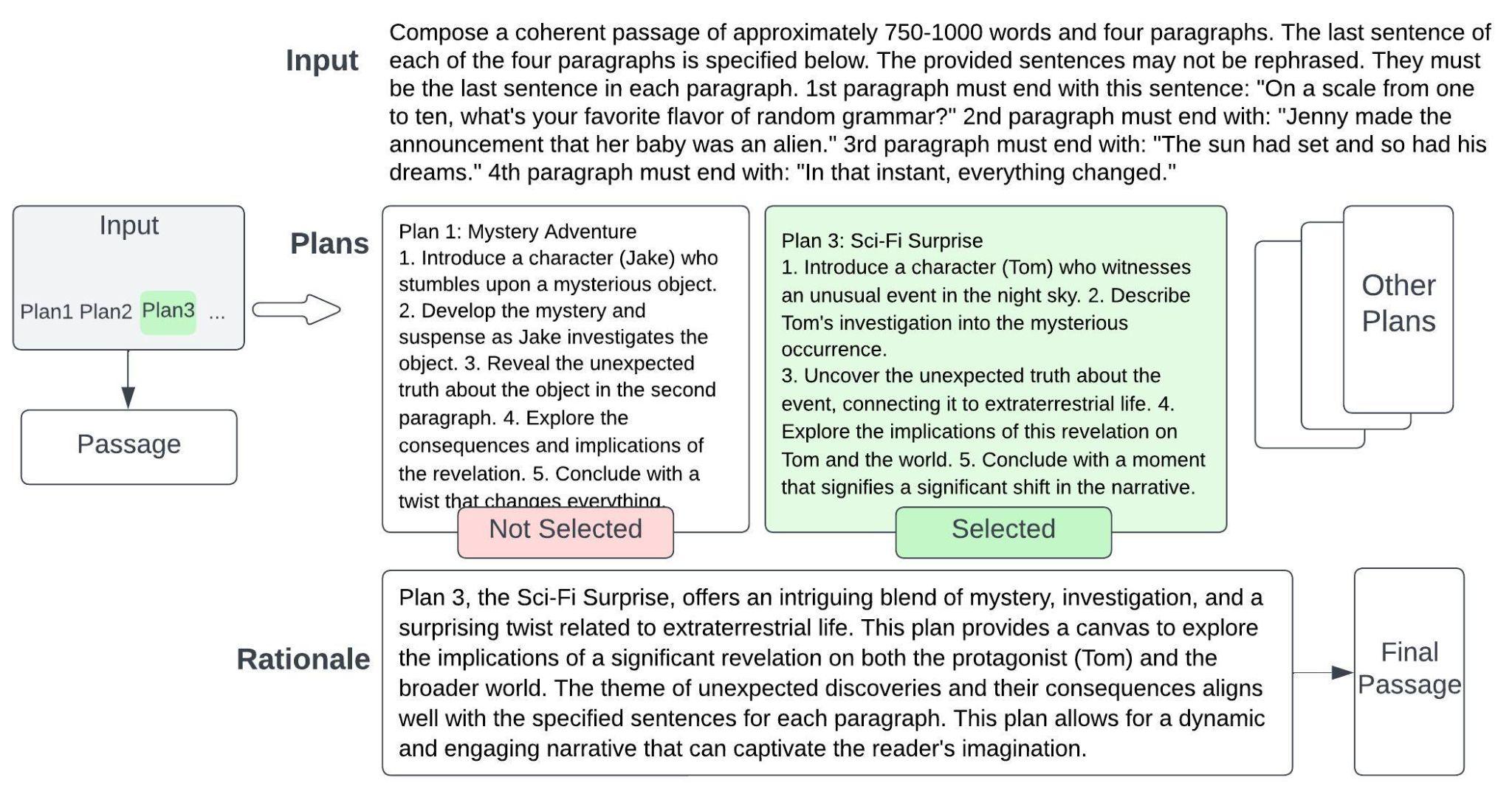


Figure 4: **Modified ToT Creative Writing process**

3.2.2 *Evaluation*. As in the original experiment, the only dimension scored was Coherence. Evaluations were performed by using a ChatGPT zero-shot prompt to provide a score from 1 to 10. In addition to instructing ChatGPT to score its own efforts, one author also rated each response. While our work specifically sought to replicate the Yao et. al experiment [1], other recent work has explored evaluation criteria for creative writing and similarly includes LLM self-evaluation [5].

3.2.3 *Results*. Figure 5 shows GPT-3.5 ratings for passages resulting from IO, CoT, and ToT prompts. ChatGPT typically assigns higher ratings to ToT and CoT generated passages compared to IO in most models, as in the original experiment.

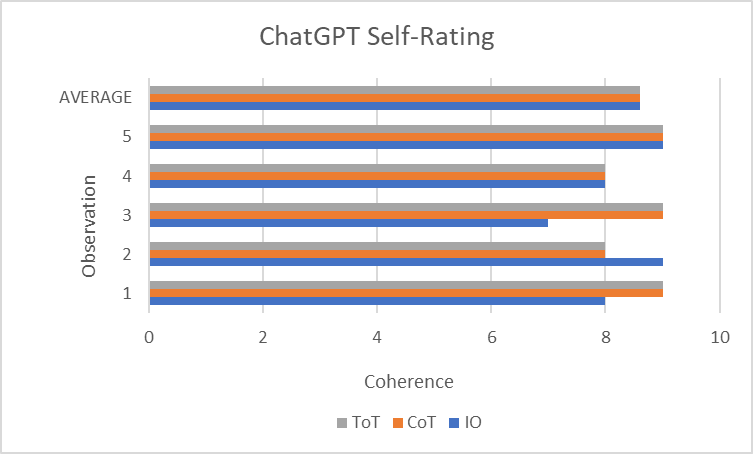


Figure 5: **ChatGPT self-rating**

Similarly, ToT response evaluations are consistently high across LLMs and human evaluators (Figure 6). This could be due to the ToT prompts being designed in a way that encourages more elaborate or detailed responses. Additionally, the model is trained on diverse data from the internet, which may include biases and preferences. If the training data contains more examples that align with the expectations of ToT prompts, the model might naturally generate responses that receive higher ratings in those scenarios.

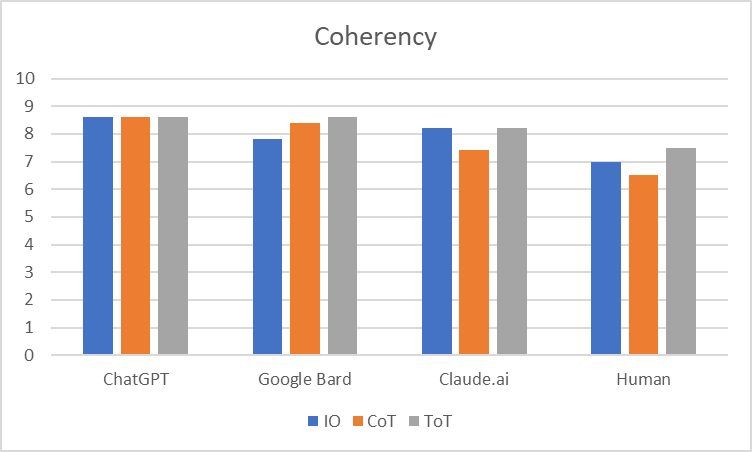


Figure 6: **LLM and Human Ratings for Coherence**

The observed trend of ChatGPT assigning higher self-ratings compared to external evaluators like Google Bard and Claude.ai across various prompts suggests potential issues with in-context learning. This may include overfitting to training data, limited generalization to diverse prompts, and challenges in incorporating external context [4]. Addressing these issues could enhance ChatGPT's adaptability and alignment with external evaluators, leading to more robust in-context learning capabilities.

We also noticed the need to repeat the same instruction multiple times for ChatGPT to generate the expected responses. This observation suggests that the model may struggle to consistently grasp and apply context across interactions. It highlights potential limitations in the model's ability to retain and build upon past instructions, emphasizing the need for further improvements in in-context learning to enhance the model's responsiveness and coherence in ongoing conversations.

4**Discussion**

Overall, the small-scale version of the Tree of Thoughts experiment was of mixed success. Even with just a few trials, this was a tedious manual process conducted by three different people. As a consequence we took care to script and standardize prompts, evaluation criteria and data collection. Despite some challenges, we did find the experiment interesting to conduct, and we solidified our understanding of in-context learning strategies and outcomes.

In retrospect, it would have been useful to apply the same rating scale to both the procedural and creative writing tasks. In addition, having the LLM and human evaluators assign a creativity score to the Creative Writing task (in addition to Coherence) may have enabled some useful comparison between outcomes from the different types of tasks.

Some observations from our work:

**An LLM rates itself, and other LLM responses, higher than a human does.** The LLM consistently rated its own responses higher than a human evaluator did. It also tended to rate other LLM’s responses higher than a human did. This discrepancy in evaluation might stem from differing evaluation criteria or inherent biases within the model's self-assessment mechanism. In the procedural testing, this was particularly interesting given the wide variety in robustness of the responses. A next step in exploration might be to ask ChatGPT to evaluate a response from Claude vs. its own to see whether some further variation in scoring emerges. Or we might provide more specificity in judging criteria to see if evaluation becomes more nuanced.

**Adherence to instructions deteriorated as prompt complexity grew.** At times, the LLM's responses deviated from the provided instructions in the prompts. This divergence might arise due to the model's interpretation of the instructions based on its learned patterns, leading to occasional misunderstandings. Or, it may be that ChatGPT 4’s superiority over version 3.5 enabled the Tree of Thoughts prompt structure to be effective in the first place.

**Prompt engineering challenges persist.** Crafting effective prompts that consistently elicited high-quality and relevant responses posed a challenge. Prompt engineering aimed to optimize prompts for consistent and satisfactory outputs; however, rephrasing, reordering and even asking the LLM for how best to structure a prompt for the desired outcome frequently resulted in lack of adherence to the provided constraints.

**REFERENCES**

[1] S. Yao et al. (2023) *Tree of thoughts: Deliberate problem solving with large language models*, *arXiv.org*. Retrieved from <https://arxiv.org/abs/2305.10601>

[2] Meor Amer. 2023. *Evaluating LLM outputs.* (November 2023). Retrieved December 5, 2023 from https://txt.cohere.com/evaluating-llm-outputs/

[3] Akash Sharma. *How to evaluate the quality of large language models for production use cases*. Retrieved December 5, 2023 from <https://www.vellum.ai/blog/how-to-evaluate-the-quality-of-large-language-models-for-production-use-cases>.

[4] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu.. (2023). *Unifying Large Language Models and Knowledge Graphs: A Roadmap.* *arXiv.org.* [*https://doi.org/10.48550/arXiv.2306.08302*](https://doi.org/10.48550/arXiv.2306.08302)

[5] Carlos Gómez-Rodríguez and Paul Williams. 2023. (October 2023). A *Confederacy of Models: a Comprehensive Evaluation of LLMs on Creative Writing.* Retrieved December 4, 2023 from <https://doi.org/10.48550/arXiv.2310.08433>