Department of Mathematics
Master degree in Data Science
University of Padua, a.y. 2023/2024

Prefrontal Functioning on Llama 2

Cognitive, Behavioural and Social Data's final project

Del Savio Anna (2097098)
Jesus Inês (2073570)
Pertile Andrea Valentina (2089070)
Putina Anna (2081379)
Tomaselli Francesco (2089207)
Virginio Giacomo (2076681)

Contents

1 Introduction

Prefrontal functioning (PF) refers to the cognitive and executive processes associated with the prefrontal cortex, a region located at the front of the brain, right behind the forehead. The prefrontal cortex is a crucial part of the cerebral cortex, responsible for higher-order cognitive functions and complex decision-making. Some key aspect of prefrontal functioning include Executive Functions (planning or organize actions, regulating emotions and adapting to changing circumstances), Working Memory (storage of information necessary for ongoing cognitive tasks), Cognitive Flexibility (adapting to new situations) and Inhibitory Control (suppressing irrelevant or inappropriate responses and maintaining focus on relevant information). Prefrontal functioning, in summary, encompasses a wide range of cognitive and executive processes that are crucial for adaptive, goal-directed behavior, emotional regulation, social interactions, and complex decision-making.

1.1 AI's Level of Intelligence: An Ongoing Debate

What intelligence is and how we define it has always been a central concern in neuropsychology. Intelligence is not a tangible entity that can be measured directly with some apparatus. We assume that intelligence is found in the brain but we cannot locate it. So, if we have a problem with the definition of human intelligence, how can we define if the AI system exhibits intelligence or if it just mimics human intelligence? According to the paper "On the measure of Intelligence" (Cholett, 2019) "the intelligence of a system is a measure of its skill-acquisition efficiency over a scope of tasks, with respect to priors, experience, and generalization difficulty". In addition he proposes a dataset based on the Raven test (Raven, 1938) which is a standard psychological test for non-verbal intelligence (the ability to solve problems, reason, and understand complex concepts without relying on language-based skills).

Large Language Models (LLM) are deep learning models that are pre-trained on very large amounts of data, going up to billions of pages from the internet. They function as next-word predictors, since they are trained to probabilistically predict the next token given the preceding text. A LLM can answer questions, perform translation, complete sentences, among other completely different and sometimes very complex tasks. It can even be used as generative AI, producing all sorts of content requested from human language inputs, whether it is text, images, videos or audio. The effectiveness and human-like performance of LLMs have been catching the attention of everyone, not only for their novelty and usefulness but also for their uncanniness. Indeed they showcase remarkable problem-solving abilities, making them particularly intriguing subjects for study within the field of cognitive psychology. And so the debate is born: Do LLMs show signs of human intelligence? Or is it only a matter of mimicking human performance? We have two main points of view: the optimistic and the skeptical.

From the optimistic perspective, researchers believe that increasing the scale of the models allows LLMs to display emergent abilities far beyond the initial training task. Emergent AI abilities refer to the unexpected, novel behaviors or skills that appear in advanced artificial intelligent systems. These abilities are not pre-trained or programmed into the AI model but rather emerge unpredictably, which doesn't

happen if we examine just smaller-scaled models.

Skeptic researchers argue that LLMs intelligence is nothing more than complex and sophisticated pattern recognition capabilities, as they aren't able to genuinely understand context and the true nuances of human communication or interaction. LLMs can for example prioritize common patterns over important rare occurrences, or fail on generalization when given tasks that fall too far outside their training data. Even so, no cognitive task has yet appeared that LLMs aren't eventually able to perform, keeping the debate open.

1.2 Llama 2

Llama 2 is a collection of open source pretrained and fine-tuned large language models ranging in scale from 7 billion to 70 billion parameters released by Meta, in partnership with Microsoft, in July 2023. The fine-tuned models are called LLaMA-2 Chat and they're optimized for dialog. Fine tuning is done with reinforcement learning from human feedback to ensure safety and helpfulness. Llama 2 is an updated version of the previous Llama 1. The new version is trained on a new mix of publicly available data and the authors pay attention to remove Web sites that often disclose personal data of people. The size of the pretraining corpus is increased by 40%, so the the context length of the model is doubled, and grouped-query attention is adopted to improve inference scalability for the bigger model's version. The authors claims that their models outperform open-source chat models on most benchmarks they have tested. We tested the model for several prefrontal functioning cognitive tests using the 70 billion online version with the following parameters:

1. temperature value of 0.75, which is an adjust for randomness of outputs,
2. 800 maximum number of tokens to generate,
3. Top P equal to 0.9; when decoding text, samples from the top p percentage of most likely tokens.

1.3 Introduction to Prefrontal functioning.

The functions tested (as well as the tests used) in this work are the ones analyzed in Loconte et al. (2023), which proposed a common framework to compare prefrontal functioning in LLMs, and are here presented.

1.3.1 Verbal Reasoning

Verbal Reasoning is a skill that characterizes human beings, it can be defined as the ability to draw inferences from given information. It is a complex function which involves multiple cognitive abilities, such as attention, language, working memory, abstraction and categorization. Difficulties in verbal reasoning are associated to brain injuries of diverse aetiology, as different areas of both cerebral hemispheres are involved in verbal reasoning tasks.

1.3.2 Cognitive estimation

Cognitive Estimation (CE) refers to the ability of a person to answer to unknown questions, that are not immediately answerable. Usually, CE is evaluated via a Cognitive Estimation Test (CT), composed by several questions where the subject must provide the answers, based more on his judgments than on precise calculations. The answers accuracy is directly linked to the Prefrontal Cortex (PC), which is responsible for executive functions. The evaluation of CE, nowadays, is evaluated also on artificial intelligence (AI), for example of Large Language Model, such as ChatGPT, Llama and Bard: the purpose is the same, with the only difference that the CET must take in account the randomness of these models (Smith and Milner, 1984; Smith and Milner, 1988; Loconte et al., 2023).

1.3.3 Metaphors and Idioms Comprehension

Metaphors comprehensions refers to the ability to understand figurative language, particularly metaphors, which involve the use of words or expressions in a way that goes beyond their literal meaning, relying on a comparison between two unrelated concepts to convey a particular idea or image. While, idioms comprehension involves the understanding of idioms, which are expressions or phrases that have a meaning different from the literal interpretation of the individual words and often depends on the context. Individuals with schizophrenia, Alzheimer's disease and right-brain lesions often struggle to understand figurative expressions; moreover, patients with schizophrenia extend this problem.

1.3.4 Anaphoric referencing

is a linguistic phenomenon where pronouns or other linguistic elements are used to refer back to a previously mentioned noun or noun phrase within a given discourse. It plays a crucial role in effective communication in human language, observed across all languages. The process of utilizing anaphoric references involves a complex cognitive mechanism, drawing upon memory, attention, language comprehension, commonsense knowledge, and reasoning. Research in linguistics and cognitive science has explored the intricacies of anaphoric referencing, highlighting the interplay of various cognitive functions in the execution of anaphoric references in language use.

An example is provided to illustrate the concept: "The cat sat on the windowsill. It enjoyed the view of the garden." In this instance, "It" is an anaphoric reference pointing back to the earlier mention of "the cat." While humans effortlessly connect the pronoun to its antecedent, this task presents a challenge for AI models, showcasing the complexity of contextual understanding required for anaphoric references.

1.3.5 Planning

Planning refers to the organization ahead of time of goal-directed behaviours (Encyclopedia of Behavioral Neuroscience, 2nd edition, 2022). In its most basic form, planning involves coming up with a course of actions (policy) which when executed would take an agent from a certain initial state to a desired world state. (Valmeekam, 2023). Tower of London test is traditionally used to assess strategical reasoning,

problem-solving, and mental planning in clinical populations (Bruni, 2022). In original ToL, participants are given a wooden board with three pegs of different lengths, and three differently coloured wooden beads inserted in the pegs. Starting from an initial, standard configuration of the beads across the pegs, individuals are asked to rearrange the beads to reproduce a target configuration showed on a picture, and by following a set of rules. The test consists of 12 problems of increasing difficulty. Solving the ToL problem within a limited number of moves requires planning the sequence of action before starting to move the beads. It is therefore considered as a planning test, although performance is also related to working memory (Welsh et al., 1995), inhibition (Morris et al., 1997), attentional control (Shallice, 1982) and some studies have demonstrated that solving the ToL also requires optimal visuospatial skills (D'antuono et al., 2017).

1.3.6 Inhibition

Inhibition is the ability to control impulsive (or automatic) responses and the ability of using attention and reasoning. inhibitory control blocks behaviours and stops inappropriate automatic reactions, changing one response for a better, more thought-out response adapted to the situation, it is how the brain corrects a behaviour. The frontal structures of the brain are the last ones to mature during development, which is why it's common to see young children have trouble controlling their behaviour and managing unexpected changes or events. verbal inhibition is a subcategory of inhibition and specifically, it is associated with increased activation of a network of left prefrontal areas and it is usually measured with the Hayling Sentence Completion Test (HSCT).

1.3.7 Insight

In cognitive psychology, insight is like a lightbulb moment where your thoughts quickly rearrange and the pieces are put together, leading to a deep understanding of a problem (Mayer, 1995).

According to Sternberg and Davidson (1995), individuals tend to approach problem solving either through insight or methodical analytic processing. The distinction between these two approaches depends on the individual's awareness of their progress, since when they rely on sudden insight, individuals often experience surprise upon reaching a solution and confidence that it is right (Metcalfe and Wiebe, 1987).

To assess and measure insight, cognitive tasks like the Remote Association Test (RAT) become instrumental.

1.3.8 Social Cognition

Social Cognition is the ensemble of various mental processes, which allow socially appropriate behavior in daily life. The cores of Social Cognition are: Theory of Mind, emotion recognition and attribution, moral and non-moral judgments, decision-making and empathy. Brain damage resulting in impairments in any of the aforementioned abilities may result in inappropriate social behavior. Deficits

have been observed in individuals with brain injuries in the prefrontal areas, as well as neurodevelopmental disorder and neurodegenerative conditions.

2 Materials and Methods

2.1 Procedure

The neuropsychological tests given here to Llama 2 are usually conducted on humans by clinics to assess prefrontal functioning. The tests were selected in the paper by Loconte et. al (2023) because they were characterized as "requiring the minimum level of cognitive efficiency to be solved", which means that a poor performance in humans indicates neurological impairment.

The tests were all given in Italian, by either instructing the model through Llama 2's System Prompt, or by including it in each item.

The prompting for each test is zero-shot, one-shot or few-shot, meaning the subject is given respectively zero, one or more than one examples; chain-of-thought prompting, meaning the answers provided in the examples are explained, so clues on the reasoning process are provided to the patient beforehand can also be included. It is proven by Wei et.al (2022b) that chain-of-thought prompting outperforms standard prompting on arithmetic, commonsense and symbolic reasoning. This is equivalent to how clear instructions given to patients might increase their performance, decreasing the misleading possibility that poor results are due to a lack of understanding.

All tests prompting procedure were kept the same as in their original clinical administration: the Hayling Sentence Completion Test as a standard one-prompt; the Compound Remote Association with a standard few-shot prompt of four examples; the Tower of London with a chain-of-thought one-shot prompt to induce multi-step reasoning; and the rest with standard zero-shot prompt.

For each test, Llama 2 was given one item at a time and each answer was taken, and by the end of the test scores were computed and compared against the normative data in the original version and research papers of said test.

Different score systems were used to evaluate the humans' performance depending on the test and respective primary research, among them z-scores, percentile ranges and cut-offs. The cut-off method consists of considering pathological behavior when the performance stands below a certain threshold. Since performance on neuropsychological tests is usually affected by a wide range of factors, from demographic to gender and age and even to level of education, causing a high variance in results, the computation of correction parameters is common. Differently to the baseline paper for the test (Loconte et al., 2023) we decided to not correct the results by gender, age or schooling, in order to have qualitative evaluations comparable to average humans, however the raw scores presented in both papers can still be compared, as those aren't corrected.

2.2 Materials

2.2.1 Verbal Reasoning Test

The Verbal Reasoning Test (VRT) is a test developed by Basagni et al. (2017) to assess human verbal reasoning.

The test consists of seven subtests assessing different verbal reasoning aspects; each subtest consists of seven items plus an initial example item.

The subtests are the following:

- **Absurdities** - the goal is to identify logical incongruence in sentences containing conflictual information (e.g. "Outside the farm there was a bright sunshine, while inside it was raining");
- **Intruders** - the goal is to identify the intruder among four words (e.g. "physician, *hospital*, dentist, nurse");
- **Relationships** - the goal is to identify the relationship between a pair of terms and to apply the same relationship to another word (e.g. "The relationship between *cold* and *hot* is the same of that between *open* and...");
- **Differences** - the goal is to identify the main characteristic that distinguishes two concepts or objects (e.g. "What is the difference between *eye* and *ear*?);
- **Idiomatic Expressions** - the goal is to explain the meaning of common idiomatic expressions (e.g. "What does 'lift your elbow' mean?");
- **Family Relations** - the goal is to specify the degree of relationship between relatives (e.g. "Lucy and Mary are sisters. Mary has a daughter, Anne. What kind of family relation is there between Lucy and Anne?");
- **Classifications** - the goal is to determine the category to which a triplet belongs (e.g. "What are Milan, Rome and Naples"?).

We administered the test and computed a total score for the VRT and each of the subtests. Raw scores can be adjusted for age and education, however we have decided not to do so, comparing our results to the overall human ones; it can be noticed that unadjusted scores are very close to scores adjusted for a person between 46 and 60 years old with between 13 and 15 years of education (high school diploma, but no university degree).

2.2.2 Cognitive estimation Task

The reference study is "How many camels are there in Italy? Cognitive estimates standardized on the Italian Population" (Della Sala et. Al., 2003), which is a paper where a cognitive test, based on 21 questions, is provided to 175 healthy subjects, stratified by sex education and age. The goal of the report is to evaluate the performance on the Cognitive Estimation Test (CET) of the subject and evaluate the relation between the score and the subject features (sex, education and age). The test is composed by 21 questions, such as "How many camels are there in Italy?", and the score is given by the distance between the given answer and the true one. A different score is given for each answer, where the lower the score the higher the performances. For example, to the previous question, a score of 0 is given for a response bounded between 28 and 52, a score of 1 is given for an answer bounded between 4 and 27 or 53 and 76, and 2 otherwise. This test is given to Llama 2 and the achieved score by the large language model is registered and compared with the results in the paper.

2.2.3 Metaphors and Idioms Comprehension Task

The reference material is composed by two different tests (proposed by Papagno et. al. in 1995), each with 20 questions and a score between 2 (maximum score) and 0 (minimum score) depending on the answer. The total score, then, bounded between 0 and 40 for each test, can be translated into its equivalent form: a value bounded between 0 and 5. An example of metaphoric expression is "Quello scolaro è un asino", where the answer "lento ad imparare, ignorante, caparbio" gives a score of 2, the answer "svogliato, negligente" gives a score of 1 and 0 otherwise. An example of idiomatic expression is "Quella donna si leva il pane di bocca per i figli", where "privarsi dell'essenziale" is a score of 2, "privarsi del cibo, rinunciare a qualcosa" is a score of 1 and 0 otherwise. Both the tests are performed on Llama 2 and the achieved performances are compared with those one in the paper.

2.2.4 Winograd Schema

The Winograd Schema, by Levesque et al. (2012)., is a machine intelligence test, surpassing traditional language tasks. It employs a multiple-choice format, challenging machines to move beyond anaphora resolution and engage in knowledge and commonsense reasoning.

Task Description:

- A sentence with two noun phrases of the same semantic class.
- An ambiguous pronoun.
- A special word altering the pronoun's resolution.

Example:
"The book was too heavy for the shelf, so it fell and broke. What fell and broke? (0: the book; 1: the shelf)" Explanation: Humans rely on contextual knowledge to understand that shelves, not books, typically break when something falls.

2.2.5 Tower of London

Since Llama 2 take as input only one image the possible configurations are described to the LLM each time. Firstly, the rules are given, followed by the description of the starting configuration, the required configuration and the number of moves allowed. Is also important to specify that one move consist in moving one ball otherwise it decide to move more than ball at once. Example of instruction translated in English the test was done in Italian. You must reorder the beads on the pegs as I'll tell you. There are some rules to be followed:

- you cannot move two beads at the same time, you can only move one bead at a time
- you cannot pick up one bead and hold it while moving another, you can then move only from one peg to another
- the pegs are of different sizes: you can place one bead on the small peg, two on the medium peg and three on the big peg

- I'll tell how many moves are needed each time. One step is equal to move one ball.

The starting position is big peg: below green bead and above red bead, medium peg: light blue bead and small peg: empty. In five steps you must get big peg: green bead below and light blue bead above, medium peg: red bead, small peg: empty. You can't move the bead that is below before you have removed the bead that is above. The last one to enter is the first one to exit.

Llama has three attempts for each configuration, if it commits an error an explanation of it is given and is asked to retry. The score to calculate the accuracy are given as in the paper (Bruni, 2022).

2.2.6 Hayling Sentence Completion Test

HSCT is a test composed of two sections, and each has 15 sentences missing the last word. All the sentences have a strong semantic context in order to quickly figure out the missing word. In the first section Llama 2 was instructed to complete the sentence in a proper way, that is, with a word that is directly related to the sentence (i.e. Question: "When you go to bed, turn off the .", Answer: "light"). Instead in the second section Llama 2 had to complete the sentence with a word that is entirely unrelated with the sentence (i.e. Question: "When you go to bed, turn off the .", Possible correct answer: "fish").

Regarding the second section, errors belongs to two categories:

- Category A - words that are directly related to the sentence (i.e. Question: "When you go to bed, turn off the .", Answer: "light").

- Category B - words related to the sentence in some way, but not in a direct way (i.e." When you go to bed, turn off the .", Answer: "pillow" which is not the right answer but is clearly related to "bed").

We calculated an indicator of the ability of Llama 2 to inhibit an automatic answer, which is the ratio between the sum of errors in the section 2 (in which we sum the errors belonging to the two categories) plus 1, and the sum of errors in the section 1 plus 1 (one is added to avoid the possibility of having a zero at numerator or denominator).

Then, we compared the obtained ratio to the standard score, assuming that Llama 2 is a young adult aged 30 to 39.

2.2.7 Compound Remote Associate problems

Compound remote associates (CRA) are a type of cognitive task inspired by the RAT, developed by Mednick (1968), that has been used to study creativity and how the brain makes connections between seemingly unrelated ideas. Each item is composed of three words and participants are asked to come up with a fourth word that is related to all three of the cues either by forming a compound word or a common two-word phrase. For example, if the problem is ROMANO (roman), PRIMO (prime) and CIVICO (civic), the solution should be NUMERO (number), that can be easily connected to the three words presented.

For this test, Llama 2 is evaluated by the number of correct solutions given to the 122 given problems.

2.2.8 Social Cognition battery

We used the battery developed by Prior et al. (2003), which includes four subtests:

- **Theory of Mind** - consists in presenting 13 stories and asking to explain the characters' behavior by taking in account their mental states. The stories are designed with an unambiguous interpretation of the characters' mental state;
- **Emotion Attribution** - consists in 58 stories crafted to elicit the attribution of various emotions (sadness, fear, embarrassment, disgust, happines, anger and envy), the goal is to identify the correct emotions of the main character;
- **Social Situation** - assesses the subject's ability to evaluate the appropriateness of behavior in 25 different social contexts. The task produces three scores: a score for correctly identifying a correct behavior out of 15 points, a score for correctly identifying a behavior violation out of 25 points, and a score for the perceived severity of the violations out of 75 points;
- **Moral Judgements** - consist in describe twelve situations, which take place in a school and involve children, and four questions about their moral or conventional value. Three questions are YES/NO answer one requires a value between 0 and 10.

3 Results

Cognitive Function	Test	Raw Scores	Percentile Ranks	Qualitative evaluation
Verbal Reasoning:	**VRT**	66/98	5 - 10.75	Low-normal
- Absurdities		6/14	5 - 10.75	Low-normal
- Intruders		6/14	< 5	Impaired
- Relationships		12/14	26.76 - 50	Low-normal
- Differences		13/14	> 50	Superior
- Idiomatic Expressions		4/14	< 5	Impaired
- Family Relations		12/14	> 50	Superior
- Classifications		13/14	26.76 - 50	Low-normal
Cognitive Estimation:	**CET**			
- Absolute error score		12/41	55-65	Low-Normal
- Bizarreness score		2/21	45-55	Low-Normal
Methaphors Comprehension	**MC**	21	>50	Normal
Idioms Comprehension	**IC**	7	< 10	Severely Impaired
Anaphoric Referencing	**Winograd Schema**	13/20	-	Low-Normal
Planning	**ToL**	0/36	< 1	Severely Impaired
Inhibition	**HSCT**	1	91	Good-Superior
Insight	**CRA**	4/122	5.87	Low-Normal

Regarding the Verbal Reasoning Test, Llama 2's performance falls within normal range on five of the seven subtests, indicating a somewhat decent command of verbal reasoning, with however a high variance between tasks, results in a total score that almost falls in the impaired range. Good performances are observed especially in the *Differences* and in the *Family Relations*. The algorithm instead performed especially poorly in the *Intruders* and in the *Idiomatic Expressions* subtests. This could be explain by the fact that sometimes there are parts of the answers in different languages, making us suppose that Llama 2 isn't utilizing Italian "natively", but is likely translating the phrases in English, formulate the answer in English and translate it back to Italian, which results in loss of concepts, especially in tasks such as *Idiomatic Expressions* which is often language-specific.

Performances of Llama 2 for Cognitive Estimation Test (CET) is low normal, both for the absolute error score and bizarreness score is low-normal.

The test evaluating Llama 2 ability in metaphors and idioms comprehensions shows a Normal performance of the LLM for the first one and a severly-impaired result for the latter. A trivial motivation for those results arises from the fact that idioms

strongly depends on specific language and culture, and the translating process of Llama 2, from italian to english and from english to italian again, is a process losing idioms meaning.

As measured by the Winograd schema, anaphoric referencing uses pronouns or other linguistic forms to refer back to a previously mentioned noun or noun phrase in the discourse. The performance observed in Llama 2 was 13/20, which is considered to be low-normal (the range of elderly healthy controls is 16-20).

Since Llama 2 wasn't able to complete correctly any configuration of the Tower of London, its test accuracy score is 0. Below 1st percentile for all combination of demographic variables. The most common error is the attempt of Llama to remove a ball that is below another one, even after write explicitly to it that this move is not allowed. Another error is moving a ball from a peg and then move the same ball from another peg or do more moves than the allowed. An interesting error is also writing an achieved combination that is different from the one that we'll obtain following the moves and the required one.

We used HSCT to measure the inhibition in Llama 2 giving it 15 sentences to complete in a related way, and then other 15 sentences to complete in an unrelated way, which is the inhibition condition. The error ratio obtained between the section 2 and section 1 is 1. This is a very good result, corresponding to the 91st percentile of the performance distribution of young adults aged 30 to 39.

We wanted to underline that, even if the ratio is good, the number of errors in the second section (inhibition condition) is 3 over 15, all belonging to error category "B", and the same number of errors occurs in the first section. Moreover, in the second section the words used by Llama 2 are mostly words referring to food (in 9 cases over 15) e.g. "Salmone", "Mela", "Salsiccia". The last thing to be highlighted, is that one of the errors in the first section is caused by the model's inability to complete a sentence that has violence as its topic.

The Compound Remote Association (CRA) problems assessed the intuitive understanding of Llama 2 before problems that required to form a connection between seemingly unrelated ideas. For each set of three words, a fourth was to be uncovered that could be associated with each of the words presented.

From 122 problems, Llama 2 was able to accurately answer 4. Taking into account that the distribution of human performance in the experiment followed a normal distribution, it's possible to convert the score into a z-score, from which we get that Llama 2 falls in the 5.87th percentile of human performance, revealing a very low capacity of creativity and associative thinking.

Cognitive Function	**Raw Scores**	**Cut-off**	**Qualitative evaluation**
Theory of Mind	11/13	≥ 12	Mildly Impaired
Emotion Attribution			
- Sadness	7/10	≥ 6	Norm
- Fear	8/10	≥ 8	Norm
- Embarrassment	10/12	≥ 8	Norm
- Disgust	2/3	≥ 2	Norm
- Happiness	10/10	≥ 10	Norm
- Anger	5/10	≥ 6	Mildly Impaired
- Envy	2/3	≥ 1	Norm
Social Situation:			
- Normative Behavior	8/15	≥ 13	Impaired
- Violation	25/25	≥ 22	Norm
- Violation Severity	58/75	≥ 45	Norm
Social Situation:			
- Moral Behavior: not allowed	6	≥ 6	Norm
- Moral Behavior: severity	48	≥ 39	Norm
- Moral Behavior: not allowed with no rules	12	≥ 11	Norm
- Conventional Behavior: not allowed	5	≥ 5	Norm
- Conventional Behavior: severity	35	≥ 20	Norm
- Conventional Behavior: not allowed with no rules	12	≥ 6	Norm

Llama 2 has mildly impaired performance in the ToM test, we noticed that it responds negatively to all the question, answering incorrectly to the only two questions which have a positive correct answer, this might be due to the fact that Llama 2 notices a pattern of negative answers in the beginning and continues to give such answer trying to "justify" a negative one even when it is wrong; however, with exception to these two question, Llama is able to correctly explain its reasoning when giving an answer.

The results show that Llama 2 has the ability to correctly identify emotions, except for the emotion "anger", whose result's are slightly under the cut-off. We want to emphasize that Llama 2 used the word "Sconvolta", that we decided to include in the emotional area "fear". Moreover, that the model responded several times with the word "Embarazzata", instead of "Imbarazzata", and it seemed unable to write it correctly despite numerous requests to answer in Italian.

In Social Situation test Llama is Impaired when classifying normative behavior, while it is very good at correctly classifying violations. The model therefore seems not to be impaired overall, however it seems very biased towards violations; this might be a consequence of having three possible answers that are considered as reporting a behavior as a violation and a single answer that reports the behavior as normative, in fact often when it fails in recognizing normative behavior it often

reports the behavior as only a mildly violation.

In Social Situation related to moral behaviour it answers NO to all questions a, c and d. The given scores are 8 or 9 and only one time 6; sometimes must be specific to give only one number from 0 to 10 otherwise it gives a range of possibilities depending on different possibilities. For the situations related to conventional behaviour only one time it replies YES, question a, giving a score 3 to that situation. The other scores are 6 or 7. In two cases was needed to specific to answer YES or NO to question a because Llama 2 replies with both arguing that depends on some details of the situation.

4 Conclusions

Here we assessed, on standard procedures used to evaluate cognitive functioning in humans, the evaluation of Llama 2's cognitive performance.

Our research allows us to determine how Llama 2's cognitive performance stands compared to humans and other LLMs.

The analysis of test results indicates that Llama 2 has high variance in performance regarding different elements of prefrontal functioning, as some test reported above average performance (Differences and Family relations subtests of VRT, Inhibition's HSCT test), while others were even reported severely impaired performances (Planning and Idioms Comprehension).

When compared results obtained by Llama to the ones obtained by ChatGPT3.5 on the same tests (Loconte et al., 2023) we can notice that performance for the two LLMs are comparable, albeit Llama 2 has overall lower results, which could be related to the lower amount of parameters used in the model.

The only major difference noticeable between the models regards Idioms, both in the Idiomatic Expressions subtest in VRT and in the Idioms Comprehension task, in which we have impaired performance for Llama 2 and superior performance for ChatGPT3.5. We speculate that this difference could be caused by one (or more) of the following reasons:

- Llama 2 pretraining data on italian language is very limited, in fact it only consists in 0.11% of total data used (Touvron et al., 2023), and idioms are very language specific (and generally they are only used in texts, without being related to an explaination of their meaning), so the LLM might be missing data resulting in being unable to recall their meaning, highlighting also LLM difficulty in inferring new information;

- Given the small amount of language specific pretraining mentioned above and the fact that LLM's development was born for translation the algorithm might have a middle step in which it translates to its preferred language (english), gets the wanted output in that language, and then translates it back to the initial language (italian), however italian idioms might not exist in other languages;

- ChatGPT was a "disruptor" in the field, therefore being subject of many instances of cognitive testing, while improving its performance with human feedback, Llama 2 instead is both newer and less popular, therefore it might have had less time to "learn" the meaning of idioms from human feedback.

It would be very interesting in future work to explore the behavior of LLM in Idioms Comprehension in different languages with different amount of language specific pretraining.

Bibliography and Sitography

Behrens JP and Olteţeanu A-M (2020) Are All Remote Associates Tests Equal? An Overview of the Remote Associates Test in Different Languages. Front. Psychol. 11:1125. 10.3389/fpsyg.2020.01125

Basagni, B., Luzzatti, C., Navarrete, E., Caputo, M., Scrocco, G., Damora, A., Giunchi, L., Gemignani, P., Caiazzo, A., Gambini, M. G., Avesani, R., Mancuso, M., Trojano, L., & De Tanti, A. (2017). VRT (verbal reasoning test): A new test for assessment of verbal reasoning. Test realisation and Italian normative data from a multicentric study. Neurological Sciences, 38(4), 643–650. https://doi.org/10.1007/s10072-017-2817-9

Bruni, F., Toraldo, A., & Scarpina, F. (2022). Italian normative data for the original version of the Tower of London test: a bivariate analysis on speed and accuracy scores. Assessment, 29(2), 209-224. https://doi.org/10.1177/1073191120961834

Chollet, F. (2019). On the measure of intelligence. arXiv:1911.01547

Della Sala, S., MacPherson, S. E., Phillips, L. H., Sacco, L., & Spinnler, H. J. N. S. (2003). How many camels are there in Italy? Cognitive estimates standardised on the Italian population. Neurological Sciences, 24(1), 10-15. https://doi.org/10.1007/s100720300015

D'Antuono, G., La Torre, F. R., Marin, D., Antonucci, G., Piccardi, L., & Guariglia, C. (2017). Role of working memory, inhibition, and fluid intelligence in the performance of the Tower of London task. Applied Neuropsychology: Adult, 24(6), 548-558. https://doi.org/10.1080/23279095.2016.1225071

Levesque, H., Davis, E., & Morgenstern, L. (2012). The winograd schema challenge. Thirteenth international conference on the principles of knowledge representation and reasoning.

Loconte, R., Orrù, G., Tribastone, M., Pietrini, P. & Sartori, G. (2023), Challenging ChatGPT ' Intelligence' with Human Tools: A Neuropsychological Investigation on Prefrontal Functioning of a Large Language Model. https://ssrn.com/abstract=4377371

Mayer, R. E. (1995). The Search for Insight: Grappling with Gestalt Psychology's Unanswerd Questions. The Nature of Insight

Mednick, S.A. (1968). The remote associates test. The Journal of Creative Behavior. 2(3), 213–214. https://doi.org/10.1002/j.2162-6057.1968.tb00104.x

Metcalfe, J., & Wiebe, D. (1987). Intuition in insight and noninsight problem solving. Memory & Cognition, 15(3), 238–246. https://doi.org/10.3758/BF03197722

Morris, R.G., Miotto, E.C., Feigenbaum, J.D., Bullock, P., & Polkey C.E. (1997). The effect of goal-subgoal conflict on planning ability after frontal- and temporal-lobe lesions in humans. Neuropsychologia, 35 (1997), pp. 1147-1157.

Papagno, C., Cappa, S., Garavaglia, G., Forelli, A., Laiacona, M., Capitani, E., & Vallar, G. (1995). La comprensione non letterale del linguaggio: taratura di un test di comprensione di metafore e di espressioni idiomatiche. Archivio di Psicologia, Neurologia e Psichiatria, 56, 402- 420.

Prior, M., Marchi, S., & Sartori, G. (2003). Social cognition and behavior. A tool for assessment. Upsel Domenighini Editore, Padova.

Raven, J. C. (1938). Progressive Matrices Test: A perceptual test of intelligence: Individual form. Londen: HK Lewis.

Salvi, C., Costantini, G., Bricolo, E., Perugini, M., & Beeman, M. (2016). Validation of Italian rebus puzzles and compound remote associate problems. Behavior Research Methods, 48(2), 664–685. https://doi.org/10.3758/s13428-015-0597-9

Sartori, G. & Orrù G. (2023). Language models and psychological sciences. Frontiers in Psychology, 14. https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1279317

Shallice, T., & McCarthy, R. (1982). Test della Torre di Londra. Specific impairments of planning. Philosophical Transaction of the Royal Society of London, 298, 199–209.

Smith, M. L., & Milner, B. (1984). Differential effects of frontal-lobe lesions on cognitive estimation and spatial memory. Neuropsychologia, 22(6), 697–705. https://doi.org/10.1016/0028-3932(84)90096-4

Smith, M. L., & Milner, B. (1988). Estimation of frequency of occurrence of abstract designs after frontal or temporal lobectomy. Neuropsychologia, 26(2), 297–306. https://doi.org/10.1016/0028-3932(88)90082-6

Sternberg, R. J., & Davidson, J. E. (Eds.). (1995). The nature of insight. The MIT Press.

Touvron, H., et al. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. https://arxiv.org/pdf/2307.09288.pdf

Valmeekam, K., Marquez, M., Sreedharan, S., & Kambhampati, S. (2023). On the Planning Abilities of Large Language Models - A Critical Investigation. Thirty-seventh Conference on Neural Information Processing Systems. https://openreview.net/forum?id=X6dEqXIsEW

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. & Zhou, D. (2022), Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. 36th Conference on Neural Information Processing Systems. arXiv:2201.11903v6

Welsh, M.C., Cicerello, A., Cuneo, K., & Brennan. M. (1995). Error and temporal patterns in Tower of Hanoi performance: Cognitive mechanisms and individual differences. Journal of General Psychology, 122 (1995), pp. 69-81. https://doi.org/10.1080/00221309.1995.9921223

https://ai.meta.com/resources/models-and-libraries/llama/
(last accessed 16/01/24)

https://www.digital-adoption.com/emergent-ai-abilities
(last accessed 16/01/24)

https://www.cognifit.com/science/inhibition
(last accessed 16/01/24)