Whiteboard-of-Thought:

Thinking Step-by-Step Across Modalities

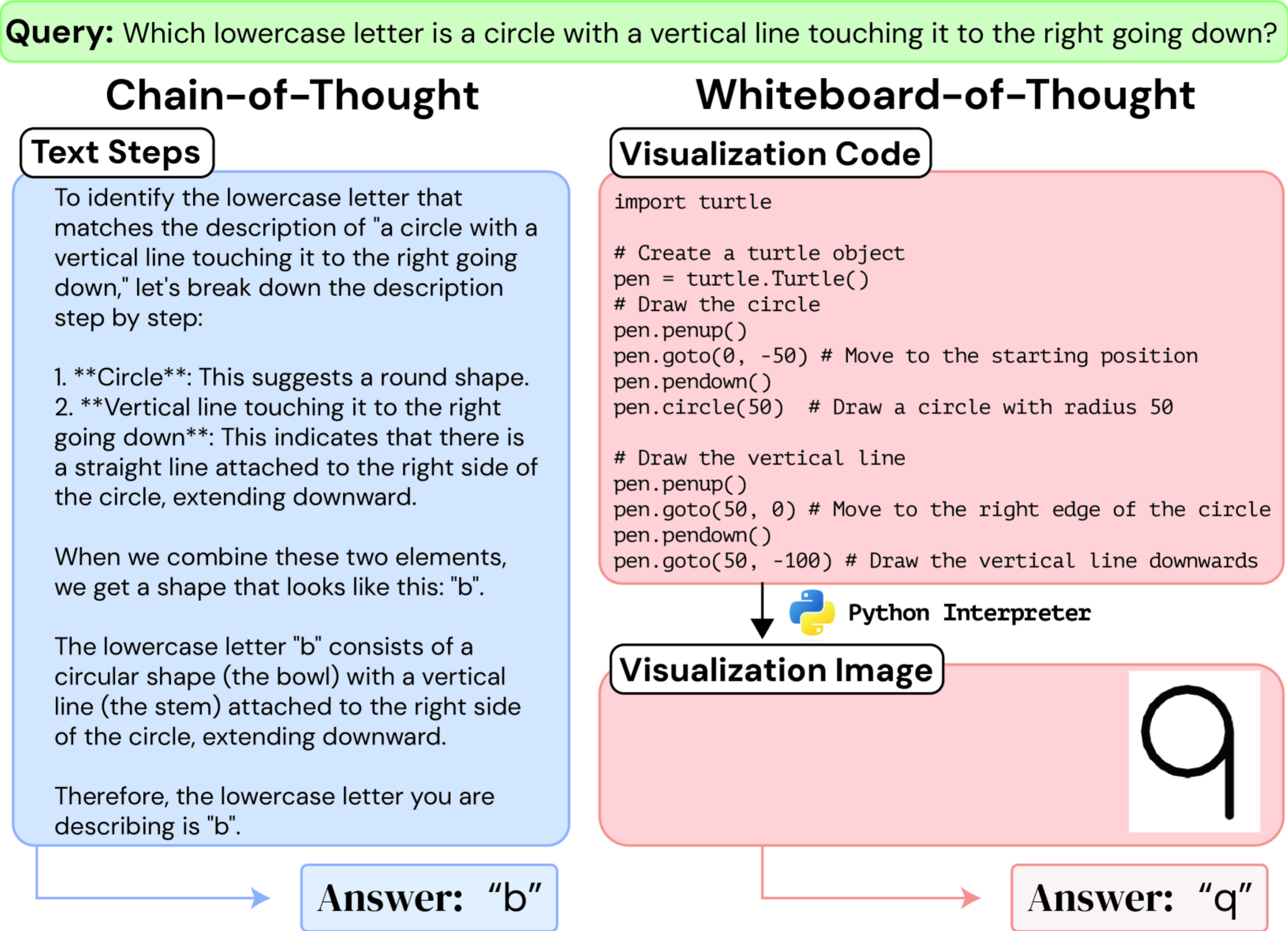
Sachit Menon, Richard Zemel, Carl Vondrick (2024)

概要

- 視覚的思考を要するタスクにおいて、マルチモーダル大規模言語モデル（MLLMs）の能力を解放するための手法「Whiteboard-of-Thought（WoT）」を提案。
- コードを生成して画像を作成し、その画像をモデルに戻してさらなる推論を実行(MatplotlibやTurtleを使用)。
- 視覚的および空間的推論が必要なタスクでCoTを上回る成果を示す。

クイズ

Which lowercase letter is a circle with a vertical line touching it to the right going down?



Query: Which lowercase letter is a circle with a vertical line touching it to the right going down?

Chain-of-Thought

Text Steps

To identify the lowercase letter that matches the description of "a circle with a vertical line touching it to the right going down," let's break down the description step by step:

1. **Circle**: This suggests a round shape.
2. **Vertical line touching it to the right going down**: This indicates that there is a straight line attached to the right side of the circle, extending downward.

When we combine these two elements, we get a shape that looks like this: "b".

The lowercase letter "b" consists of a circular shape (the bowl) with a vertical line (the stem) attached to the right side of the circle, extending downward.

Therefore, the lowercase letter you are describing is "b".

Answer: "b"

Whiteboard-of-Thought

Visualization Code

```
import turtle

# Create a turtle object
pen = turtle.Turtle()
# Draw the circle
pen.penup()
pen.goto(0, -50) # Move to the starting position
pen.pendown()
pen.circle(50)  # Draw a circle with radius 50

# Draw the vertical line
pen.penup()
pen.goto(50, 0) # Move to the right edge of the circle
pen.pendown()
pen.goto(50, -100) # Draw the vertical line downwards
```

Python Interpreter

Visualization Image

Answer: "q"

背景と動機

- 人間は、問題解決時に言語と視覚の推論を織り交ぜて利用。
- LLMsは算術や記号的推論では成功しているが、視覚的推論には課題が残る。
- 提案手法WoTは、比喩的な「ホワイトボード」を使用して視覚化を行い、視覚的推論能力を解放。

手法

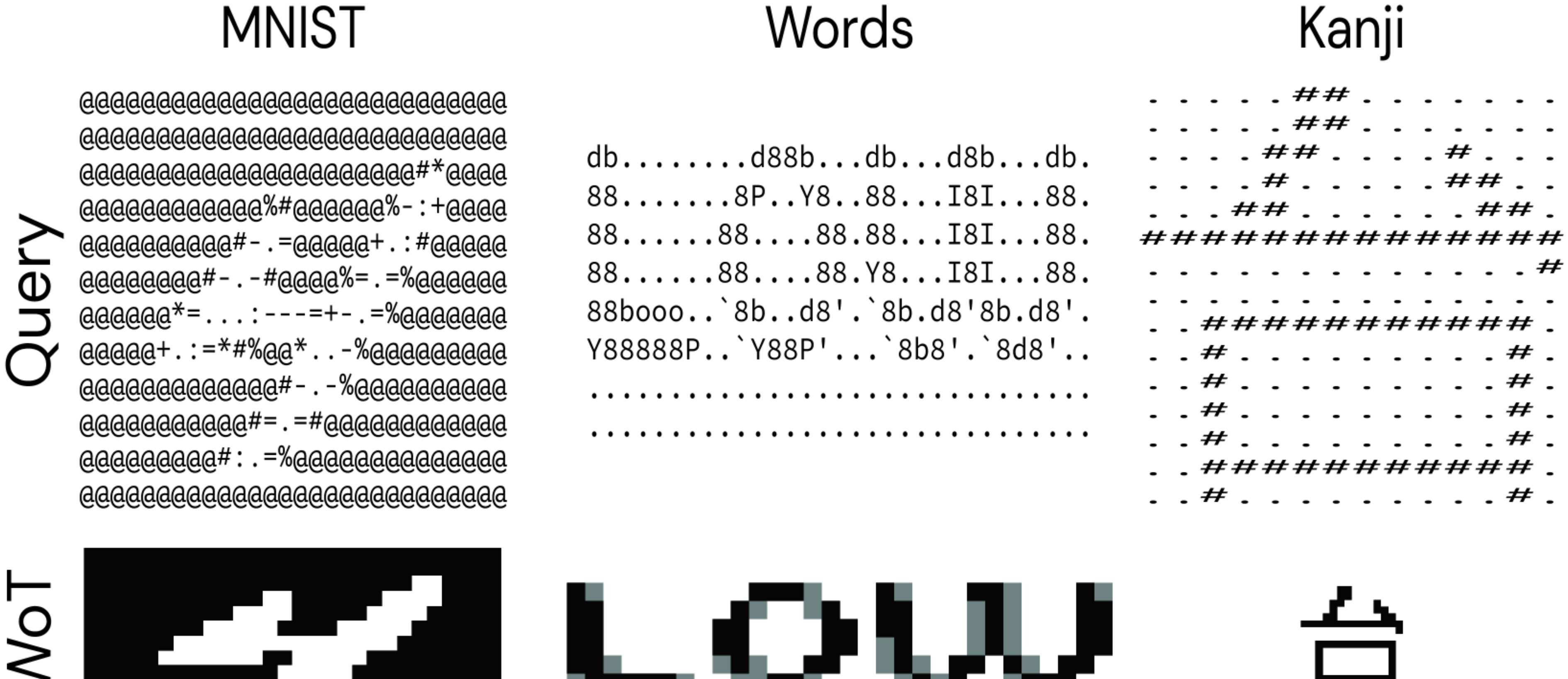
Whiteboard-of-Thoughtの流れ

1. **クエリ**をモデルに入力。
2. モデルが**視覚化を生成するコード**を生成。
3. コードを実行して**画像を作成**。
4. 生成された画像をモデルに戻してさらなる推論。

実験

タスク1: アスキーアートの理解

- **BIG-Benchデータセット**（例：MNIST数字認識、単語認識、漢字認識）を使用。



- 人間にとってはテキストも画像も同じモダリティだが、LLMにとっては異なるモダリティ。

Task	Direct (%)	CoT (%)	WoT (%)
MNIST	19.6	21.6	66.0
Word	24.8	27.2	66.4
Kanji	1.1	1.1	73.8

- CoTでは性能が不十分だが、WoTは視覚化により大幅な性能向上を達成。

タスク2: 空間ナビゲーション

- 異なる空間構造（円形、六角形、正方形など）でのナビゲーション指示を評価。

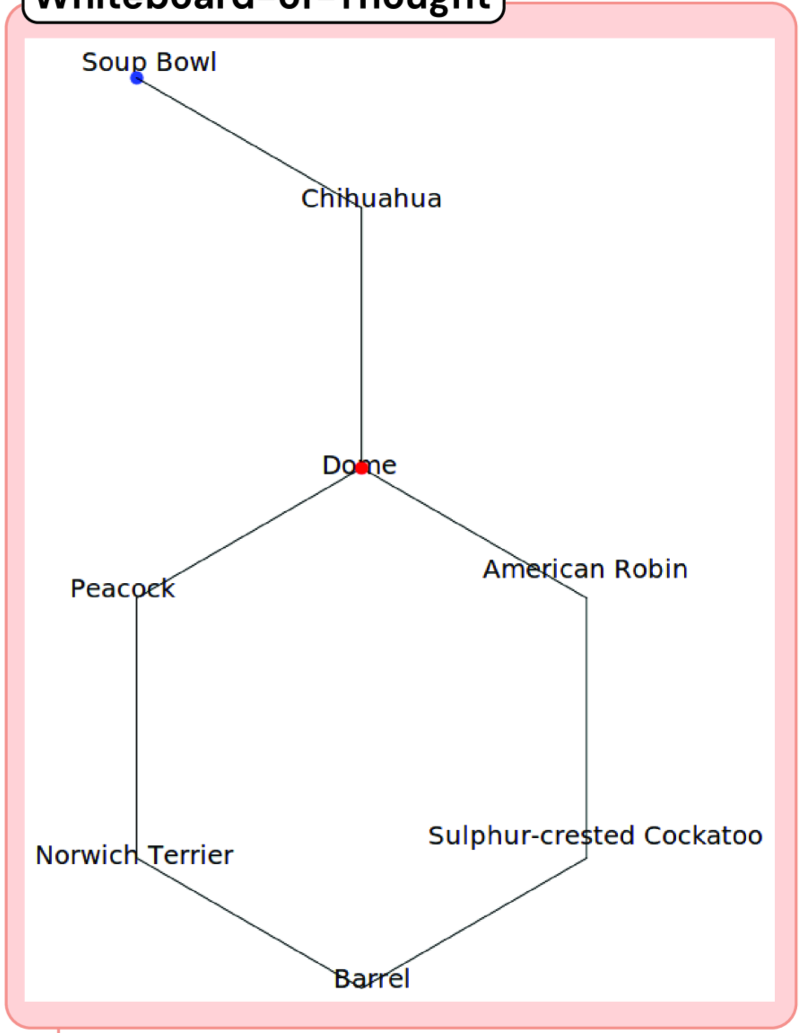
Query:

You have been given a pointy-topped regular hexagonal tile map consisting of 2 rows, where the first row has one tile and the second row has two tiles.

Starting from a vertex, you will move along the edges of these tiles.

Initially, you are positioned at the top corner of the map, where you find a soup bowl. You move down-right by one step, where you find a Chihuahua. You move down by one step, where you find a dome. You move down-left by one step, where you find a peacock. You move down by one step, where you find a Norwich Terrier. You move down-right by one step, where you find a barrel. You move up-right by one step, where you find a sulphur-crested cockatoo. You move up by one step, where you find an American robin. You move up-left by one step.

Whiteboard-of-Thought



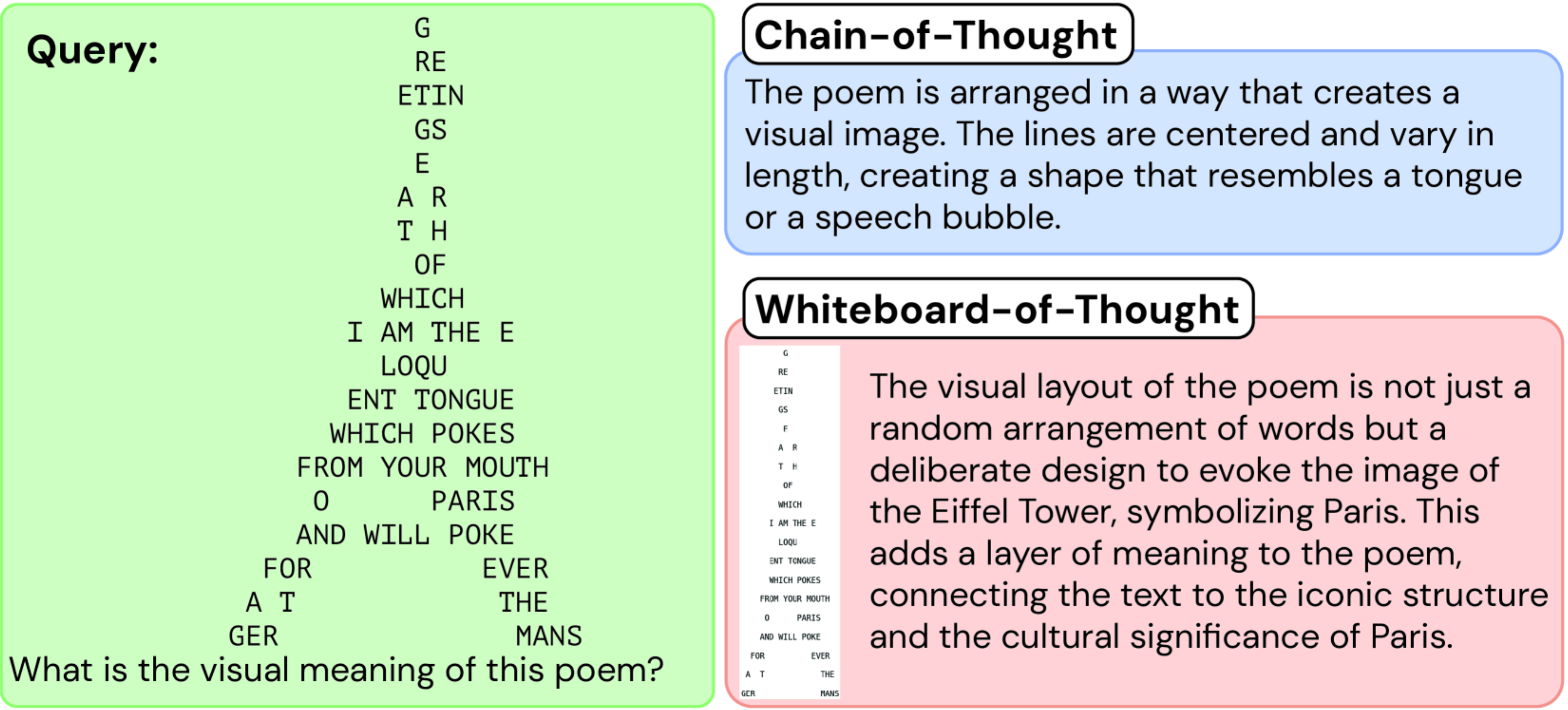
Structure	**Circle (%)**	**Hexagon (%)**	**Triangle (%)**	**Square (%)**	**Rhombus (%)**	**Overall Avg (%)**
Direct	14	3	16	68	63	33
CoT	25	8	26	98	51	42
WoT	41	61	55	50	52	52

- グリッドはテキストで表現しやすいことに加えてデータが多いので従来手法でも高性能。
- WoTはグリッド以外の形状でも一貫した性能を発揮。

実際の使用例

カリグラム

- 文字の配置が詩の意味に寄与する



Query:

G
RE
ETIN
GS
E
A R
T H
OF
WHICH
I AM THE E
LOQU
ENT TONGUE
WHICH POKES
FROM YOUR MOUTH
O PARIS
AND WILL POKE
FOR EVER
A T THE
GER MANS

What is the visual meaning of this poem?

Chain-of-Thought

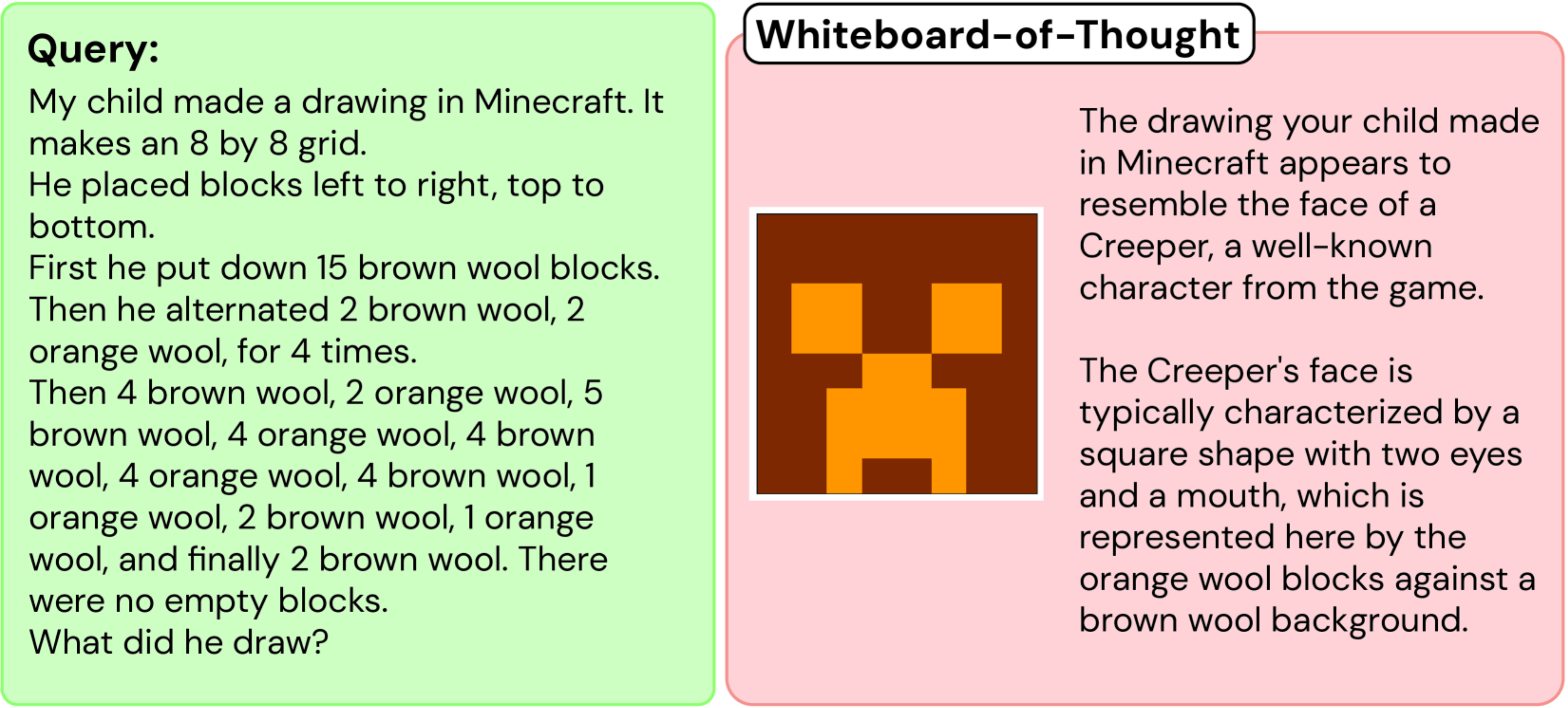
The poem is arranged in a way that creates a visual image. The lines are centered and vary in length, creating a shape that resembles a tongue or a speech bubble.

Whiteboard-of-Thought

The visual layout of the poem is not just a random arrangement of words but a deliberate design to evoke the image of the Eiffel Tower, symbolizing Paris. This adds a layer of meaning to the poem, connecting the text to the iconic structure and the cultural significance of Paris.

ゲーム内のアート作品

- プレイヤーの生成したコンテンツが適切かどうか行動ログから判断



Query:

My child made a drawing in Minecraft. It makes an 8 by 8 grid.
He placed blocks left to right, top to bottom.
First he put down 15 brown wool blocks.
Then he alternated 2 brown wool, 2 orange wool, for 4 times.
Then 4 brown wool, 2 orange wool, 5 brown wool, 4 orange wool, 4 brown wool, 4 orange wool, 4 brown wool, 1 orange wool, 2 brown wool, 1 orange wool, and finally 2 brown wool. There were no empty blocks.
What did he draw?

Whiteboard-of-Thought

The drawing your child made in Minecraft appears to resemble the face of a Creeper, a well-known character from the game.

The Creeper's face is typically characterized by a square shape with two eyes and a mouth, which is represented here by the orange wool blocks against a brown wool background.

実装してみた

https://github.com/HKodai/WoT

まとめ

- 視覚的および空間的推論が必要なタスクにおいて、WoTはCoTを上回る成果を示す。
- モデルのコード生成能力、視覚入力の理解能力、一般的な推論能力が向上するにつれて、WoTの結果も同様に向上すると期待される。

参考文献

Whiteboard-of-Thought: Thinking Step-by-Step Across Modalities
(https://whiteboard.cs.columbia.edu/)