Fleeting Connections: An Inquiry into Analogical Reasoning

Sofia Bazzan

Simone Bianchi

The study:

Connecting long distance: semantic distante in analogical reasoning Modulates frontpolar cortex activity

Adam E. Green1, David J. M. Kraemer2,3, Jonathan A. Fugelsang4, Jeremy R. Gray1 and Kevin N. Dunbar5

Purpose

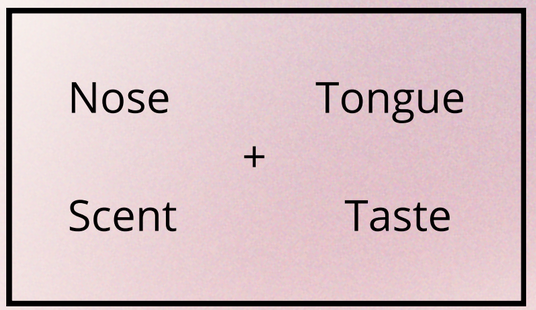
See if **increasing semantic distance** in an analogical reasoning task **increase** the **activity** in the **front polar cortex**.

Experiment

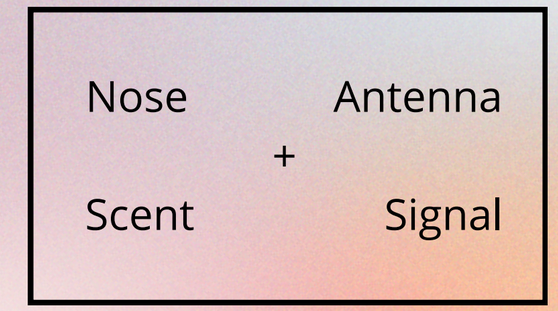
23 partecipants performed **120 analogy trials** during a 4-event related fMRI runs. On each trial, participants indicated whether a **4-word set** constituted a **valid analogy** (left word pair analogous to right word pair), responding ‘‘true’’ or ‘‘false’’ by button press with the index or middle finger of the right hand

Three kind of analogies

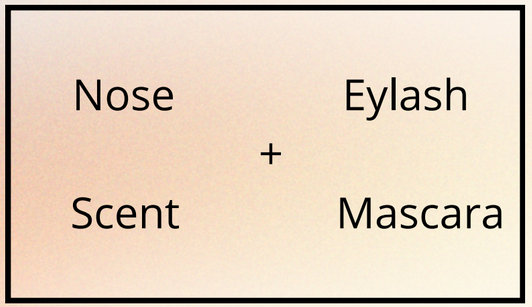
Within-Domain Analogy



Cross-Domain Analogy



False Analogy



For **each trial** was computed a **semantic distance value** and assigned the class by asking to 84 independent raters to **separate the true analogies** between **cross-domain analogies** (involving mapping between items taken from disparate semantic domains), and **within-domain analogies** (involving mapping between items taken from proximal semantic domains). In addition, the 84 independent raters used a 7-point scale to **score** all analogy stimuli for **difficulty** (“How difficultis it to identify the analogical connection?”) .

Results

Accuracy

Participants performed at a response **accuracy level of 92.97%** ± standard deviation (SD) = 4.62%, standard error (SE) = 0.42% overall. Also item analysis revealed that response **accuracy** and **semantic distance values** were **not correlated** (r = –0.18, P = 0.11).

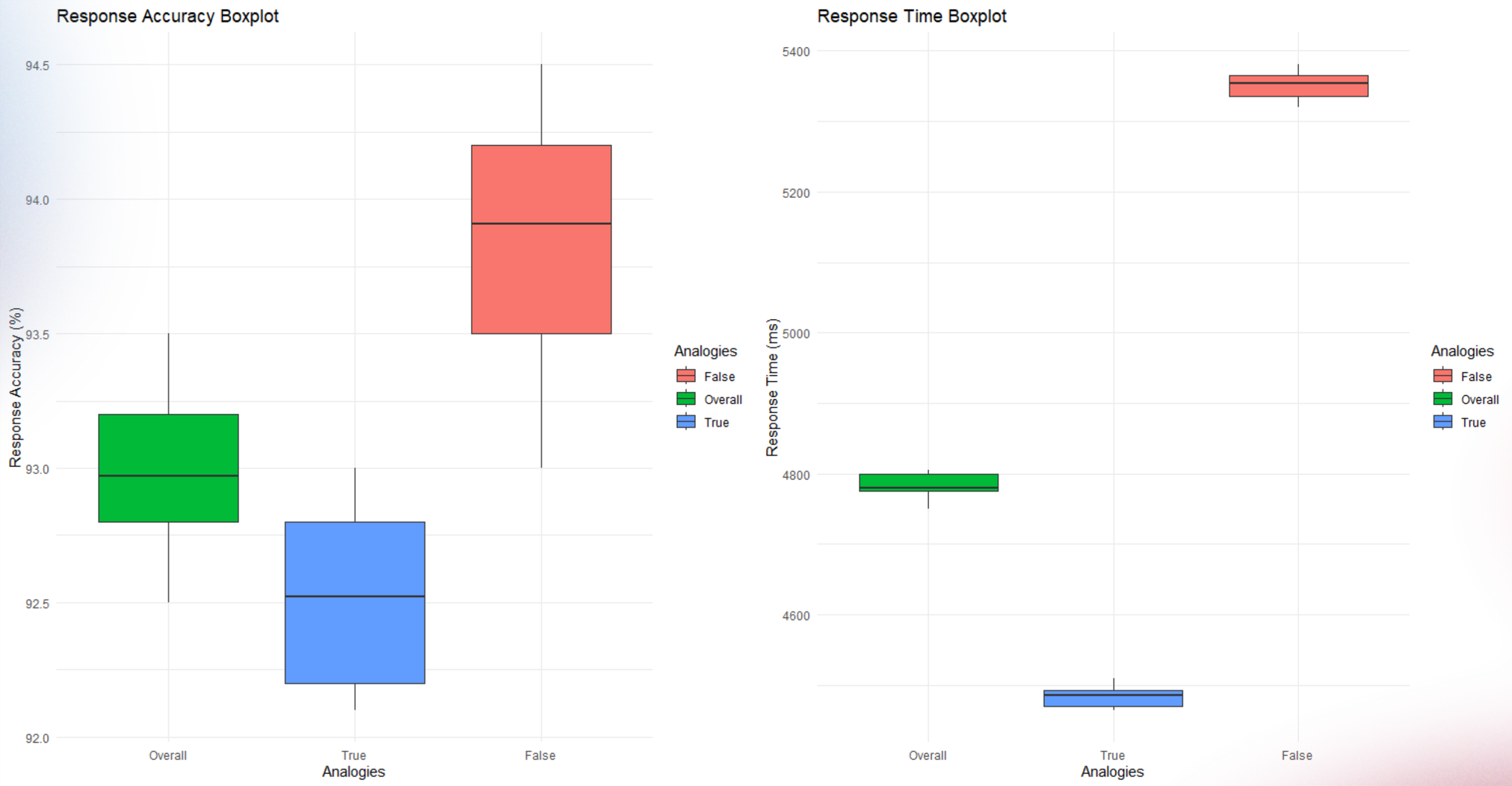
Response time and semantic distance

Participants performed with a mean response time of 4780 ± SD = 503, SE = 54 ms overall. **Response time** was **positively correlated** with **semantic distance** (r = 0.36, P = 0.001).

Frontpolar activity and semantic distance

Frontopolar recruitment strengthened as a function of **increasing semantic distance** of analogical mapping. Also from the study was found that **difficulty** related factors **cannot explain** the **relation** between **semantic distance** and **frontpolar activity.**

Boxplot Of Human Perfomance



Our Investigation

We tested **llama2,** with different parameters and different prompts, on the same trials of the experiment, in order to see if the **performance of the artificial intelligence** were **similar** to the one **achieved by umans**.

Methodology

1. **Create a dataset:** First we created a dataset were each data is an **instruction phrase** and a **response** ("Cross-Domain", "Within-Domain" or "False").
2. **Model interaction**: We accessed to Llama2 via the **Gradient platform** and for each element in the dataset was given the input **prompt** to the model to generate its response. We employed **three** prompts.
3. **Store model-generated responses**: Each response was then checked for **specific words** ("Cross-Domain," "Within-Domain," or "FALSE") and inserted into a **vector**
4. **Comparison with the groundtruth**: The final vector was then **compared** with the **true answer** for the question.

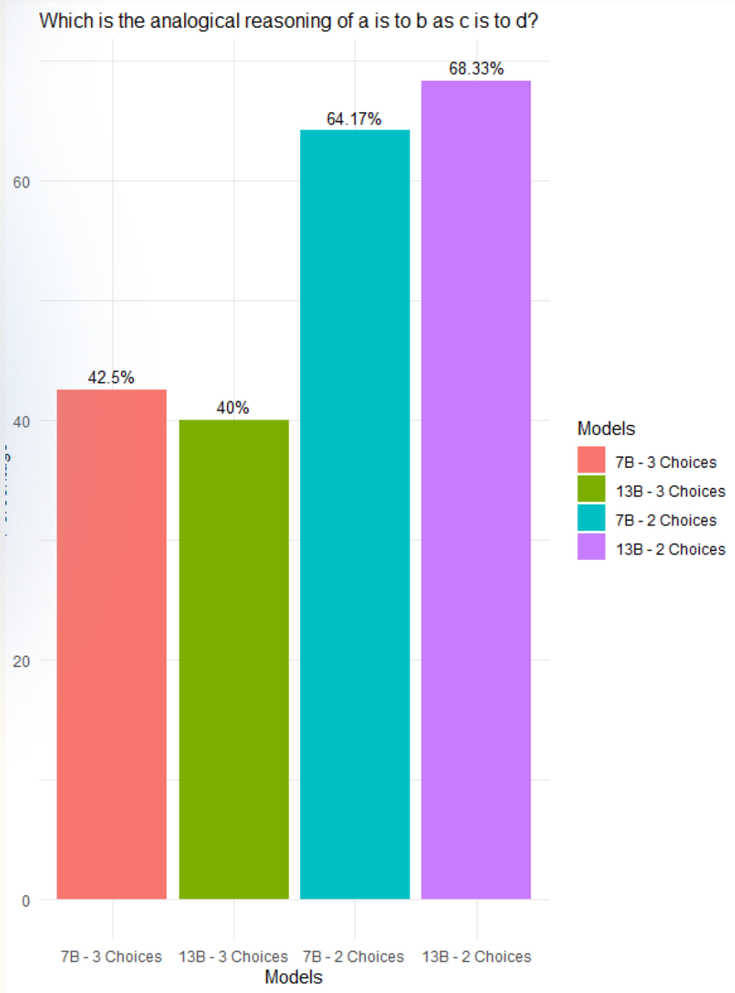
Prompts

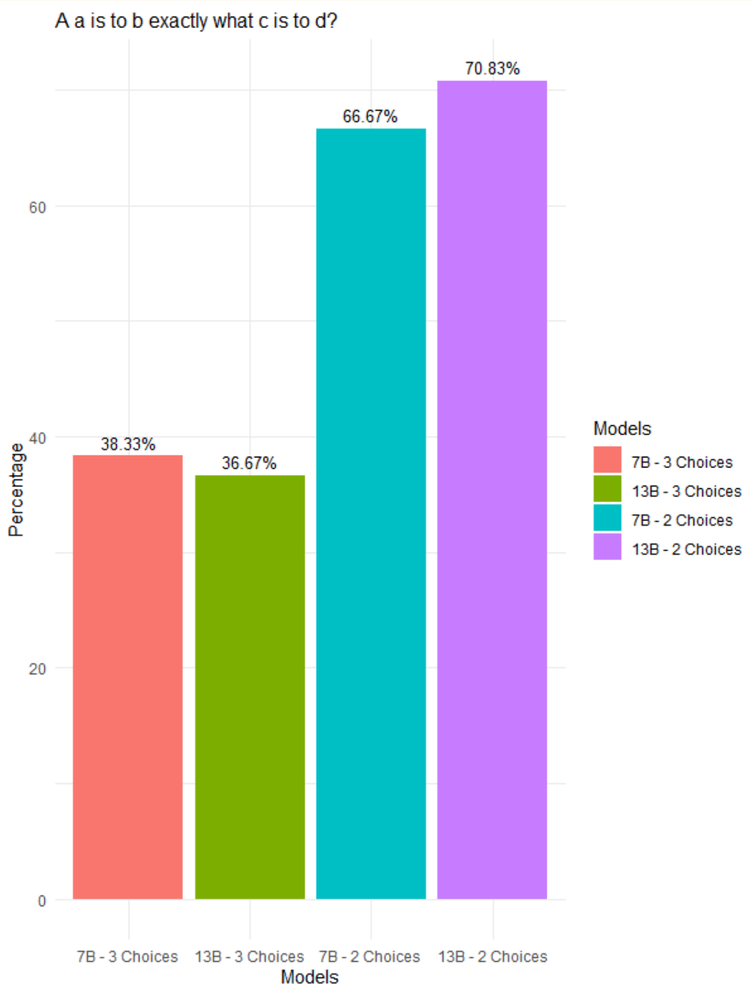
Prompt 1: Which is the analogical reasoning of a is to b as c is to d?

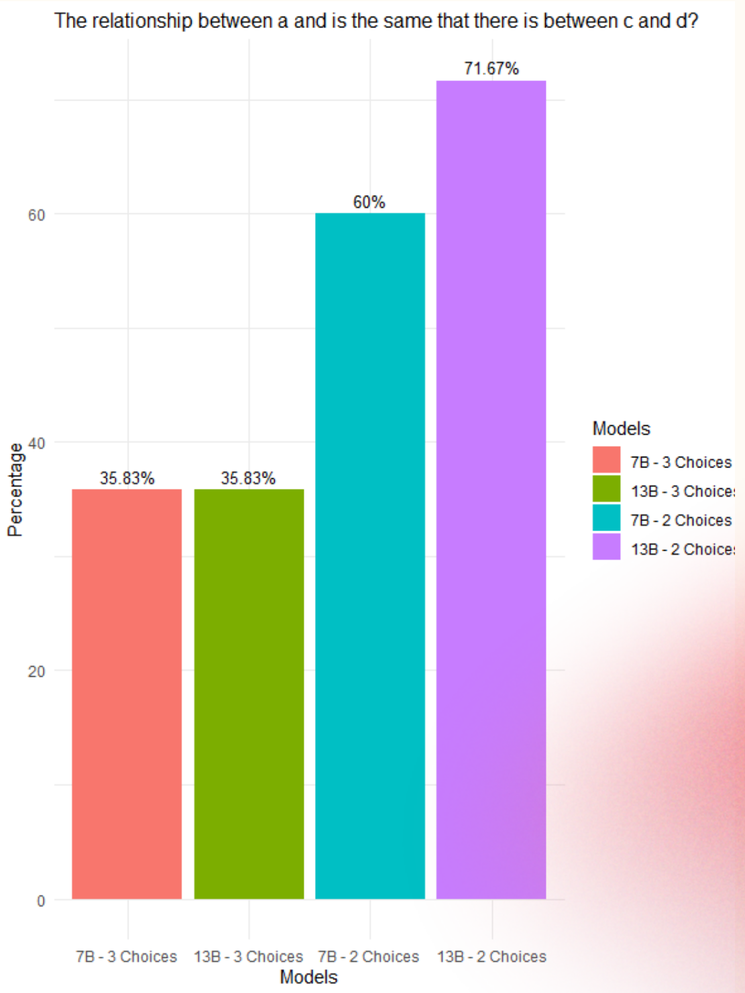
Prompt 2: A a is to b exactly what a c is to d?

Prompt 3: The relationship between a and b is the same that there is between c and d?

Results For The Different Models







Results For The Different Models

Prompt 1		
	False Positive	False Negative
7B - 3 choices	69	0
13B - 3 choices	64	8
7B - 2 choices	39	3
13B - 2 choices	38	0

Prompt 2		
	False Positive	False Negative
7B - 3 choices	74	0
13 - 3 choices	71	5
7B - 2 choices	35	5
13B - 2 choices	34	1

Prompt 3		
	False Positive	False Negative
7B - 3 choices	73	4
13B - 3 choices	68	9
7B - 2 choices	9	39
13B - 2 choices	23	11

In the scenario where **three possible choices** are considered, the **total number** of **False Positives** is obtained from the **sum** of False Positives associated with **Cross-Domain** and those associated to **Within-Domain**.

Major issue

The **accuracy** obtained is **very low**, expecially **compared with human performance** and that is due to the fact that it **assign** the **same class** to the vast **majority of the data**. In particular:

- In the case with **3 possible categories** (“Within-domain”, “Cross-domain” and “False”) it classify the majority of the data as “**Cross-Domain**” or “**Within-Domain**” based on the prompt
- In the case with **2 possible categories** (“**True**” or “**False**”) it classify the majority of the data as **True**.

Our attempts to solve it

Changing prompts:

By asking the question “**The relationship** between **a** and **b** is the same that there is between **c** and **d**?” instead of “Which is the analogical reasonign od a is to b as c is to d?” the **performance** of Llama 2 with 13 Bilions parameters **improved**, for the binary classification task, from **68.33%** to **71,67%.**

However, performance of Llama 2 with 7 Billions and 13 Billions parameters drops with three possible analogies, from **42,5%** to **35.83%** and from **40%** to **35.83%**, respectively.

Fine-Tuning Approach

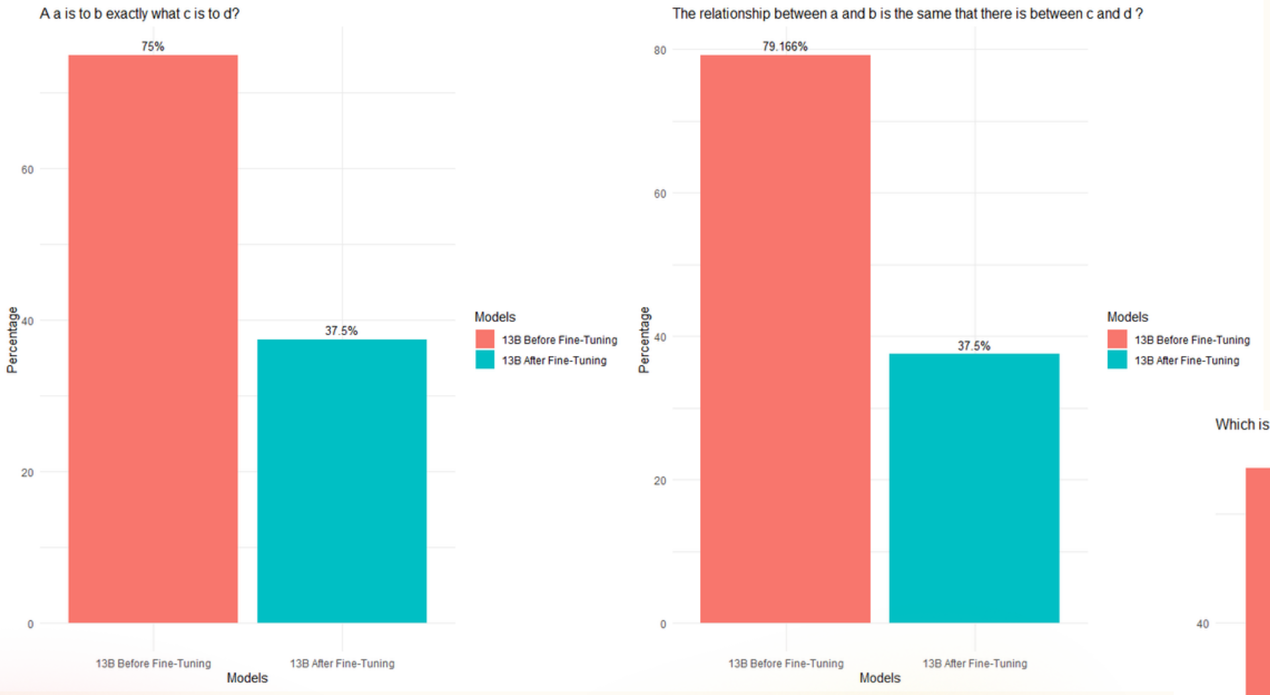
In attempt to refine model performance we also tried to **fine-tune** the models. We experimented with diverse prompts and scenarios for two and three potential choices. Accessing Llama-2, a model with **13 billion parameters** for **two-choices** scenarios and **7 billion parameters** for **three-choices scenarios**, via the Gradient platform. Due to time and computational constraints we curated subsets of **72 trials** for **two choices** and **40 trials** for **three choices** as our **training set** and **24 trials** as **test set**. Two **prompts** for each experiment were **selected** based on their promising **results** obtained **without fine-tuning.**

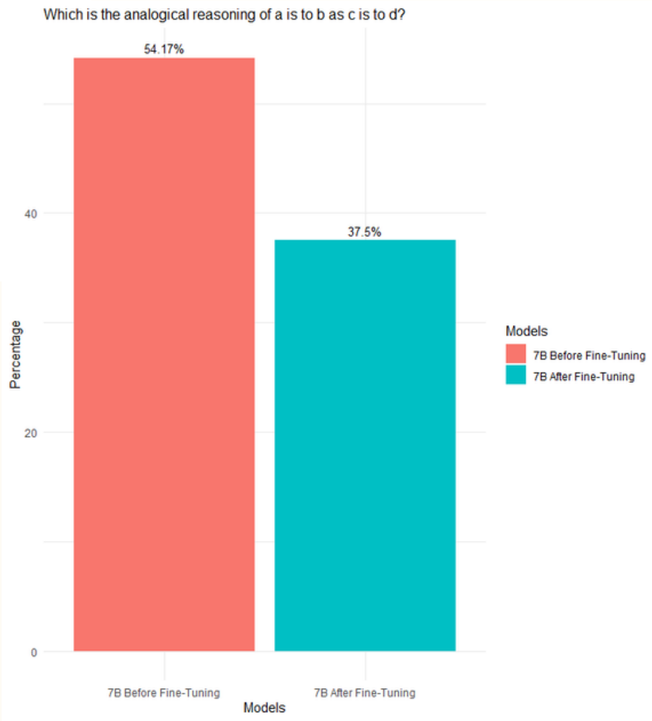
Results

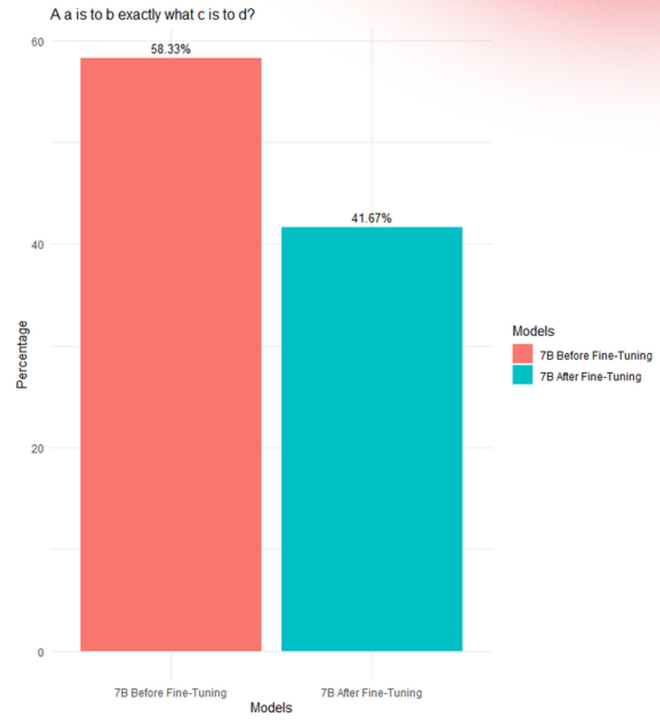
Our efforts to enhance model performance through fine-tuning were met with **specific challenges**. In a scenario with **two choices**, despite initially promising prompts, we observed a significant **decrease in accuracy**. This decline may be attributed to the complexities of fine-tuning, such as potential overfitting and deviation from the pre-trained model. Similarly, in a context involving **three choices**, the fine-tuning process unexpectedly led to a substantial **decrease** in overall **model performance.**

Overall, we anticipated an improvement in performance from this experiment. However, the reality was a **decrease in accuracy** in **both cases**, resulting in worse outcomes instead of closing the gap with human performance as we had hoped

Results For The Best Different Models







Results For The Best Different Models

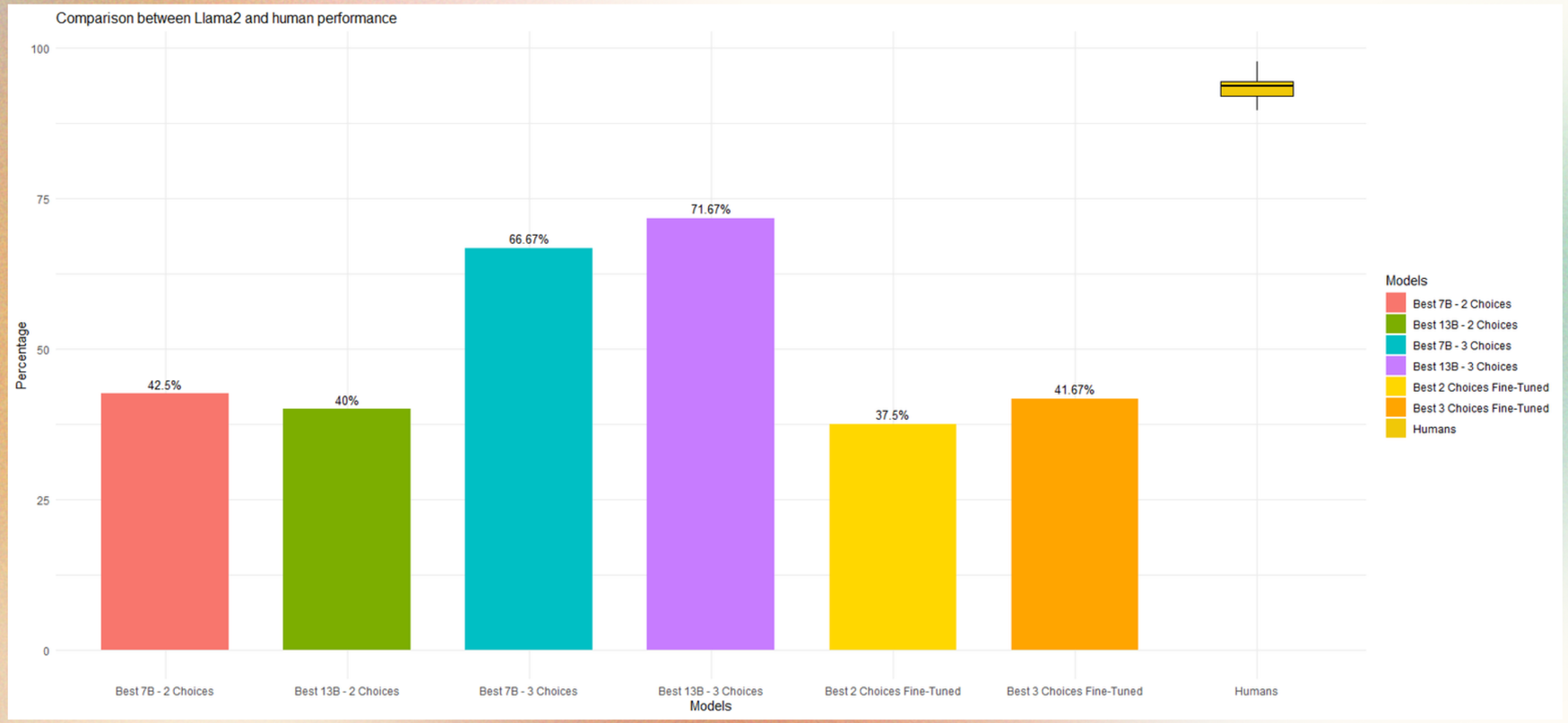
Prompt 2		
	Before fine-tuning	After fine-tuning
False positive	6	2
False negative	0	13
Accuracy	75%	37,5%

Prompt 3		
	Before fine-tuning	After fine-tuning
False positive	3	0
False negative	2	15
Accuracy	79,166%	37,5%

Prompt 1		
	Before Fine-Tuning	After Fine-Tuning
False Positive	8	10
False Negative	2	4
Accuracy	54,17%	37,50%

Prompt 2		
	Before Fine-Tuning	After Fine-Tuning
False Positive	11	7
False Negative	0	8
Accuracy	58,33%	41,67%

Final Considerations



In conclusion **Llama2's performance** consistently fell **below human levels**, even ranking **lower** than the **worst human performance**. **Despite efforts** to **enhance** its **capabilities**, including prompt variations and fine-tuning, Llama2's accuracy remained significantly behind that of humans. This highlights the challenges in achieving human-level accuracy with current AI models and emphasizes the need for ongoing advancements in the field.

Thank you for the attention