Beyond Strawberry – gpt-o1
By John Tan
Link: https://openai.com/index/learning-to-reason-with-llms/

gpt-o1 just dropped.

It is a huge upgrade in reasoning capabilities for LLMs!

Imagine a model that can break down the task into smaller steps to execute at inference time.
Imagine a model that can correct itself and find more promising paths to take if current steps do not work out.
Imagine a model that can repeatedly sample various outputs and consolidate its results.
That is gpt-o1.

Finally, we have a model that can count the number of ‘r’s in strawberry successfully.

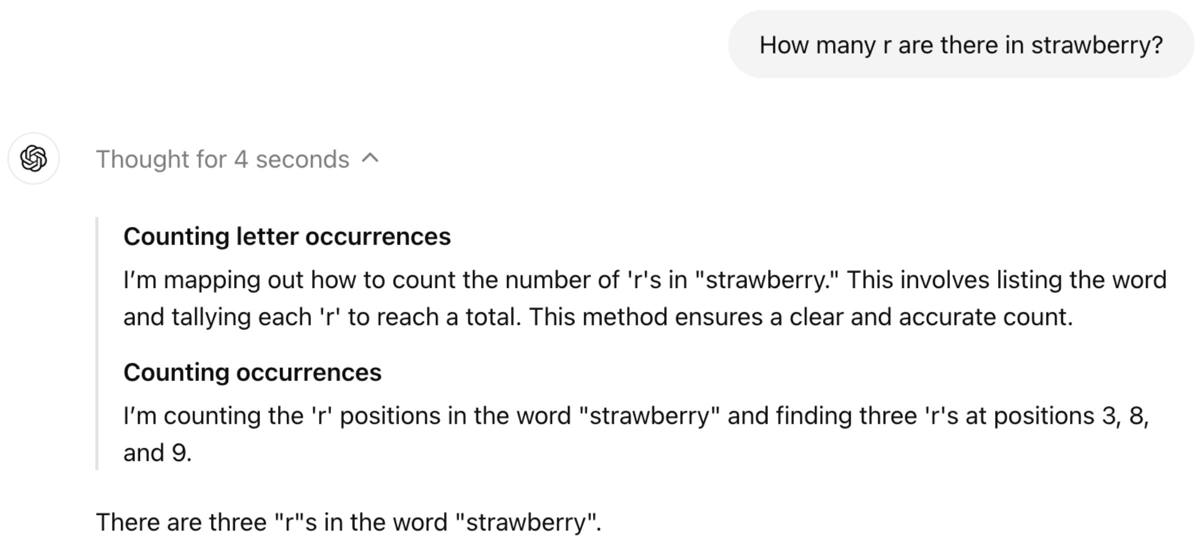


Figure 1 gpt-o1 performance on the strawberry task

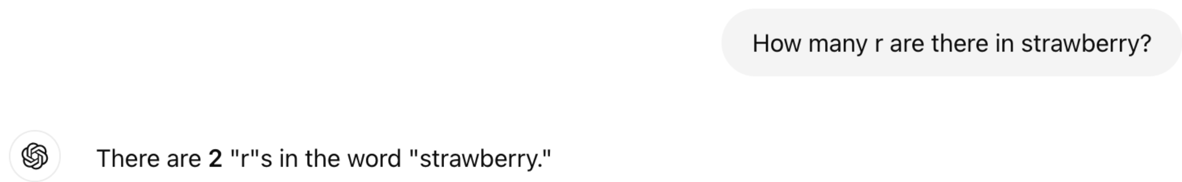
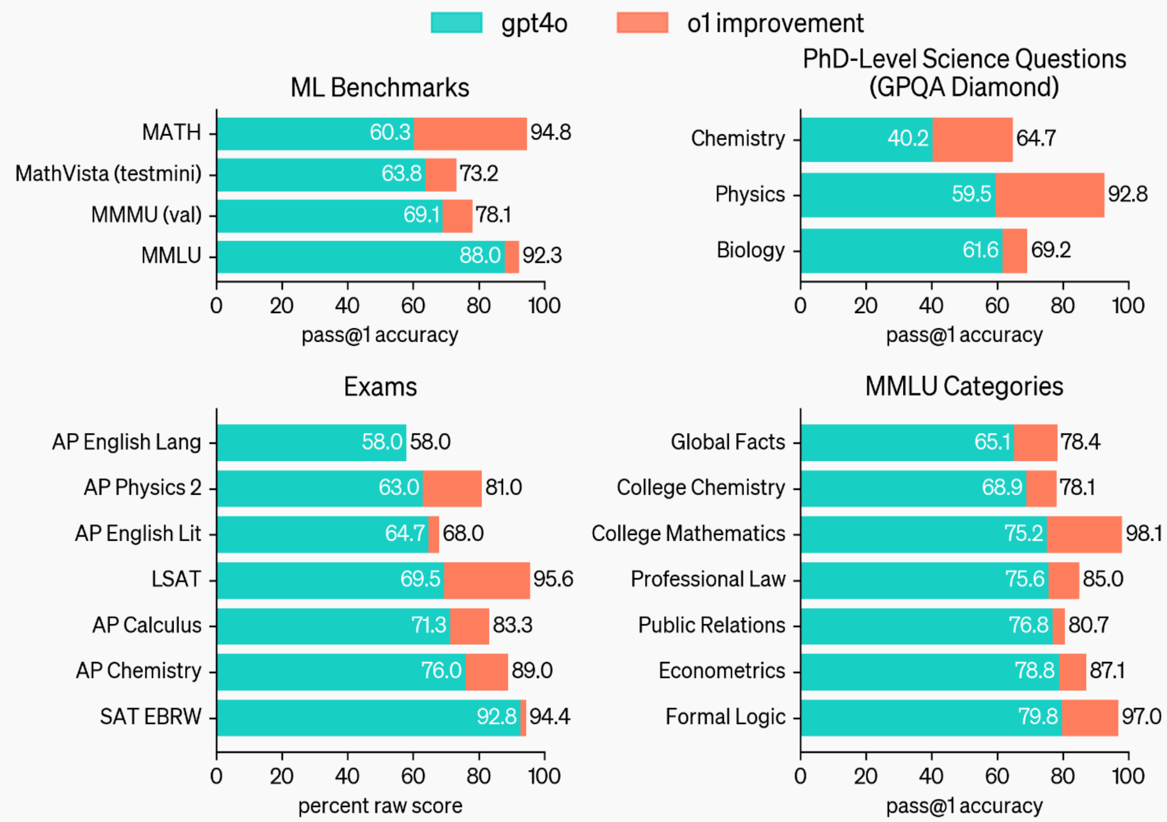


Figure 2 gpt-4o performance on the strawberry task

As can be seen from Fig. 1 and Fig. 2, this new model is able to perform logical tasks like counting the letters in a word, something which gpt-4o struggles with. This is a marked improvement for LLM-based models requiring to perform complex reasoning and understanding, and showcases what this new “Chain of Thought” approach can do.

Superior Performance to gpt-4o



o1 improves over GPT-4o on a wide range of benchmarks, including 54/57 MMLU subcategories. Seven are shown for illustration.

Figure 3 gpt-o1 vs gpt-4o performance on multiple benchmarks

gpt-o1 demonstrates a marked improvement over gpt-4o across multiple benchmarks like PhD-Level Science Questions (GPQA), and multiple examination benchmarks. The performance gains are impressive and highlights that this approach is very promising.

Typical Chain-of-Thought

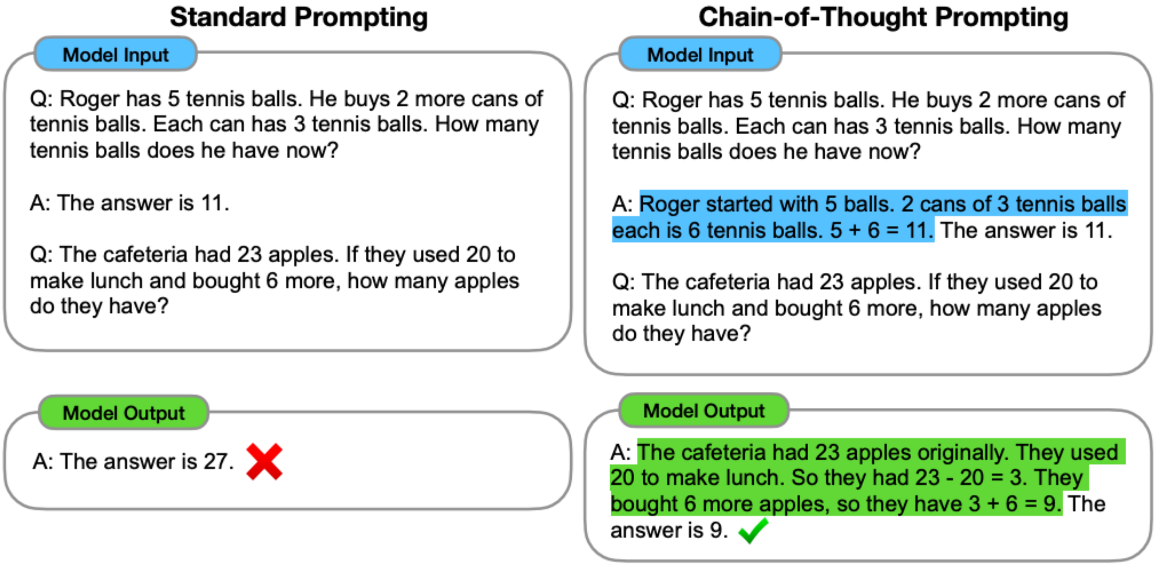


Figure 4 Fig. 1 from Chain of Thought paper: ***https://arxiv.org/pdf/2201.11903***

Let us first understand what is the typical Chain-of-Thought used in the literature. As shown in Fig. 3, the typical Chain of Thought involves giving some examples with extended reasoning steps to encourage the model to explain and elaborate their results. This behaviour can also be replicated to a certain extent with prompts like "Let's think step by step" (https://arxiv.org/pdf/2205.11916).

The key issue of such Chain of Thought approaches is that the model **only generates the answer once**, and if there is any error in the answer, it may not be corrected before displaying to the user.

Agentic Chain-of-Thought (Speculation)

What gpt-o1 appears to do is an agentic version of Chain of Thought (see Fig. 5). An agent will process the given user input / task, and break it down into steps to solve. Each step can probably use tools (although the web version of gpt-o1-preview does not seem to use tools at all), and the output of each step will be used to decide what the next step should be. Finally, we will terminate when the task is completed, or when a maximum number of steps has been taken.

The benefit of doing this agentic chain-of-thought as opposed to traditional chain-of-thought is that the model can potentially **correct itself** and try a different approach if it is not getting the results it needs using one approach.

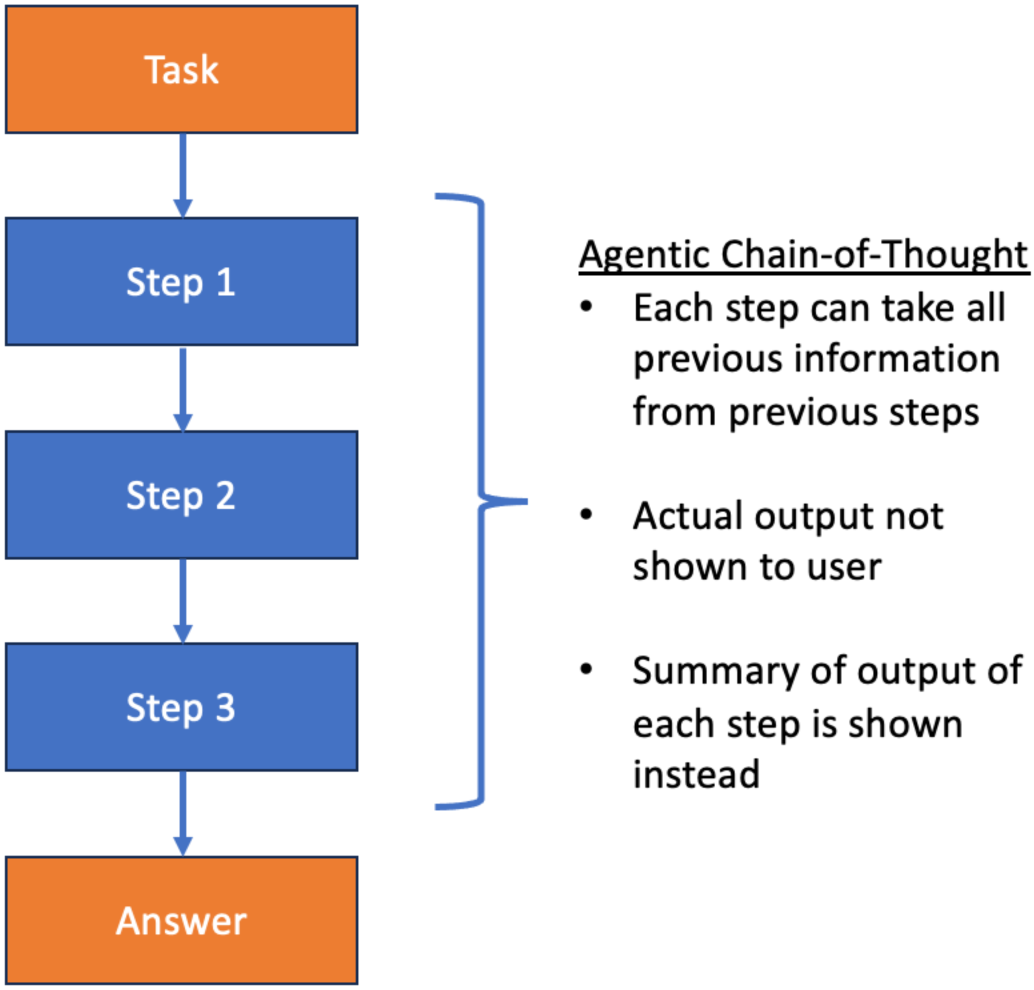


Figure 5 Agentic Chain of Thought

Multiple Sampling and Consolidation

In the benchmarks provided, there is an improved performance with increased sampling and consolidation (see Fig. 6). This sampling and consolidation can be done at each step, or it could be done at the entire agent level.

For example:
```On the 2024 AIME exams, GPT-4o only solved on average 12% (1.8/15) of problems. o1 averaged 74% (11.1/15) with a single sample per problem, 83% (12.5/15) with consensus among 64 samples, and 93% (13.9/15) when re-ranking 1000 samples with a learned scoring function. ```

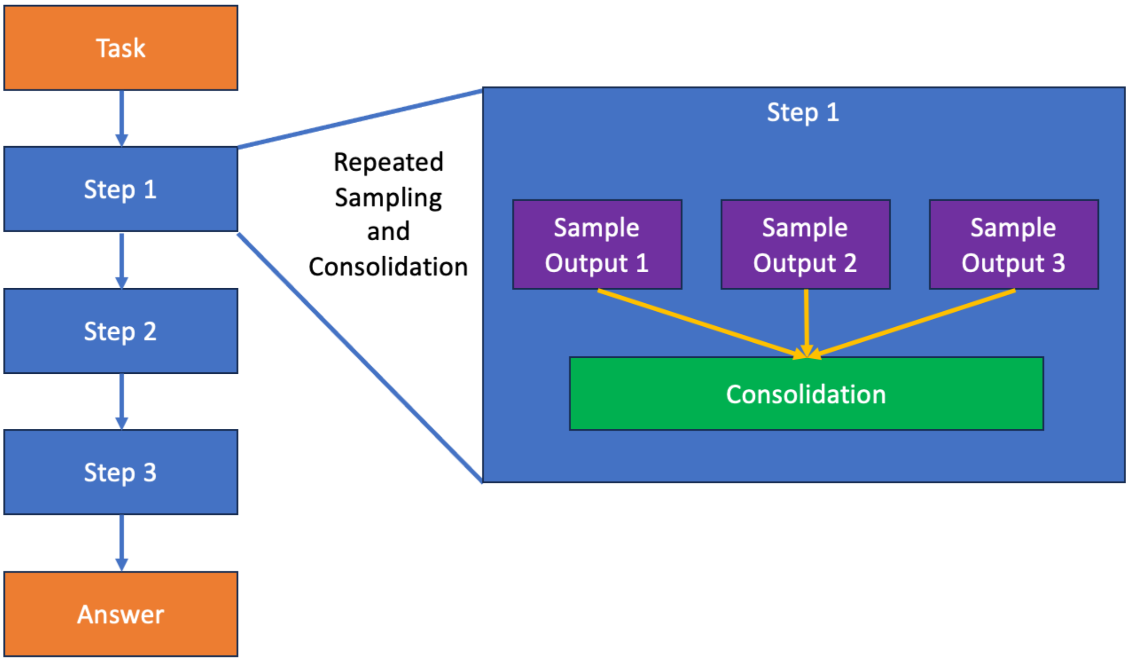


Figure 6 Repeated Sampling and Consolidation

(Speculation) In the web version, it could be 2 different samples used per question, as shown by the two similar steps observed in Fig. 7.
```

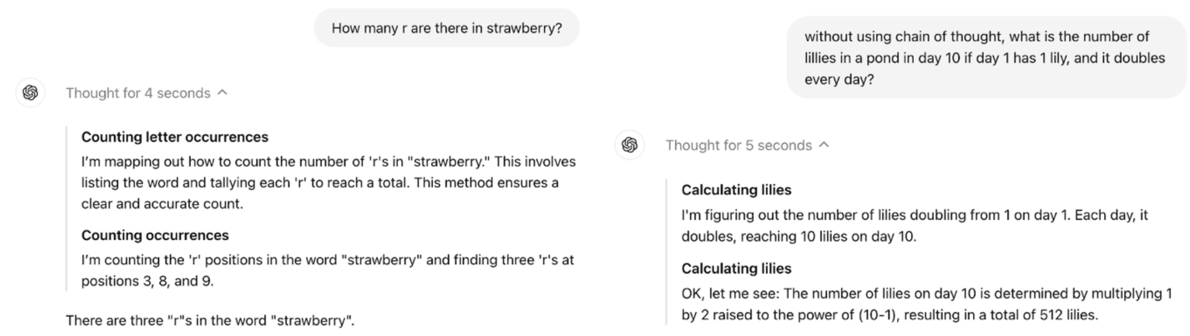


Figure 7 Likely 2 samples used per generation in gpt-o1

The more samples generated, the higher the chance of something being correct, and hence, the more robust the generation.

Conclusion

gpt-o1 is a very promising architecture. The way the model performs chain-of-thought in a sequential, (potentially) agentic manner makes it better able to correct its earlier mistakes.

Furthermore, the use of repeated sampling and consolidation helps to improve reliability of the system.

gpt-o1 is doing something right here, and its performance gains are probably a combination of the more deliberate process of inference-time thinking together with fine-tuning on higher quality reasoning datasets.

The true test of gpt-o1 will be when it is tested on out-of-domain reasoning problems. Only time will tell whether LLM-based reasoning is good enough, or an integration with an external reasoning module is needed.