BEYOND GPT-3

PATHWAYS LANGUAGE MODEL (PALM)

DALL•E 2

Gene Olafsen

REFERENCES

- Google AI Blog: Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance (googleblog.com)
- DALL·E 2 (openai.com)

GPT-3 OVE

- Generative Pre-trained Transformer 3 (GPT-3) is the largest (175 billion trainable parameters), most capable, language model (a statistical tool to predict the next word(s) in a sequence) that leverages deep learning to generate human-like text. But GPT-3 doesn't stop there... it can also generate code, stories, poems, etc. GPT-3 falls into the machine learning category known as natural language processing (NLP).

Where does GPT-3 come from?
OpenAI is an AI research and deployment company.
Whose mission is to ensure that artificial general intelligence benefits all of humanity.

ENTER PALM

- Google's AI research group introduces the Pathways Language Model (PaLM).
- Google Research announced our vision for Pathways, a single model that could generalize across domains and tasks while being highly efficient.
- A 540-billion parameter, dense decoder-only Transformer model trained with the Pathways system.
- PaLM demonstrates the first large-scale use of the Pathways system to scale training to 6,144 chips, the largest TPU-based system configuration used for training to date.

PALM PERFORMS

- PaLM was evaluated on hundreds of language understanding and generation tasks, and found that it achieves state-of-the-art few-shot performance across most tasks, by significant margins in many cases.

TRAINING EFFICIENCY

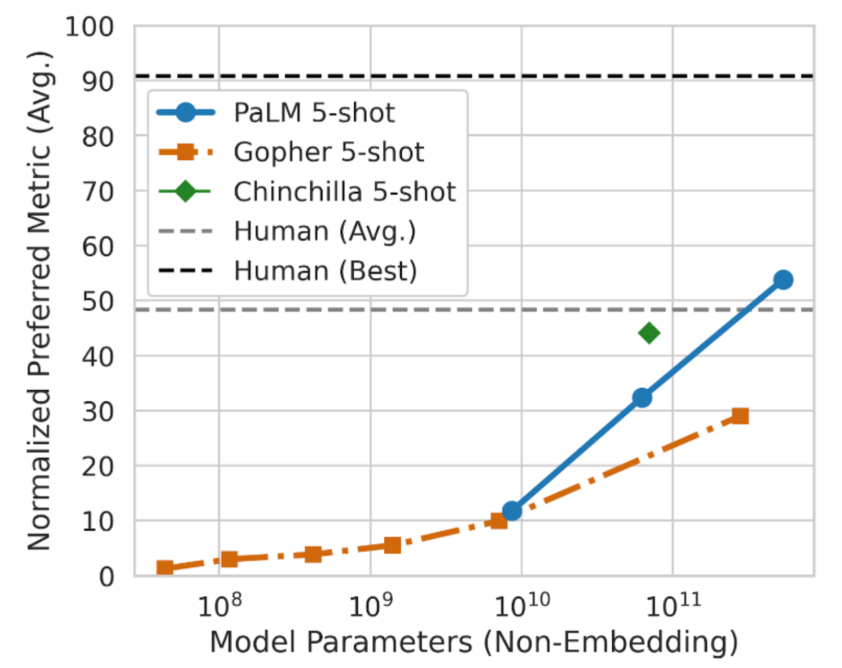
- PaLM achieves a training efficiency of 57.8% hardware FLOPs utilization, the highest yet achieved for LLMs at this scale. This is due to a combination of the parallelism strategy and a reformulation of the Transformer block that allows for attention and feedforward layers to be computed in parallel, enabling speedups from TPU compiler optimizations.

TRAINING CORPUS

- A combination of English and multilingual datasets:
    - High-quality web documents
    - Books
    - Wikipedia
    - Conversations
    - GitHub code
- Google Research also created a “lossless” vocabulary that preserves all whitespace (especially important for code), splits out-of-vocabulary Unicode characters into bytes, and splits numbers into individual tokens, one for each digit.

PERFORMANCE

- PaLM’s performance as a function of scale follows a log-linear behavior similar to prior models, suggesting that performance improvements from scale *have not yet plateaued*.
- PaLM 540B 5-shot also does better than the average performance of people asked to solve the same tasks.

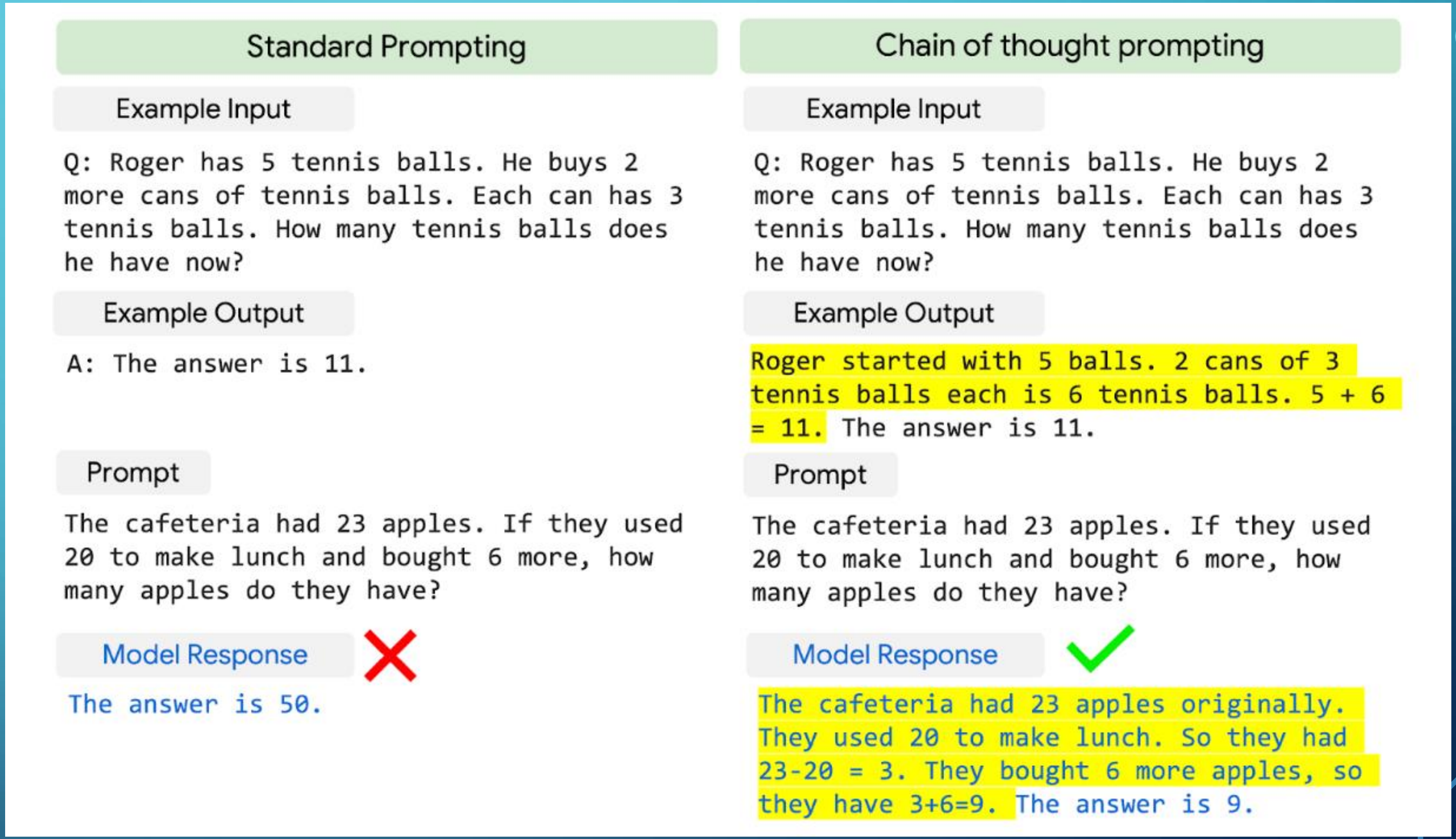


1-SHOT PERFORMANCE

- Open tab showing 1-shot performance.

REASONING

- By combining model scale with chain-of-thought prompting, PaLM shows breakthrough capabilities on reasoning tasks that require multi-step arithmetic or common-sense reasoning.



Standard Prompting

Example Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

Example Output

A: The answer is 11.

Prompt

The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Response ✗

The answer is 50.

Chain of thought prompting

Example Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

Example Output

Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Prompt

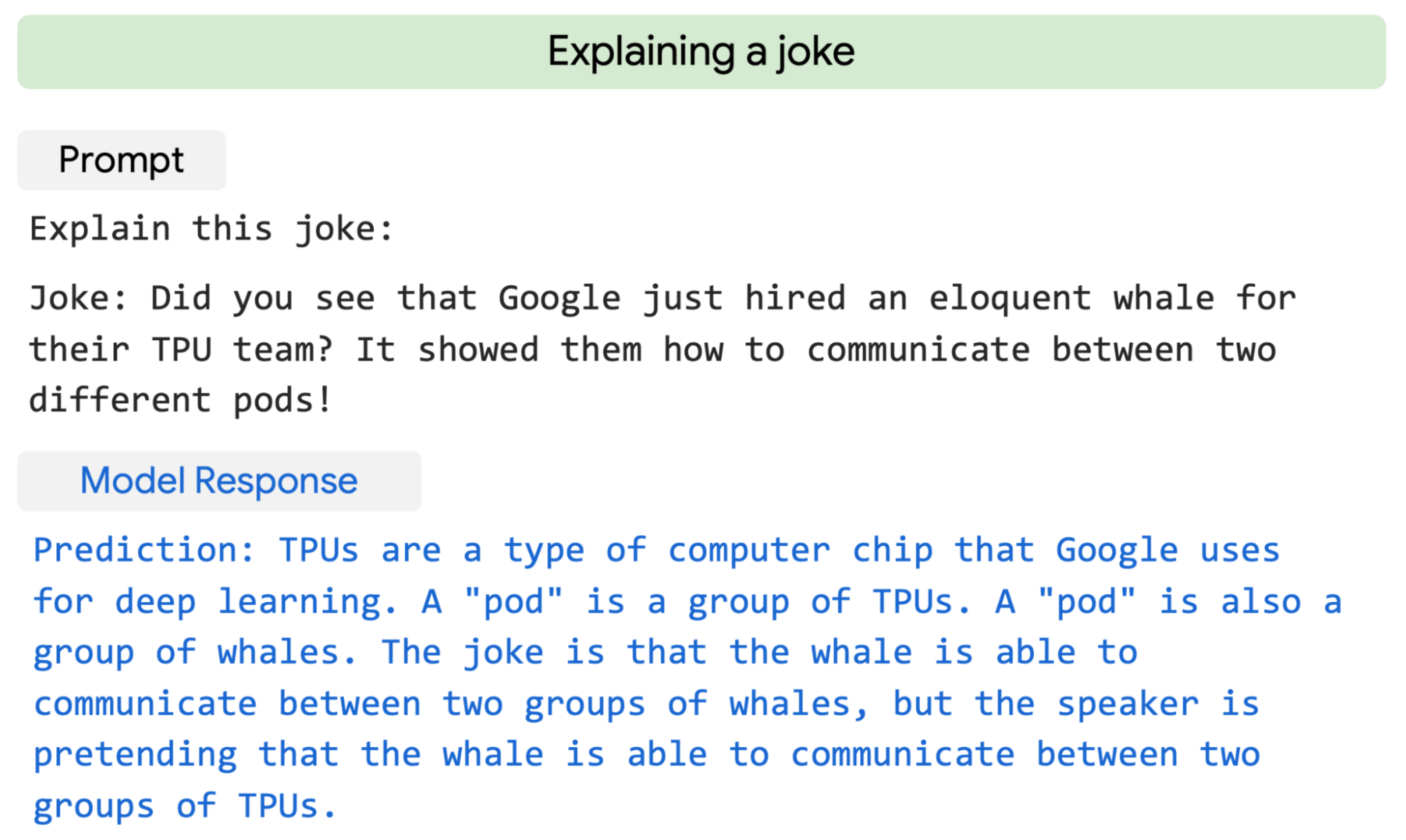
The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Response ✓

The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23-20 = 3. They bought 6 more apples, so they have 3+6=9. The answer is 9.

EXPLAINING A JOKE

- PaLM can provide high quality explanations for novel jokes not found on the web



Explaining a joke

Prompt

Explain this joke:

Joke: Did you see that Google just hired an eloquent whale for their TPU team? It showed them how to communicate between two different pods!

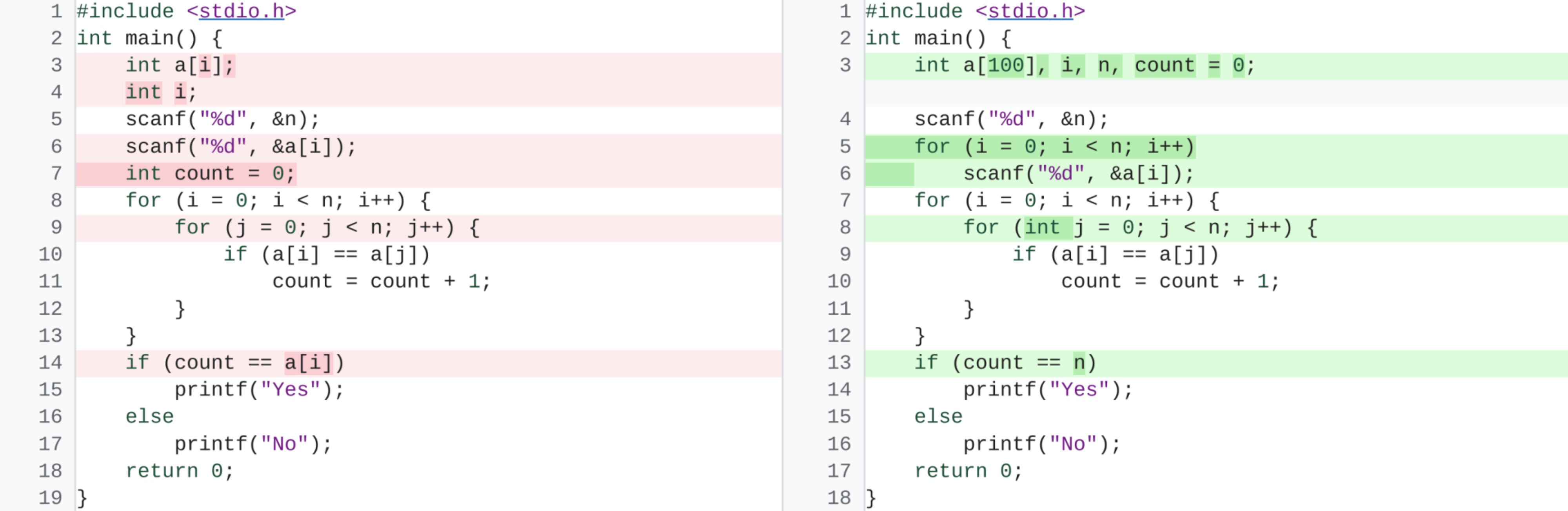
Model Response

Prediction: TPUs are a type of computer chip that Google uses for deep learning. A "pod" is a group of TPUs. A "pod" is also a group of whales. The joke is that the whale is able to communicate between two groups of whales, but the speaker is pretending that the whale is able to communicate between two groups of TPUs.

CODING

- PaLM 540B shows strong performance across coding tasks and natural language tasks in a single model, even though it has only 5% code in the pre-training dataset.
- Open tab showing coding examples.

CODE REPAIR



```c
#include <stdio.h>
int main() {
    int a[i];
    int i;
    scanf("%d", &n);
    scanf("%d", &a[i]);
    int count = 0;
    for (i = 0; i < n; i++) {
        for (j = 0; j < n; j++) {
            if (a[i] == a[j])
                count = count + 1;
        }
    }
    if (count == a[i])
        printf("Yes");
    else
        printf("No");
    return 0;
}
```

```c
#include <stdio.h>
int main() {
    int a[100], i, n, count = 0;
    scanf("%d", &n);
    for (i = 0; i < n; i++)
        scanf("%d", &a[i]);
    for (i = 0; i < n; i++) {
        for (int j = 0; j < n; j++) {
            if (a[i] == a[j])
                count = count + 1;
        }
    }
    if (count == n)
        printf("Yes");
    else
        printf("No");
    return 0;
}
```

- An example code repair task called DeepFix, where the objective is to modify initially broken C programs until they compile successfully, PaLM-Coder 540B demonstrates impressive performance, achieving a compile rate of 82.1%, which outperforms the prior 71.7% state of the art.

ETHICS

- "Recent research has highlighted various potential risks associated with LLMs trained on web text. It is crucial to analyze and document such potential undesirable risks through transparent artifacts such as model cards and datasheets, which also include information on intended use and testing. To this end, our paper provides a datasheet, model card and Responsible AI benchmark results, and it reports thorough analyses of the dataset and model outputs for biases and risks. While the analysis helps outline some potential risks of the model, domain- and task-specific analysis is essential to truly calibrate, contextualize, and mitigate possible harms."

DALL•E 2

- DALL·E 2 is a new AI system that can create realistic images and art from a description in natural language. It can combine concepts, attributes, and styles.