UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO MATEMATICA

Cognitive, Behavioral and
Scoial Data: Report

Fleeting Connections: An Inquiry into Analogical Reasoning

University of Padua

Department of Mathematics

Sofia Bazzan	mat. 2089076
Simone Bianchi	mat. 2076677

Abstract

*Addressing issues often requires identifying new **connections** between concepts or events that initially seemed unrelated. Innovative solutions of this kind are based on **analogical reasoning**, which is a relationship-centered thinking process and plays a crucial role in human **problem-solving** and **decision-making**. The challenges in analogical reasoning research span from understanding the neural mechanisms involved to developing computational models that can mimic human-like reasoning processes. Bridging the gap between seemingly **unrelated domains** often requires identifying and representing the underlying structural similarities.*
*In this analysis, we applied a similar procedure as described by Green et al. 2010 [4]. However, we gave the analogies of the original experiment to the Large Language Model **(LLM) Llama2** instead of humans. For this reason, while the authors also measured and evaluated human reaction times in relation to the semantic distance between words, we decided to evaluate the model only on response **accuracy** with the **purpose** of comparing Llama2's performance with that achieved by humans.*

1 Introduction

Analogical reasoning is a cognitive process that involves finding similarities between two or more things that are otherwise dissimilar. It is an important tool for problem-solving because it allows us to transfer knowledge and insights from one domain to another, even when the two domains seem unrelated at first glance. By identifying commonalities between seemingly disparate concepts, we can generate new ideas, make predictions, and solve problems more effectively. In this context, Green et al. [4], article delves into the study of analogical reasoning, contributing valuable insights to our understanding of this cognitive process and its role in effective problem-solving. Their research explores the identification of similarities between seemingly dissimilar entities. In their work they tried to find answers to the following questions:

1. *How does **semantic distance** affect analogical mapping, and what are the **neural correlates** of this process?*

2. *What are the implications of this research for **understanding creativity** and **innovation** in the brain?*

In response to the first question, the insights gleaned from the article written by Green et al. [4] show that semantic distance affects analogical mapping by modulating activity in the frontopolar cortex, a region of the brain that is important for relational reasoning and cognitive flexibility. Specifically, the study found that as the semantic distance between two concepts in an analogy increased, activity in the frontopolar cortex also increased. This suggests that the brain recruits this region to integrate and map semantically distant concepts in the service of analogical reasoning. These findings provide a first empirical characterization of how the brain mediates semantically distant analogical mapping.
Moving on to the second question, the implications of the article written by Green et al. [4] fall within the realms of creativity and innovation in the brain. Specifically, the study found that identifying similarities across greater semantic distance reveals connections that support more innovative solutions and models. This suggests that the ability to engage in semantically distant analogical mapping is a key cognitive process that underlies creative thinking and problem-solving. By shedding light on the neural substrates of this process, this research provides a foundation for

future studies that seek to understand the cognitive and neural mechanisms that support creativity and innovation in the brain.

2 The Study

The study involved 107 people: 23 all right-handed and native English speakers (12 males, M = 22.2 years), recruited from the nearby university community to take part in the functional magnetic resonance imaging (**fMRI**) investigation; 84 native English-speaking university students (18 males, M = 21.8 years) took part in the evaluation of the stimuli.

2.1 Stimuli and Procedure

During four event-related fMRI sessions, participants completed 120 analogy tasks. In each task, participants determined whether a set of four words formed a valid analogy by indicating their response as **"true"** or **"false"** using the index or middle finger of their right hand. For the experiment, the attendees have up to 8 s to respond, but they have to prioritize precision over speed, so they should strive to respond as accurately as possible. The properties of the visual system remained consistent across different stimuli; all stimuli were presented in sets of exactly four words, resulting in the formation of a rectangle measuring $15x6$ cm from the midpoints of the four words, in a visual angle of 7.5 degrees.
Latent semantic analysis (**LSA**) was used to determine the semantic distance value for each analogy item (Landauer and Dumais 1997 [8]; Landauer et al. 1998 [10]). Specifically, the analysis involved performing pairwise comparisons between the pairs of words that make up the left and right halves of each analogy. The Latent Semantic Analysis (LSA) application, available at http://lsa.colorado.edu, determines the similarity between the contextual meanings of words by evaluating the cosine of the angle between the vectors assigned to those words in a "semantics" high dimensional space. This space includes a large corpus of English texts, providing a quantitative measure of the semantic relationships between words based on their contextual usage. A vector is included for input of multiple words, such as the word pairs that make up our analogy stimuli. For the main parametric analysis, semantic distance values were used: in particular, these values enabled to identify brain areas where there was a parametric correlation between semantic distance and activity related to the stimulus. Moreover, a total of 84 independent raters used a 7-point scale to evaluate the difficulty of all analogy stimuli. These ratings were used as a parametric regressor in subsequent fMRI analysis, in which semantic distance values were assessed for their correlation with rated difficulty. The stimuli were labeled into two classes: **within-domain** analogies (involving mapping between items from similar semantic domains) and **cross-domain** analogies (involving mapping between items from disparate semantic domains). The classification was based on a binary rating obtained from 84 independent assessors, with over 90% agreement. The experiment aimed to test the hypothesis that the frontopolar cortex reflects a taxonomy in analogy literature, specifically the distinction between within-domain and cross-domain analogies (Holyoak and Thagard 1995 [7]; Barnett and Ceci 2002 [5]; Bowdle and Gentner 2005 [6]). Equal numbers of stimuli for each class were used, and the stimuli were also categorized as true or false with over 90% agreement. **False** analogies were included as a manipulation check. To minimize confounding, the same "base" word pair was used in one trial of each stimulus type, with all words standardized for various linguistic features using the MRC Psycholinguistic Database (Wilson 1988 [9]). The complete set of stimuli was divided into subsets of 3 stimuli: they shared the same left word pair (1 cross-domain analogy, 1 within-domain analogy,

and 1 false analogy; see Fig. 1). The order of trials was pseudorandom, ensuring that consecutive trials did not involve the same pair of words on the left (Avoid consecutive presentations from the same 3-stimuli subset). In addition, the order was counterbalanced to ensure equal likelihood of the cross-domain analogy being presented as the first, second, or third trial within its respective 3-stimuli subset.

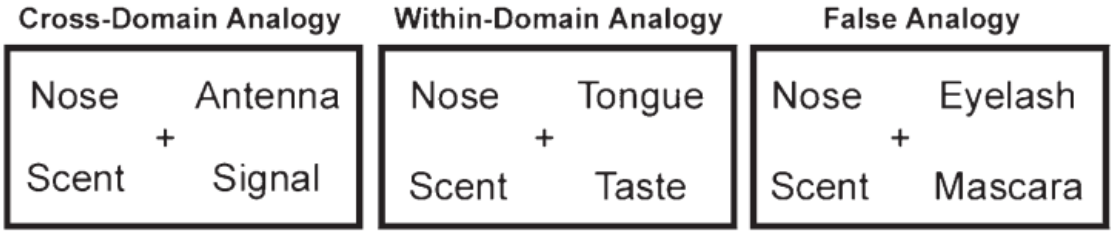


Figure 1: Example stimuli. The figure displays examples of the 3 classes of trials presented to participants during fMRI. The same word pair appeared on the left side in 1 trial of each class.

2.2 Dataset description

The dataset provided is a classification of the various pairs of stimuli based on certain criteria. Each row represents a pair of words (e.g., "answer:riddle::key:lock") along with information about the relationship between the words and a measure of difficulty for a given task.
Here's a breakdown of the columns:

1. **Stimulus Class:**
    - "Cross-Domain": Pairs of words with different semantic domains.
    - "Within-Domain": Pairs of words within the same semantic domain.
    - "FALSE": Pairs of words with no clear semantic relationship.
2. **Semantic Distance:**
    - • A numerical value, calculated using Latent Semantic Analysis, indicating the semantic distance between the word pairs. This value measures how closely related or associated the words are.
3. **Rated Difficulty:**
    - • A numerical measure, computed by aggregating the difficulty scores assigned by 84 independent raters, indicating the perceived task difficulty related to word pairs. This subjective score, assigned to each pair, reflects the level of difficulty in establishing the relationship between the words.

In particular, the dataset includes several pairs of words, and for each pair, there is information on the semantic relationship, the distance between them and the perceived difficulty of the task. The categories "Cross-Domain" and "Within-Domain" suggest an interest in exploring how semantic

relatedness and difficulty might vary depending on whether words belong to different or the same semantic domains. The "FALSE" category is a control or base set without a clear semantic relationship.

Stimuli	Stimulus Class	Semantic Distance	Rated Difficulty
answer:riddle::key:lock	Cross-Domain	0.82	3.53
ash:fireplace::lint:pocket	Cross-Domain	0.81	3.82
aspirin:pain::muffler:noise	Cross-Domain	0.89	4.55
baker:cake::scientist:discovery	Cross-Domain	0.92	3.14
basket:picnic::holster:gun	Cross-Domain	0.81	4.43
basketball:hoop::traveler:destination	Cross-Domain	0.94	2.8
blindness:sight::poverty:money	Cross-Domain	0.88	3.97
blizzard:snowflake::army:soldier	Cross-Domain	0.99	3.87
bracelet:wrist::moat:castle	Cross-Domain	0.88	4.23
burger:bun::book:cover	Cross-Domain	0.92	3.47
cleanser:face::absolution:sinner	Cross-Domain	0.83	4.16
eraser:pencil::amnesia:memory	Cross-Domain	0.91	2.54
father:son::inventor:invention	Cross-Domain	0.91	3.39
flock:goose::constellation:star	Cross-Domain	0.96	3.73
foresight:future::x-ray:bone	Cross-Domain	0.98	4.54
foundation:house::premise:argument	Cross-Domain	0.97	5.78
furnace:coal::stomach::food	Cross-Domain	0.91	3.77
hoof:hoofprint::introduction:impression	Cross-Domain	0.91	6.04
immunization:disease::forewarning:surprise	Cross-Domain	0.98	5.5
jacket:zipper::wound:suture	Cross-Domain	0.72	3.72
ketchup:tomato::fuel:petroleum	Cross-Domain	0.98	4.52
kitten:cat::spark:fire	Cross-Domain	0.93	3.82
knee:kneepad::snail:shell	Cross-Domain	0.9	2.07
lambchop:lamb::chapter:book	Cross-Domain	0.95	4.28
landscaper:lawn::stylist:hair	Cross-Domain	0.83	3.68
launchpad:helicopter::divingboard:diver	Cross-Domain	0.83	3.98
lawschool:lawyer::vineyard:wine	Cross-Domain	0.95	6.74
movie:screen::lightning:sky	Cross-Domain	0.89	5.44
multiplication:product::brewing:beer	Cross-Domain	0.9	6.94
nose:scent::antenna:signal	Cross-Domain	0.93	4.92
orchard:apple::neighborhood:apartment	Cross-Domain	0.9	6.78
painting:canvas::birthmark:skin	Cross-Domain	0.97	3.85
pen:pig::reservoir:water	Cross-Domain	0.89	5.06
rectangle:perimeter::nation:border	Cross-Domain	0.97	3.73
revising:manuscript::evolving:species	Cross-Domain	0.98	3.4
saxophone:jazz::typewriter:poetry	Cross-Domain	0.85	3.36
sugar:coffee::incentive:deal	Cross-Domain	0.86	6.23
thermometer:temperature::polygraph:honesty	Cross-Domain	0.99	3.73
train:track::signal:wire	Cross-Domain	0.86	4.6
watermelon:rind::cigarette:butt	Cross-Domain	0.99	4.21
answer:riddle::solution:problem	Within-Domain	0.55	3.55
ash:fireplace::soot:chimney	Within-Domain	0.34	2.42
aspirin:pain::antacid:heartburn	Within-Domain	0.91	2.22
baker:cake::chef:meal	Within-Domain	0.54	1.69
basket:picnic::lunchbox:lunch	Within-Domain	0.57	2.27
basketball:hoop::soccerball:goal	Within-Domain	0.6	2.83
blindness:sight::deafness:hearing	Within-Domain	0.57	3.8
blizzard:snowflake::monsoon:raindrop	Within-Domain	0.78	3.68
Bracelet:wrist::ring:finger	Within-Domain	0.52	2.83
burger:bun::sub:roll	Within-Domain	0.59	3.56
cleanser:face::soap:body	Within-Domain	0.85	2.49
eraser:pencil::whiteout:pen	Within-Domain	0.61	2.47
father:son::mother:daughter	Within-Domain	0.41	1.49
flock:goose::wolfpack:wolf	Within-Domain	0.56	2.1
foresight:future::hindsight:past	Within-Domain	0.43	2.64
foundation:house::base:structure	Within-Domain	0.47	4.78
furnace:coal::woodstove:wood	Within-Domain	0.75	2.01
hoof:hoofprint::paw:pawprint	Within-Domain	0.81	3.12
immunization:disease::vaccination:infection	Within-Domain	0.11	2.3
jacket:zipper::overcoat:button	Within-Domain	0.45	2.32
ketchup:tomato::guacamole:avocado	Within-Domain	0.57	2.75
kitten:cat::puppy:dog	Within-Domain	0.61	1.45
knee:kneepad::elbow::elbowpad	Within-Domain	0.26	2.95
lambchop:lamb::porkchop:pig	Within-Domain	0.35	2.7
landscaper:lawn::gardener:garden	Within-Domain	0.44	3.41
launchpad:helicopter::runway:airplane	Within-Domain	0.35	3.05
lawschool:lawyer::medschool:doctor	Within-Domain	0.43	2.04
movie:screen::gameshow:television	Within-Domain	0.44	2.46
multiplication:product::addition:sum	Within-Domain	0.68	2.57
nose:scent::tongue:taste	Within-Domain	0.41	2.47
orchard:apple::grove:orange	Within-Domain	0.48	2.7
painting:canvas::drawing:paper	Within-Domain	0.81	2.84
pen:pig::coop:chicken	Within-Domain	0.62	2.81
rectangle:perimeter::circle:circumference	Within-Domain	0.56	2.16
revising:manuscript::editing:story	Within-Domain	0.83	2.76
saxophone:jazz::harmonica:blues	Within-Domain	0.18	2.88
sugar:coffee::honey:tea	Within-Domain	0.54	2.13
thermometer:temperature::barometer:pressure	Within-Domain	0.69	4.51
train:track::trolley:rail	Within-Domain	0.43	2.96

watermelon:rind::orange:peel	Within-Domain	0.75	3.7
answer:riddle::jersey:number	FALSE	0.84	3.34
ash:fireplace::harness:climber	FALSE	0.81	2.41
aspirin:pain::foot:sock	FALSE	0.85	3.61
baker:cake::muffin:blueberry	FALSE	0.5	3.13
basket:picnic::knife:napkin	FALSE	0.71	2.02
basketball:hoop::serve:volley	FALSE	0.85	3.78
blindness:sight::wall:paint	FALSE	0.79	3.59
blizzard:snowflake::tornado:cloud	FALSE	0.66	4.37
bracelet:wrist::skill::practice	FALSE	0.85	3.92
burger:bun::onion:lettuce	FALSE	0.81	3.87
cleanser:face::curtain:shower	FALSE	0.63	2.53
eraser:pencil::glue:paper	FALSE	0.33	3.1
father:son::nephew:cousin	FALSE	0.59	2.09
flock:goose::pond:turtle	FALSE	0.71	2.38
foresight:future::letter:mailman	FALSE	0.96	4.48
foundation:house::duplex:renter	FALSE	0.84	3.65
furnace:coal::beach:ocean	FALSE	0.97	2.39
hoof:hoofprint::battery:toy	FALSE	0.98	2.34
immunization:disease::hotel:innkeeper	FALSE	0.98	3.47
jacket:zipper::actor:film	FALSE	0.92	2.63
ketchup:tomato::shoelace:skate	FALSE	0.9	2.65
kitten:cat::hamster:wheel	FALSE	0.94	3.88
knee:kneepad::flag:flagpole	FALSE	0.8	2.85
lambchop:lamb::fillet:skillet	FALSE	0.38	3.53
landscaper:lawn::fence:field	FALSE	0.69	2.37
launchpad:helicopter::thorn:rose	FALSE	0.76	1.46
lawschool:lawyer::beard:razor	FALSE	0.93	3.37
movie:screen::metal:rust	FALSE	0.95	5.79
multiplication:product::sleep:pajamas	FALSE	0.97	2.5
nose:scent::eyelash:mascara	FALSE	0.75	2.62
orchard:apple::cantaloupe:farmstand	FALSE	0.88	4
painting:canvas::mistake:regret	FALSE	0.94	5.96
pen:pig::hay:horse	FALSE	0.73	2.5
rectangle:perimeter::octagon:angle	FALSE	0.57	3.45
revising:manuscript::price:sale	FALSE	0.98	2.16
saxophone:jazz::document:copier	FALSE	0.99	2.67
sugar:coffee::grinder:bean	FALSE	0.92	3.85
thermometer:temperature::table:leg	FALSE	0.96	3.2
train:track::conductor:whistle	FALSE	0.69	2.23
watermelon:rind::raspberry:bush	FALSE	0.78	2.05

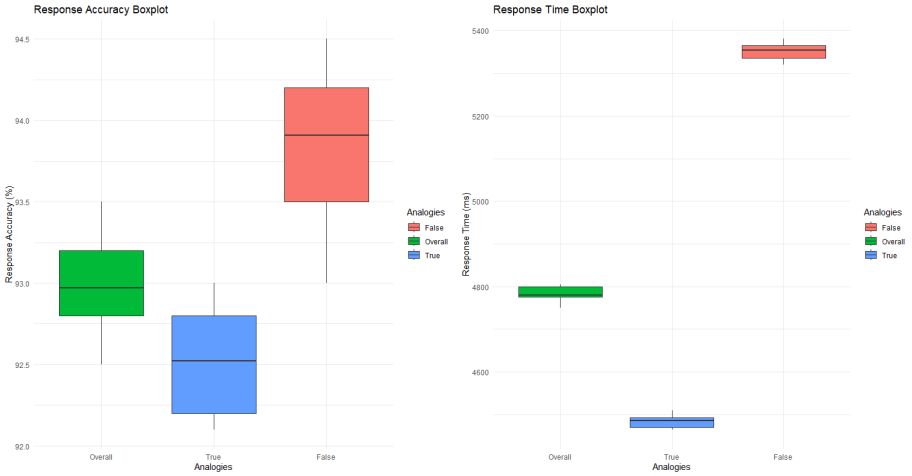
2.3 fMRI Data Acquisition and Analysis

Since our primary objective is to compare the model's performance with human performance on analogical reasoning trials, we provide a concise overview of the acquisition and analysis of the fMRI data. This brief explanation aims to offer a clearer overview of the work conducted in the study.
Data were collected using a General Electric Medical Systems Signa 1.5T full-body scanner: subjects underwent various preprocessing steps such as realignment, coregistration, normalization, and spatial smoothing.
On the other hand, in order to analyze the functional imaging was using SPM99 software, employing a mixed blocked/event-related design to distinguish different analogy tasks. Contrast images were produced for individual subjects, which were then used in a random effects analysis to create group average images, and a small volume correction within the frontopolar cortex region of interest (ROI) was applied to specific statistical maps. During the primary analysis, semantic distance values for analogies served as a parametric regression to investigate their predictive impact on brain activity within the frontopolar region of interest (ROI): this analysisadjusted the predicted hemodynamic response according to the semantic distance value associated with each analogy.
Results were subjected to a second-level random effects group analysis, with specific thresholds applied to identify significant findings. Furthermore, less conservative parametric analyses were performed, considering several factors related to the difficulty of the analogies to evaluate their influence on frontopolar cortex activity.

2.4 Results

2.4.1 Behavioral Outcomes

Participants carried out a response accuracy level of 92.97% ± 4.62% **standard deviation** (SD) and a **standard error** (SE) of 0.42% overall. For true trials, accuracy was 92.52%± SD = 4.21%, SE = 0.47%, and for false trials it was 93.91%± SD = 5.38%, SE = 0.85%. Item analysis revealed no correlation between response accuracy and semantic distance values (r = –0.18, P = 0.11). In terms of response time, participants showed an average of 4780 SD = 503, SE = 54 ms overall. For true trials, the response time was 4493 SD = 430, SE = 51 ms, while for false trials it was 5354 SD = 591, SE = 93 ms. A positive correlation was found between response time and semantic distance (r = 0.36, P = 0.001).



2.4.2 fMRI Outcomes

The study of Green et al. [4] found that the recruitment of the frontopolar cortex strengthened with an increase in semantic distance during analogical mapping (see Fig. 2). To investigate the relationship between semantic distance and frontopolar activity, the researchers used semantic distance values as a parametric regressor in the design matrix for each analogy stimulus item. The analysis focused on a predefined region of interest (ROI) in the left frontopolar cortex associated with analogical reasoning (Green, Fugelsang,Kraemer, et al. 2006 [2]). The results, obtained through small volume correction (SVC), indicated that semantic distance positively influenced neural activity within the specified ROI.

To dissociate the effect of semantic distance from difficulty, the researchers analyzed response time, correctness, and perceived difficulty in relation to the semantic distance of the stimuli. Using the residual variances from this analysis as a parametric regressor in the SPM, they confirmed that these residuals significantly predicted neural activity in the frontopolar region of interest, suggesting that difficulty-related factors may not account for the relationship between semantic distance and frontopolar activity. In contrast, activity in other brain regions such as the anterior cingulate, caudate head, and inferior occipital gyrus was not significantly affected by semantic distance once difficulty-related factors were taken into account.

In addition, focusing on the analysis of semantic distance in analogical mapping, researchers examined the neural basis of two types of analogies: cross-domain and within-domain (Barnett and Ceci 2002 [5]; Bowdle and Gentner 2005 [6]; Green, Fugelsang, and Dunbar 2006 [3]; Green et al. 2008 [1]). Cross-domain analogies, which involve concepts from different domains, show significantly greater recruitment in the frontopolar region than within-domain analogies, which involve concepts within the same domain. A direct comparison revealed in the main parametric analysis highlights that cross-domain analogies show significantly higher semantic distance values than within-domain analogies.This relationship is further confirmed by the significant correlation between semantic distance and activity in the frontopolar region, as indicated by separate analyzes of between-domain and within-domain analogy studies. Moreover, the investigation extends to the analysis of true and false trials, revealing that semantic distance positively modulates activity in the frontopolar region for both categories of trials. A direct contrast between true and false trials indicates a marginally higher level of frontopolar activity in true trials, suggesting that the analogical mapping involved in these trials may require a greater relational integration component (Green, Fugelsang, Kraemer, et al. 2006 [2]).
Furthermore, the relationship between semantic distance and difficulty in participants' responses to analogical stimuli was evaluated. Response times of each participant were used as an individual-level parameter and subsequently aggregated at the group level. A parametric analysis of semantic distance was conducted at the group level, using an exclusive mask related on response times. The results indicate that the preferential recruitment of the frontopolar cortex for semantically more distant analogies is not affected by response time: similar analyzes based on difficulty and average percentage of correct responses showed no significant modulations in the frontopolar cortex.

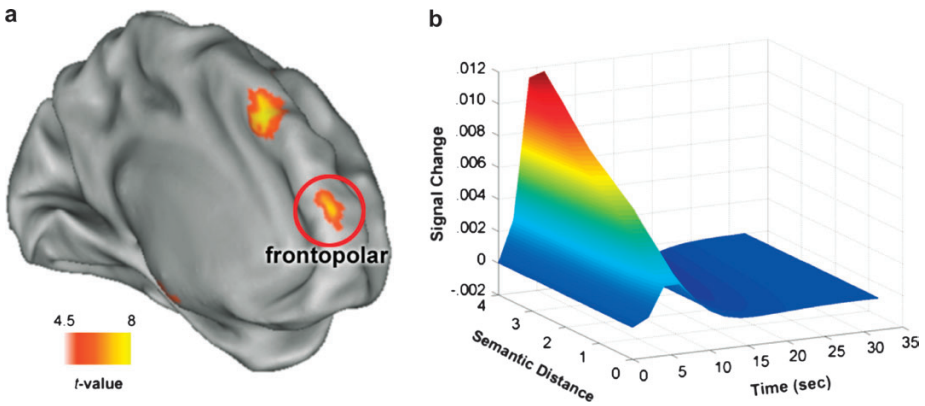


Figure 2: Neural response to semantic distance in reasoning. (A) The orange areas on an inflated cortical representation of the left hemisphere indicate brain activity; Parametric analysis identified regions with increased activation for analogies that were more semantically distant. (B) The signal change (y-axis) within the frontopolar region of interest (ROI) rises progressively over time (x-axis) in correlation with the growing semantic distance (z-axis)

2.5 Final Considerations

The results obtained from the study by Green et al. in 2010 [4] provide an initial empirical characterization of how the brain manages semantically distant analogical mapping. The findings suggest a direct relationship between the semantic distance of analogical mapping and the activation

of the frontopolar cortex. Analyses show that the semantic distance values of analogical stimuli significantly predict activity in a specific frontopolar region previously associated with analogical mapping. Fundamentally, it is the semantic distance of the analogical mapping that influences activity in this region, not the task difficulty measured by response time, correctness, and perceived difficulty.

3 Our Investigation

As mentioned before our purpose is to compare the model's performance with human performance. For our analyses we used **Llama-2** model, a family of generative text models that are optimized for assistant-like chat use cases or can be adapted for a variety of natural language generation tasks: code Llama models are fine-tuned for programming tasks.
We conducted tests on Llama-2, adjusting parameters and prompts during the same analogy trials given to the participants in Green's experiment. The goal was to assess whether the artificial intelligence's performance was comparable to that achieved by humans:

- We used two models, Llama-2 7B and Llama-2 13B (those available on the Gradient platform).
- We employed three prompts, namely, "Which is the analogical reasoning of **a** is to **b** as **c** is to **d**?" (Prompt 1), "A **a** is to **b** exactly what a **c** is to **d**?"
  (Prompt 2) and "The relationship between **a** and **b** is the same that there is between **c** and **d**?" (Prompt 3) in order to try to improve accuracy.
- We employed two distinct experimental setups, each with its own categorization. In the first experiment, we presented the model with the task of choosing among three categories: "Cross-Domain" analogies, "Within-Domain" analogies, and "False" analogies. In the second experiment, the model was tasked with selecting between two categories: "True" and "False" analogies. This dual approach allowed us to assess the model's performance under different classification scenarios.

Where **a**, **b**, **c** and **d** are the words to be compared.

3.1 Methodology

The following sub-section provides a detailed insight into the approach adopted to conduct our analysis:

1) **Dataset construction:** initially, we formed a dataset by pairing instructional sentences with corresponding responses, ensuring a wide range of linguistic scenarios (e.g., "inputs": "`###` Instruction: Which is the relation of answer is to riddle as key is to lock? `###` Response: Cross-Domain"). In addition, to achieve a more authentic replication of the experiment, we randomly mixed up the dataset.
2) **Model interaction:** We accessed the Llama-2 model through the Gradient platform (`https://gradient.ai`), enabling us to provide a unique input prompt for each individual element within the dataset. This allowed us to receive responses from the model, generating comprehensive results tailored to the specific characteristics of each dataset element.

3) **Store model-generated responses:** Each response was carefully examined for specific terms such as "Cross-Domain," "Within-Domain," or "FALSE," when the task involved choosing between three categories or "TRUE"/"FALSE" with two possible categories. This research phase aims to categorize each response based on the identified keywords. Subsequently, the categorized responses are stored into a vector. This approach not only ensures the precise identification of relevant information but also significantly facilitates comparison with responses provided by humans.

4) **Comparison with the groundtruth:** The resultant vector from the previous step was then compared in relation to the accurate and correct response corresponding to the posed question. It is noteworthy to mention that, considering the model's predominant use of "True" or "False" responses with occasional explanations, we assigned a value of 1 to a correct prediction and 0 to an incorrect prediction during the calculation of accuracy. Additionally, we computed the number of false positives and false negatives to provide a more comprehensive understanding of the model's performance and potential areas of improvement. This method allows for an appropriate evaluation of the model's performance based on its response patterns.

3.2 Dataset

For example, we present a portion of the dataset in which we have considered two distinct categories("TRUE" and "FALSE") to optimize the responsiveness of the Llama2 model:

```
dataset = [
 {'inputs': '### Instruction: A burger is to bun exactly what a book is to cover? ### Response: TRUE'},
 {'inputs': '### Instruction: A basket is to picnic exactly what a knife is to napkin? ### Response: FALSE'},
 {'inputs': '### Instruction: A foundation is to house exactly what a duplex is to renter? ### Response: FALSE'},
 {'inputs': '### Instruction: A revising is to manuscript exactly what a price is to sale? ### Response: FALSE'},
 {'inputs': '### Instruction: A father is to son exactly what a nephew is to cousin? ### Response: FALSE'},
 {'inputs': '### Instruction: A orchard is to apple exactly what a neighborhood is to apartment? ### Response: TRUE'},
 {'inputs': '### Instruction: A flock is to goose exactly what a pond is to turtle? ### Response: FALSE'},
 {'inputs': '### Instruction: A blizzard is to snowflake exactly what a tornado is to cloud? ### Response: FALSE'},
 {'inputs': '### Instruction: A furnace is to coal exactly what a beach is to ocean? ### Response: FALSE'},
 {'inputs': '### Instruction: A nose is to scent exactly what a eyelash is to mascara? ### Response: FALSE'},
 {'inputs': '### Instruction: A blizzard is to snowflake exactly what a monsoon is to raindrop? ### Response: TRUE'},
 {'inputs': '### Instruction: A watermelon is to rind exactly what a orange is to peel? ### Response: TRUE'},
 {'inputs': '### Instruction: A bracelet is to wrist exactly what a skill is to practice? ### Response: FALSE'},
 {'inputs': '### Instruction: A pen is to pig exactly what a reservoir is to water? ### Response: TRUE'},
 {'inputs': '### Instruction: A furnace is to coal exactly what a stomach is to food? ### Response: TRUE'},
 {'inputs': '### Instruction: A knee is to kneepad exactly what a elbow is to elbowpad? ### Response: TRUE'},
 {'inputs': '### Instruction: A aspirin is to pain exactly what a foot is to sock? ### Response: FALSE'},
 {'inputs': '### Instruction: A hoof is to hoofprint exactly what a battery is to toy? ### Response: FALSE'},
 {'inputs': '### Instruction: A eraser is to pencil exactly what a whiteout is to pen? ### Response: TRUE'},
 {'inputs': '### Instruction: A multiplication is to product exactly what a addition is to sum? ### Response: TRUE'},
 {'inputs': '### Instruction: A bracelet is to wrist exactly what a moat is to castle? ### Response: TRUE'},
 {'inputs': '### Instruction: A kitten is to cat exactly what a puppy is to dog? ### Response: TRUE'},
 {'inputs': '### Instruction: A immunization is to disease exactly what a hotel is to innkeeper? ### Response: FALSE'},
 {'inputs': '### Instruction: A multiplication is to product exactly what a sleep is to pajamas? ### Response: FALSE'},
 {'inputs': '### Instruction: A blizzard is to snowflake exactly what a army is to soldier? ### Response: TRUE'},
 {'inputs': '### Instruction: A ash is to fireplace exactly what a lint is to pocket? ### Response: TRUE'},
 {'inputs': '### Instruction: A train is to track exactly what a trolley is to rail? ### Response: TRUE'},
 {'inputs': '### Instruction: A orchard is to apple exactly what a grove is to orange? ### Response: TRUE'},
 {'inputs': '### Instruction: A lawschool is to lawyer exactly what a vineyard is to wine? ### Response: TRUE'},
 {'inputs': '### Instruction: A answer is to riddle exactly what a jersey is to number? ### Response: FALSE'},
 {'inputs': '### Instruction: A kitten is to cat exactly what a spark is to fire? ### Response: TRUE'},
 {'inputs': '### Instruction: A father is to son exactly what a mother is to daughter? ### Response: TRUE'},
 {'inputs': '### Instruction: A ash is to fireplace exactly what a soot is to chimney? ### Response: TRUE'},
 {'inputs': '### Instruction: A cleanser is to face exactly what a soap is to body? ### Response: TRUE'},
 {'inputs': '### Instruction: A lawschool is to lawyer exactly what a medschool is to doctor? ### Response: TRUE'},
 {'inputs': '### Instruction: A watermelon is to rind exactly what a cigarette is to butt? ### Response: TRUE'},
 {'inputs': '### Instruction: A movie is to screen exactly what a gameshow is to television? ### Response: TRUE'},
 {'inputs': '### Instruction: A revising is to manuscript exactly what a evolving is to species? ### Response: TRUE'},
 {'inputs': '### Instruction: A ketchup is to tomato exactly what a guacamole is to avocado? ### Response: TRUE'},
 {'inputs': '### Instruction: A train is to track exactly what a signal is to wire? ### Response: TRUE'},
 {'inputs': '### Instruction: A jacket is to zipper exactly what a overcoat is to button? ### Response: TRUE'},
 {'inputs': '### Instruction: A foundation is to house exactly what a base is to structure? ### Response: TRUE'},
 {'inputs': '### Instruction: A sugar is to coffee exactly what a honey is to tea? ### Response: TRUE'},
 {'inputs': '### Instruction: A knee is to kneepad exactly what a snail is to shell? ### Response: TRUE'},
 {'inputs': '### Instruction: A lambchop is to lamb exactly what a porkchop is to pig? ### Response: TRUE'},
 {'inputs': '### Instruction: A launchpad is to helicopter exactly what a runway is to airplane? ### Response: TRUE'},
 {'inputs': '### Instruction: A lambchop is to lamb exactly what a chapter is to book? ### Response: TRUE'},
 {'inputs': '### Instruction: A answer is to riddle exactly what a solution is to problem? ### Response: TRUE'},
 {'inputs': '### Instruction: A jacket is to zipper exactly what a actor is to film? ### Response: FALSE'},
```

```
{'inputs': '### Instruction: A pen is to pig exactly what a hay is to horse? ### Response: FALSE'},
{'inputs': '### Instruction: A landscaper is to lawn exactly what a stylist is to hair? ### Response: TRUE'},
{'inputs': '### Instruction: A blindness is to sight exactly what a poverty is to money? ### Response: TRUE'},
{'inputs': '### Instruction: A furnace is to coal exactly what a woodstove is to wood? ### Response: TRUE'},
{'inputs': '### Instruction: A cleanser is to face exactly what a curtain is to shower?### Response: FALSE'},
{'inputs': '### Instruction: A immunization is to disease exactly what a forewarning is to surprise? ### Response: TRUE'},
{'inputs': '### Instruction: A burger is to bun exactly what a sub is to roll? ### Response: TRUE'},
{'inputs': '### Instruction: A train is to track exactly what a conductor is to whistle? ### Response: FALSE'},
{'inputs': '### Instruction: A aspirin is to pain exactly what a antacid is to heartburn? ### Response: TRUE'},
{'inputs': '### Instruction: A aspirin is to pain exactly what a muffler is to noise? ### Response: TRUE'},
{'inputs': '### Instruction: A lambchop is to lamb exactly what a fillet is to skillet? ### Response: FALSE'},
{'inputs': '### Instruction: A pen is to pig exactly what a coop is to chicken? ### Response: TRUE'},
{'inputs': '### Instruction: A rectangle is to perimeter exactly what a circle is to circumference? ### Response: TRUE'},
{'inputs': '### Instruction: A sugar is to coffee exactly what a incentive is to deal? ### Response: TRUE'},
{'inputs': '### Instruction: A blindness is to sight exactly what a deafness is to hearing? ### Response: TRUE'},
{'inputs': '### Instruction: A thermometer is to temperature exactly what a polygraph is to honesty? ### Response: TRUE'},
{'inputs': '### Instruction: A cleanser is to face exactly what a absolution is to sinner? ### Response: TRUE'},
{'inputs': '### Instruction: A revising is to manuscript exactly what a editing is to story? ### Response: TRUE'},
{'inputs': '### Instruction: A foundation is to house exactly what a premise is to argument? ### Response: TRUE'},
{'inputs': '### Instruction: A immunization is to disease exactly what a vaccination is to infection? ### Response: TRUE'},
{'inputs': '### Instruction: A basketball is to hoop exactly what a traveler is to destination? ### Response: TRUE'},
{'inputs': '### Instruction: A rectangle is to perimeter exactly what a nation is to border? ### Response: TRUE'},
{'inputs': '### Instruction: A basket is to picnic exactly what a lunchbox is to lunch? ### Response: TRUE'},
{'inputs': '### Instruction: A basketball is to hoop exactly what a soccerball is to goal? ### Response: TRUE'},
{'inputs': '### Instruction: A father is to son exactly what a inventor is to invention? ### Response: TRUE'},
{'inputs': '### Instruction: A Bracelet is to wrist exactly what a ring is to finger? ### Response: TRUE'},
{'inputs': '### Instruction: A kitten is to cat exactly what a hamster is to wheel? ### Response: FALSE'},
{'inputs': '### Instruction: A orchard is to apple exactly what a cantaloupe is to farmstand? ### Response: FALSE'},
{'inputs': '### Instruction: A watermelon is to rind exactly what a raspberry is to bush? ### Response: FALSE'},
{'inputs': '### Instruction: A ketchup is to tomato exactly what a shoelace is to skate? ### Response: FALSE'},
{'inputs': '### Instruction: A jacket is to zipper exactly what a wound is to suture? ### Response: TRUE'},
{'inputs': '### Instruction: A basketball is to hoop exactly what a serve is to volley? ### Response: FALSE'},
{'inputs': '### Instruction: A baker is to cake exactly what a chef is to meal? ### Response: TRUE'},
{'inputs': '### Instruction: A landscaper is to lawn exactly what a fence is to field? ### Response: FALSE'},
{'inputs': '### Instruction: A ketchup is to tomato exactly what a fuel is to petroleum? ### Response: TRUE'},
{'inputs': '### Instruction: A answer is to riddle exactly what a key is to lock? ### Response: TRUE'},
{'inputs': '### Instruction: A saxophone is to jazz exactly what a document is to copier? ### Response: FALSE'},
{'inputs': '### Instruction: A launchpad is to helicopter exactly what a thorn is to rose? ### Response: FALSE'},
{'inputs': '### Instruction: A hoof is to hoofprint exactly what a paw is to pawprint? ### Response: TRUE'},
{'inputs': '### Instruction: A rectangle is to perimeter exactly what a octagon is to angle? ### Response: FALSE'},
{'inputs': '### Instruction: A flock is to goose exactly what a wolfpack is to wolf? ### Response: TRUE'},
{'inputs': '### Instruction: A movie is to screen exactly what a metal is to rust? ### Response: FALSE'},
{'inputs': '### Instruction: A burger is to bun exactly what a onion is to lettuce? ### Response: FALSE'},
{'inputs': '### Instruction: A baker is to cake exactly what a muffin is to blueberry? ### Response: FALSE'},
{'inputs': '### Instruction: A eraser is to pencil exactly what a glue is to paper? ### Response: FALSE'},
{'inputs': '### Instruction: A launchpad is to helicopter exactly what a divingboard is to diver? ### Response: TRUE'},
{'inputs': '### Instruction: A painting is to canvexactly what a exactly what a drawing is to paper? ### Response: TRUE'},
{'inputs': '### Instruction: A knee is to kneepad exactly what a flag is to flagpole? ### Response: FALSE'},
{'inputs': '### Instruction: A thermometer is to temperature exactly what a barometer is to pressure? ### Response: TRUE'},
{'inputs': '### Instruction: A landscaper is to lawn exactly what a gardener is to garden? ### Response: TRUE'},
{'inputs': '### Instruction: A nose is to scent exactly what a antenna is to signal? ### Response: TRUE'},
{'inputs': '### Instruction: A movie is to screen exactly what a lightning is to sky? ### Response: TRUE'},
{'inputs': '### Instruction: A lawschool is to lawyer exactly what a beard is to razor? ### Response: FALSE'},
{'inputs': '### Instruction: A thermometer is to temperature exactly what a table is to leg? ### Response: FALSE'},
{'inputs': '### Instruction: A basket is to picnic exactly what a holster is to gun? ### Response: TRUE'},
{'inputs': '### Instruction: A foresight is to future exactly what a hindsight is to past? ### Response: TRUE'},
{'inputs': '### Instruction: A saxophone is to jazz exactly what a harmonica is to blues? ### Response: TRUE'},
{'inputs': '### Instruction: A eraser is to pencil exactly what a amnesia is to memory? ### Response: TRUE'},
{'inputs': '### Instruction: A nose is to scent exactly what a tongue is to taste? ### Response: TRUE'},
{'inputs': '### Instruction: A painting is to canvas exactly what a mistake is to regret? ### Response: FALSE'},
{'inputs': '### Instruction: A blindness is to sight exactly what a wall is to paint? n### Response: FALSE'},
{'inputs': '### Instruction: A flock is to goose exactly what a constellation is to star? ### Response: TRUE'},
{'inputs': '### Instruction: A sugar is to coffee exactly what a grinder is to bean? ### Response: FALSE'},
{'inputs': '### Instruction: A hoof is to hoofprint exactly what a introduction is to impression? ### Response: TRUE'},
{'inputs': '### Instruction: A multiplication is to product exactly what a brewing is to beer? ### Response: TRUE'},
{'inputs': '### Instruction: A painting is to canvas exactly what a birthmark is to skin? ### Response: TRUE'},
{'inputs': '### Instruction: A saxophone is to jazz exactly what a typewriter is to poetry? ### Response: TRUE'},
{'inputs': '### Instruction: A foresight is to future exactly what a letter is to mailman? ### Response: FALSE'},
{'inputs': '### Instruction: A baker is to cake exactly what a scientist is to discovery? ### Response: TRUE'},
{'inputs': '### Instruction: A foresight is to future exactly what a x-ray is to bone? ### Response: TRUE'},
{'inputs': '### Instruction: A ash is to fireplace exactly what a harness is to climber? ### Response: FALSE'}
]
```

In particular, we made a modification in the formulation of the question to improve clarity and standardization. The original expression, "**###** Response: TRUE/FALSE", will be transformed when the question is asked to the Llama2 model. This change involves replacing the initial expression with a more explicit prompt: "Choose one among two: TRUE or FALSE".

3.3 Performance Outcomes Across Different Models

In this section, we explore the landscape of results obtained through the two models available on the Gradient platform: Llama2-7B and Llama2-13B.

The main challenge we have encountered involve achieving decent levels of accuracy, particularly when compared to human performance. The problem arises because Llama2 assigns the same class to the overwhelming majority of the data. Inparticular

- In scenarios with three potential categories ("Cross-domain", "Cross-domain", and "False"), the model predominantly classifies the data as "Cross-domain" or "Within-Domain" based on the prompt.
- Likewise, in situations with two possible categories ("True" or "False"), the model tends to classify most of the data as "True".

As previously mentioned, in order to address this issue, we used different prompts. Each model has been systematically tested across all prompts and categories, ensuring a thorough evaluation of its performance under different conditions. We showcase the outcomes through the following tables and barplots.
We recalled that the prompts that we utilized are:

- 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**?

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

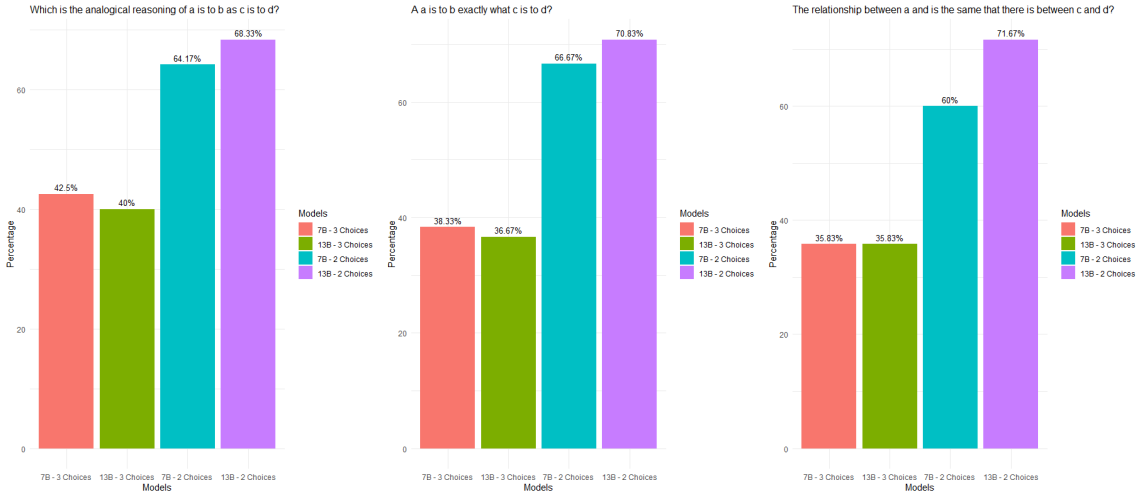
Prompt 2

	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.



3.4 Fine-tuning of the Models

Another method we attempted to improve performance involved fine-tuning the model. In this instance, we experimented with different prompts and scenarios that presented two or three potential choices. Unfortunately, due to time constraints and computational limitations, we were only able to examine a subset of the analogy trials. All relevant details are outlined below.

3.4.1 Fine-tuning with Two Possible Choices

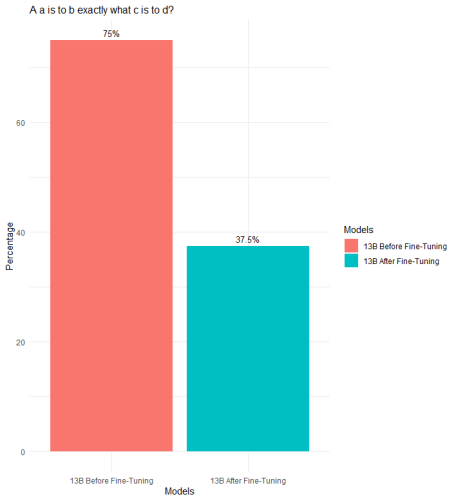
First, we attempted to fine-tune the model with two possible choices on Llama2 with 13 billion parameters. In this scenario, we initially worked with a training set comprising 96 analogy trials and a test set of 24 trials. However, due to computational constraints, we decided to focus on 72 analogy trials for the fine-tuning proces. Considering the previous outcomes, we opted to fine-tune the model using pompt2 and prompt3 since they yielded better performance (structured as "A **a** is to **b** exactly what a **c** is to **d**" and "The relationship between **a** and **b** is the same that there is between **c** and **d**?", respectively).
The results we obtained fell below our expectations. Surprisingly, the models that had previously demonstrated excellent performance had a drop in accuracy after tuning, as illustrate in the tables below:

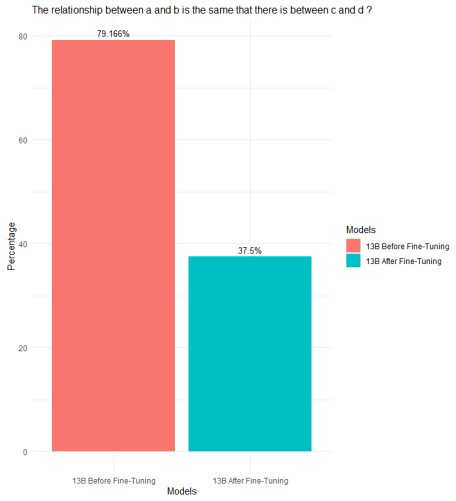
Prompt 2

	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%





In our opinion, the decrease in accuracy observed during fine-tuning may be attributed to two possible issues: overfitting and divergence from the pre-trained model.

- **Overfitting**: Overfitting occurs when the model excessively adapts to the details of the training dataset, losing the ability to generalize to new data. In this case, the model may have memorized specific features of the fine-tuning dataset but lacks the ability to apply this knowledge in broader contexts.
- **Divergence from the Pre-trained Model**: Divergence indicates a significant deviation from the original pre-trained model. If fine-tuning is too aggressive, the model may lose coherence with the knowledge acquired during pre-training, compromising overall performance.

We believe that these issues may have occurred in our investigation due to the limited size of the dataset.

3.4.2 Fine-tuning with Three Possible Choices

In addition, we attempted to fine-tune the model with three possible choices on Llama2 with 7 billion parameters. As in the scenario with two possible choices, we initially worked with a training set comprising 96 analogy trials and a test set of 24 trials. However, due to computational constraints, we decided to focus only on 40 analogy trials for the fine-tuning proces. Considering the previous outcomes, we opted to fine-tune the model using pompt1 and prompt2 since they yielded better performance (structured as "Which is the analogical reasoning of **a** is to **b** as **c** is to **d**?"and "A **a** is to **b** exactly what a **c** is to **d**", respectively).
Also in this case the results we obtained fell below our expectations. We illustrate the outcomes in the tables below:

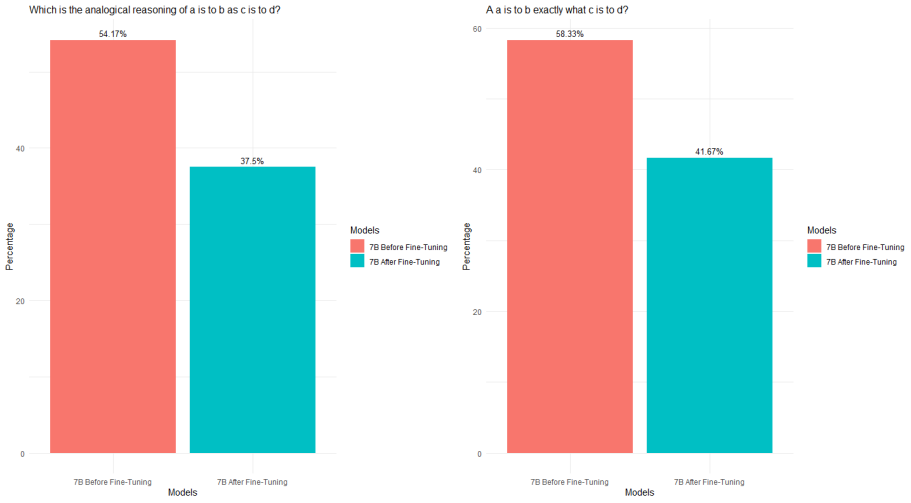
Prompt 1

	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%

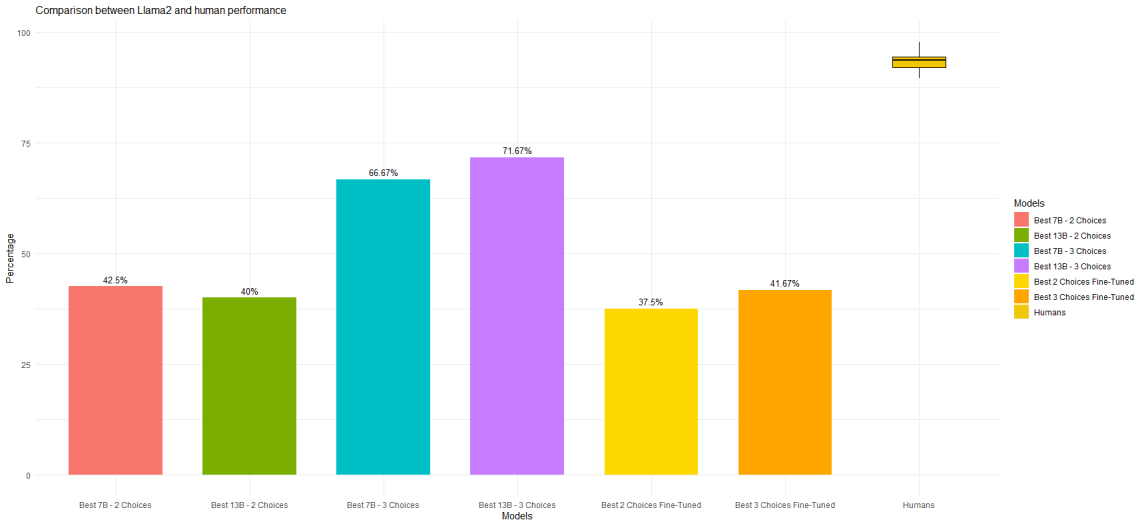
The total number of False Positives is obtained from the sum of False Positives associated with Cross-Domain and those associated to Within-Domain.



As in the context with two possible choices, the drop in accuracy observed during fine-tuning could be attributed to overfitting and divergence from the pre-trained model, therefore these challenges might have arisen in our study due to the restricted size of the dataset.

4 Final conclusions

In conclusion, our experiment has unequivocally demonstrated the superior accuracy of human performance compared to that of Llama2. Despite our diligent efforts to enhance its capabilities through prompt variations, such as the introduction of prompt3 ("The relationship between **a** and **b** is the same that there is between **c** and **d**?"), the outcomes still did not match human capabilities. It is noteworthy that humans consistently outperformed Llama2, even when using a simple analogy square, highlighting the inherent superiority of human reasoning in analogy tasks. Despite our efforts to fine-tune the model, there was no significant improvement in accuracy. Even after fine-tuning, Llama2 remained significantly behind human performance levels, as illustrated in the plot below.



Our investigation into different prompts and fine-tuning techniques has offered valuable insights into the behavior of the model, highlighting the complexity involved in bridging the gap between artificial intelligence and human-like reasoning.
The limitations we encountered in reaching human-level accuracy highlight the challenges inherent in current AI models, thus underscoring the need for continued advancements in the field.

References

[1] David J. M. Kraemer Adam E. Green, Jonathan A. Fugelsang and Kevin N. Dunbar. The micro-category account of analogy. *Cognition*, 106(2):1004–1016, 2008.

[2] David JM Kraemer Noah A. Shamosh Adam E. Green, Jonathan A. Fugelsang and Kevin N. Dunbar. Frontopolar cortex mediates abstract integration in analogy. *Brain research*, 1096(1):125–137, 2006.

[3] Jonathan A. Fugelsang Adam E. Green and Kevin N. Dunbar. Automatic activation of categorical and abstract analogical relations in analogical reasoning. *Memory & cognition*, 34:1414–1421, 2006.

[4] Jonathan A. Fugelsang Jeremy R. Gray Adam E. Green, David JM. Kraemer and Kevin N. Dunbar. Connecting long distance: semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cerebral cortex*, 20(1):70–76, 2010.

[5] Susan M. Barnett and Stephen J. Ceci. When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological bulletin*, 128(4):612–637, 2002.

[6] Brian F. Bowdle and Dedre Gentner. The career of metaphor. *Psychological review*, 112(1):193–216, 2005.

[7] Holyoak KJ. and Thagard P. Mental leaps. 1998.

[8] Thomas K. Landauer and Susan T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211–240, 1997.

[9] Wilson MD. Mrc psycholinguistic database: Machine-usable dictionary, version 2. behav res methods instrum comput. *Behavior research methods, instruments, & computers*, 20(1):6–11, 1988.

[10] Darrell Laham Thomas K. Landauer, Peter W. Foltz et al. An introduction to latent semantic analysis. *Discourse processes*, 25(2&3):259–284, 1998.