Weak-eval-Strong: Evaluating and Eliciting Lateral Thinking of LLMs with Situation Puzzles

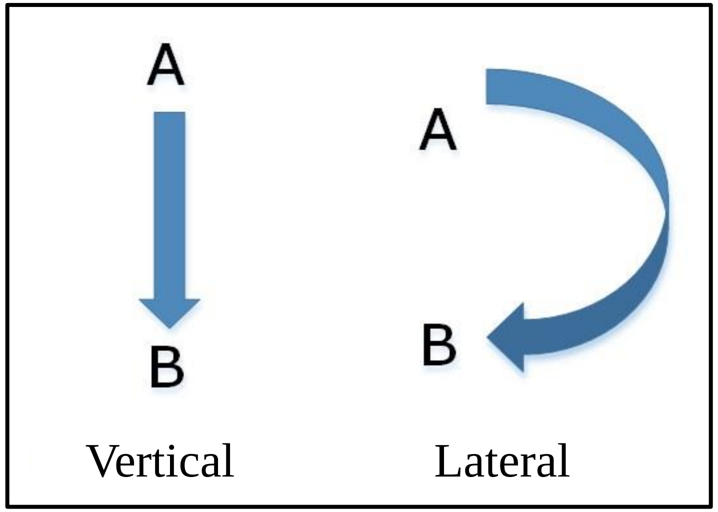
Qi Chen, Bowen Zhang, Gang Wang, Qi Wu*

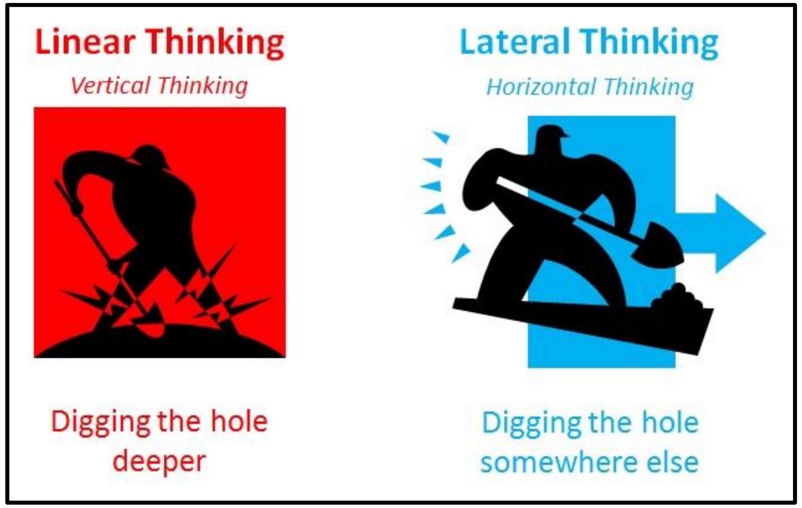
Australian Institute for Machine Learning (AIML), The University of Adelaide

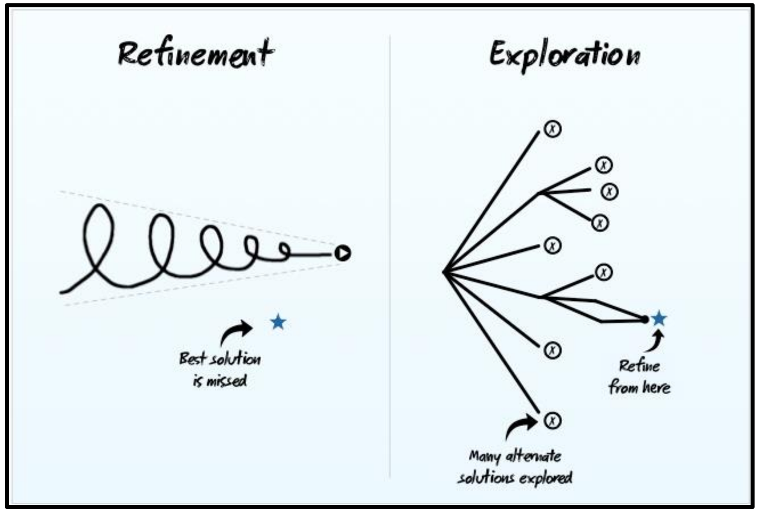
Funded by the Centre for Augmented Reasoning

Vertical Thinking vs. Lateral Thinking

- ☐ **Vertical Thinking** follows a sequence of direct and logical steps to reach a conclusion.
- ☐ **Lateral Thinking** uses indirect and creative approaches to find solutions.

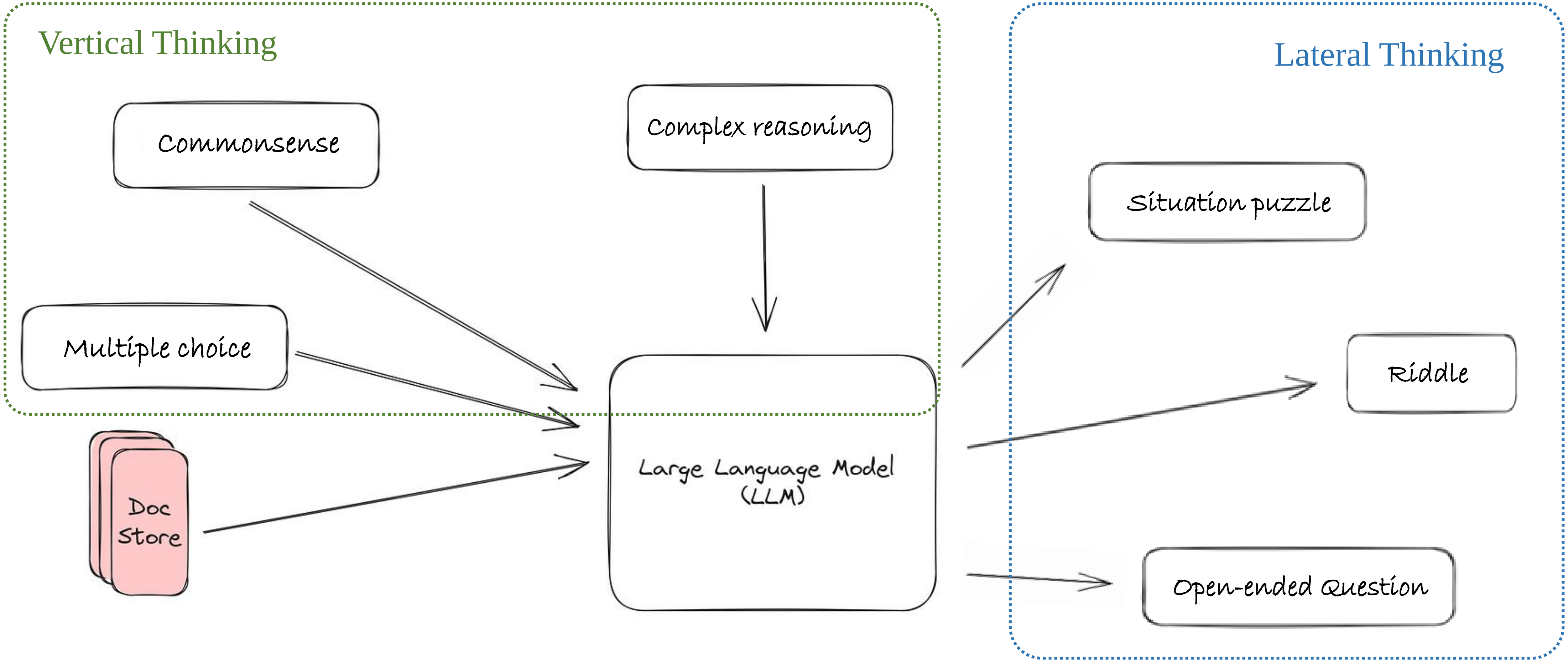






*Figures are from https://www.oventhal.com/blog/2019/4/18/vertical-vs-lateral-thinking and https://andyeklund.com/is-linear-thinking-bad/

Lateral Thinking Benchmark for LLMs

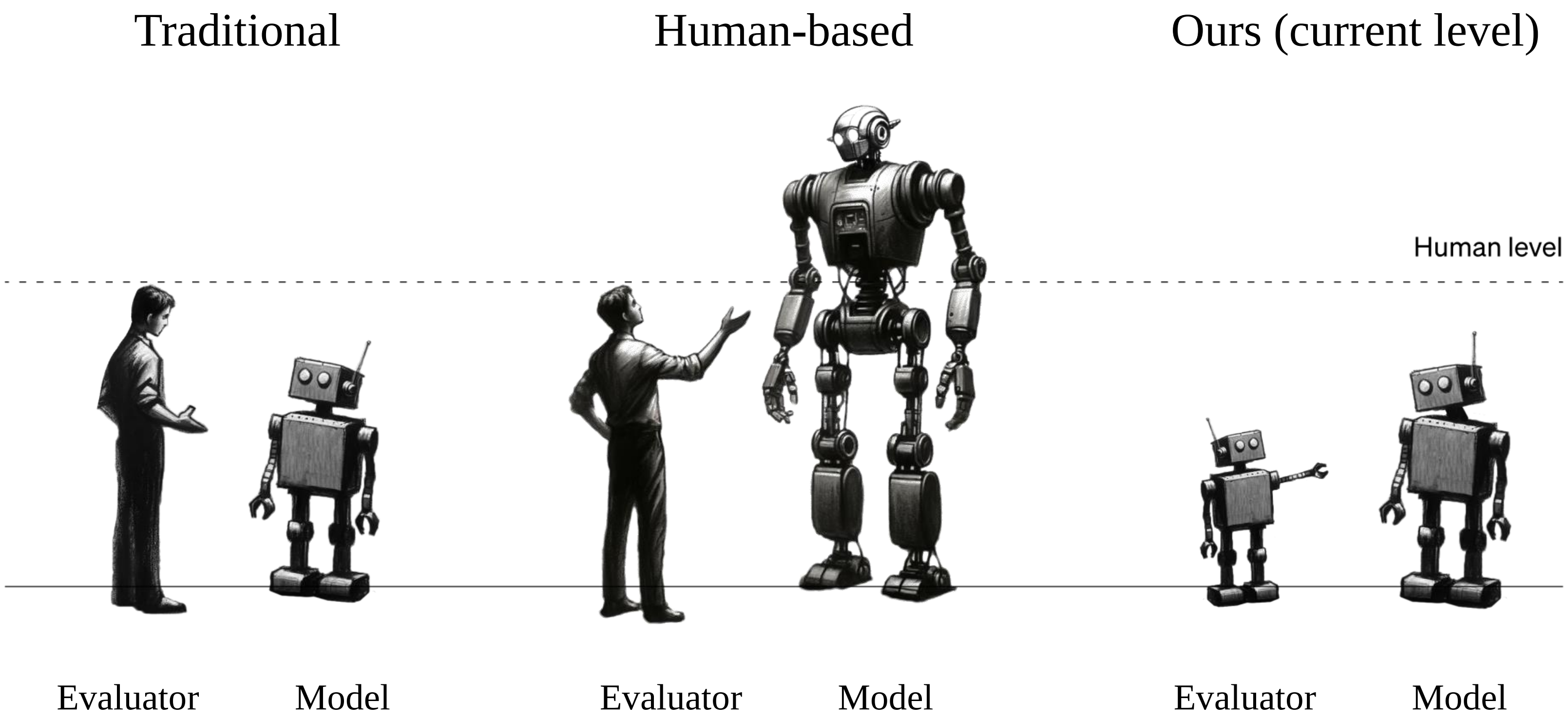


Weak-eval-Strong Requirement

Benchmarks increasingly aim for tasks challenging even for humans, with outputs that are typically open-ended

(Need)

- Human-based evaluation
- Model-based evaluation



*Figures are from https://openai.com/index/weak-to-strong-generalization/

Our SPLAT Benchmark for LLMs

Easy

Question/Story: A black cat is walking through the middle of a road where there is a broken streetlamp. A man is driving his vehicle with headlights that do not work. Despite all of this, he is able to avoid harming the cat. Why?

Reference Answer: The man avoided the cat because it was daylight.

Medium

Question/Story: A hunter aimed his gun carefully and fired. Seconds later, he realized his mistake. Minutes later, he was dead. Why?

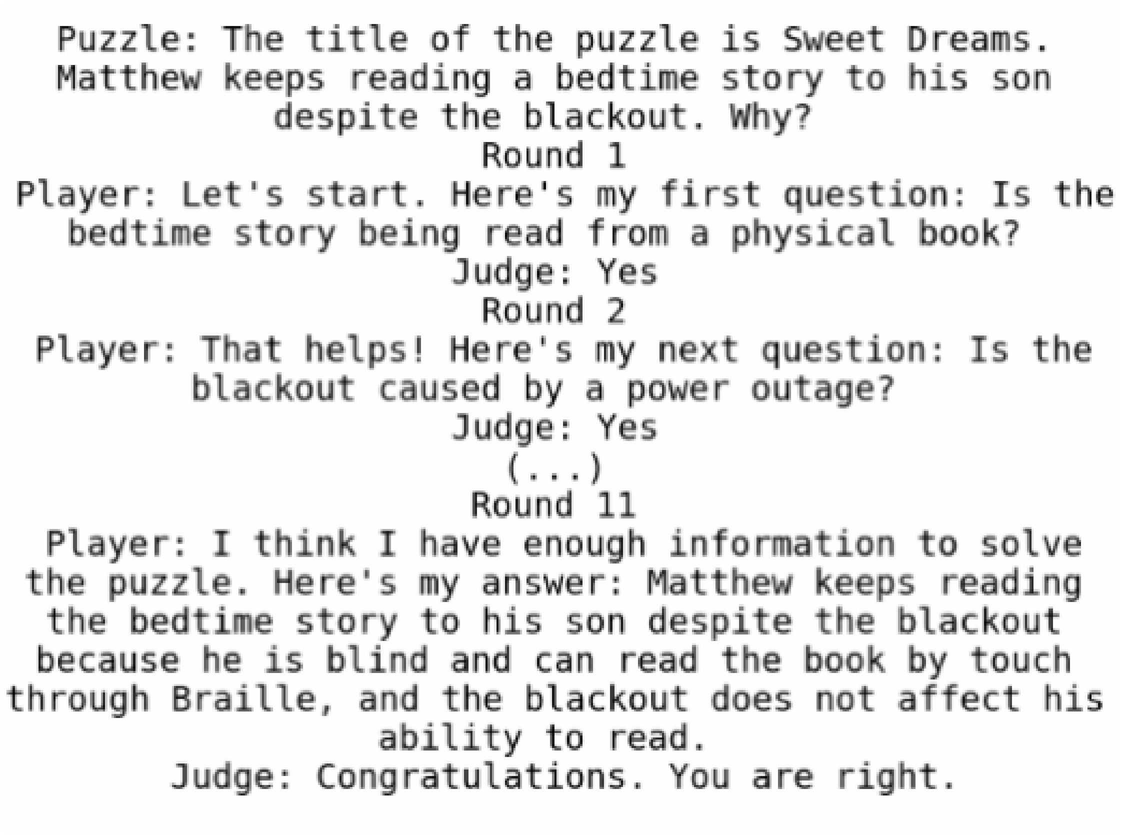
Reference Answer: He hunted in snow-capped mountains. The shot provoked an avalanche, which covered the man. He died of strangulation.

Hard

Question/Story: Two men are beating each other up and both of them suddenly fall to the floor. Why?

Reference Answer: The daughter of one of the boxers had been kidnapped and she would be killed if her father won the fight. At the end of the fight, the boxer's opponent fainted and the other boxer - in order not to win the match - threw himself to the floor.

Example of Game



Puzzle: The title of the puzzle is Sweet Dreams. Matthew keeps reading a bedtime story to his son despite the blackout. Why?

Round 1

Player: Let's start. Here's my first question: Is the bedtime story being read from a physical book?

Judge: Yes

Round 2

Player: That helps! Here's my next question: Is the blackout caused by a power outage?

Judge: Yes

(...)

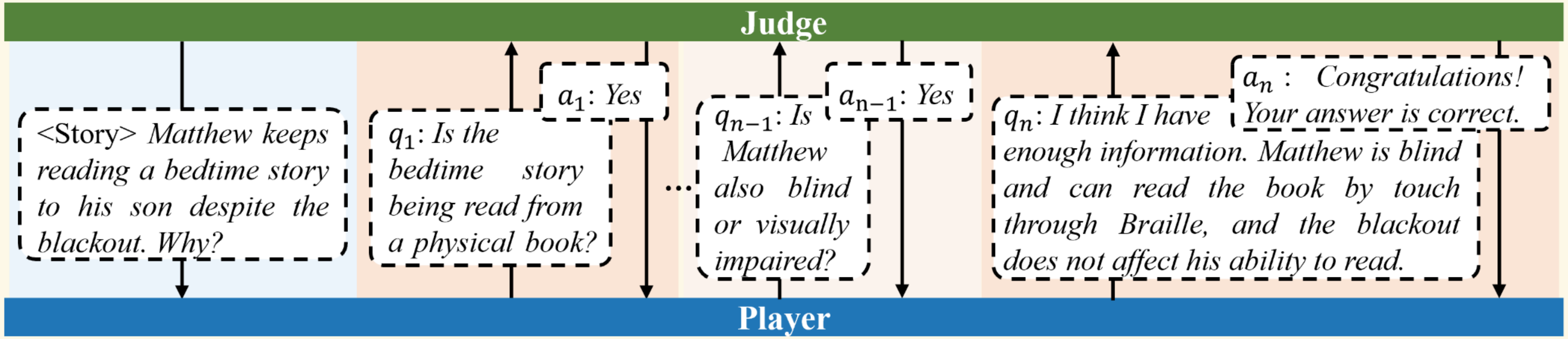
Round 11

Player: I think I have enough information to solve the puzzle. Here's my answer: Matthew keeps reading the bedtime story to his son despite the blackout because he is blind and can read the book by touch through Braille, and the blackout does not affect his ability to read.

Judge: Congratulations. You are right.

Weak-eval-Strong: multi-turn player-judge framework

Character Initialisation (Judge):
You are a judge, and your role is to challenge the user with a given puzzle in the form of a short, logically self-consistent story. The story are complete but may intentionally omit some details that the user needs to deduce or inquire about.
Here are the guidelines for your role: 1. 2. 3. 4.
Now, the system will give your first puzzle. The game starts when the user indicates they are ready. Enjoy the challenge of testing and expanding the user's deductive reasoning!
Puzzle Title: Sweet Dreams
Story: Matthew keeps reading a bedtime story to his son despite the blackout. Why?
Reference Answer: Matthew was blind, and he usually read bedtime stories to his son from a braille book. That night there was a blackout, but this did not stop him from finishing the story.



Judge

<Story> *Matthew keeps reading a bedtime story to his son despite the blackout. Why?*

q_1: *Is the bedtime story being read from a physical book?*

a_1: *Yes*

…

q_{n-1}: *Is Matthew also blind or visually impaired?*

a_{n-1}: *Yes*

q_n: *I think I have enough information. Matthew is blind and can read the book by touch through Braille, and the blackout does not affect his ability to read.*

a_n : *Congratulations! Your answer is correct.*

Player

Character Initialisation (Player):
You are a player and going to play an interesting situation puzzle game.
The rules of the game are as follows: 1. 2. 3. 4. 5. 6.
Are you ready? Let's begin! User will give you the first puzzle.

Elicit Lateral Thinking by Our Benchmark and Framework

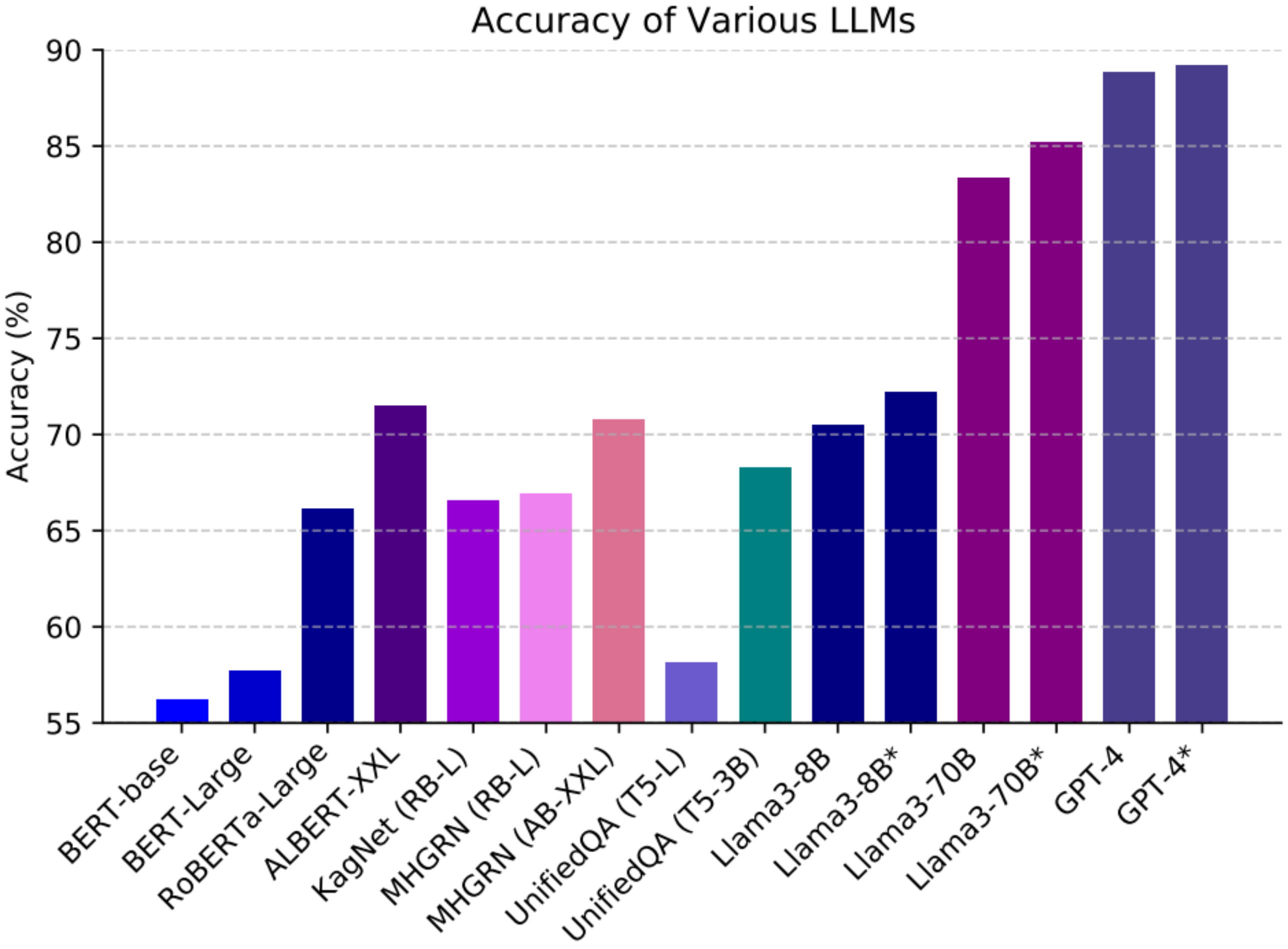


Table 6: Impact of our data and reasoning processes. Models with '†' mean using our data only while with '*' mean using both data and reasoning processes. 'RS' refers to the results on RiddleSense (Dev). 'BT (S.)' and 'BT (W.)' are the overall results on BrainTeaser (Sentence) and BrainTeaser (Word), respectively.

	RS	BT (S.)	BT (W.)	Avg.
Llama3-8B	70.51	67.65	46.20	61.45
Llama3-8B†	70.32	69.62	52.28	64.07
Llama3-8B*	72.18	69.89	58.07	66.71
Llama3-70B	83.34	87.76	71.20	80.76
Llama3-70B†	82.95	91.12	77.27	83.78
Llama3-70B*	85.21	91.51	79.54	85.42

Thank you for your attention!