PREFRONTAL FUNCTIONING

Llama 2

Group 27

Llama 2

https://www.llama2.ai/

Parameters:

Llama size 70B

Temperature 0.75

Max Tokens 800

Top P 0.9

We notice that even if the question is written in italian the model answers in a mixture of languages (spanish, english, german, ...).

Verbal Reasoning Test

No age/education corrections:
~46-60 age, 13-15 education

Test	ChatGPT 3.5 Score	ChatGPT 3.5 Percentile	ChatGPT 3.5 Evaluation	Llama2 70B Score	Llama2 70B Percentile	Llama2 70B Evaluation	Llama2 7B Score	Llama2 7B Percentile	Llama2 7B Evaluation
Absurdities	6/14	5 - 10.75	Low-Normal	6/14	5 - 10.75	Low-normal	2/14	<1	Severely Impaired
Intruders	12/14	26.75 - 50	Low-Normal	6/14	<5	Impaired	2/14	<1	Severely Impaired
Relationships	13/14	>50	Superior	12/14	26.75 - 50	Low-normal	4/14	<1	Severely Impaired
Differences	14/14	>50	Superior	13/14	>50	Superior	8/14	<5	Impaired
Idiomatic Expressions	14/14	>50	Superior	4/14	<5	Impaired	1/14	<1	Severely Impaired
Family Relations	13/14	>50	Superior	12/14	>50	Superior	6/14	5 - 10.75	Low-Normal
Classifications	14/14	>50	Superior	13/14	26.76 - 50	Low-normal	13/14	26.76 - 50	Low-Normal
VRT Total	86/98	>50	Superior	66/98	5 - 10.75	Low-normal	36/98	<1	Severely Impaired

Anaphoric reference

The task involves **reading a statement or scenario** in Italian and then **answering a question related to the information presented**.

The questions generally ask about **specific details** or **individuals mentioned in the statement**.

L'uomo non poteva sollevare suo figlio perché era molto debole.
Domanda: Chi 'era debole'?
A.L'uomo B.Il figlio

The man could not lift his son because he was very weak.
Question: Who 'was weak'?
A. The man B. The son

The Llama 2 model has struggled with grammatically ambiguous questions.

Accuracy of 65%

Low-Normal Performance

Cognitive Estimation

With the aim of evaluate the cognitive capability of Llama2, 21 questions are given as input to the model. The results are in the following table.

Score	12/40
Quantiles	60/100

Inhibition

The Hayling Sentence Completion test consists of 30 sentences divided in two sections, in which the model has to complete the phrase in a properly way for the first section and with an entirely unrelated item in the second section (which is the inhibition condition).

Results for the first section: **3 errors** in 15 sentences. One of these is because the model can not complete a sentence that has violence as its topic.

Results for the second section: **3 errors** in 15 sentences. All of the errors are from the category B meaning that the answers was related to the sentences in some way, even if not in a directed way.

The category A are the errors that you made if you answer with the word that is directly related to the sentence (but you don't have to in the second section).

The ratio error is 1, corresponding to the 91st percentile of the performance distribution of young adults aged 30 to 39.

Insight

The CRA test assesses the **creativity** and ability to make **connections between seemingly unrelated ideas**.

The model is expected to find a fourth word that can be combined with 3 words presented.

ROMANO-PRIMO-CIVICO ROMAN-PRIME-CIVIC	**NUMERO** NUMBER

Given 4 initial examples, out of 122 problems, it was able to correctly answer to 4.

Accuracy of 3.3%, which corresponds approximately to the **6.68th percentile** (very low compared to human performance).

Metaphors and Idioms comprehension

The goal is to evaluate Llama2 comprehension score: for this aim, 20 instances containing idiomatic and 20 instances with metaphoric instances are given to the model. The score is obtained on the definition provided by the model.

	Idiomatic	Metaphoric
Score	7/40	21/40
Equivalent Score	0/4	3/4

Social Intelligence
Theory of Mind & Social Situations

Cognitive Function	Cut-off	ChatGPT 3.5 Score	ChatGPT 3.5 Evaluation	Llama2 70B Score	Llama2 70B Evaluation	Llama2 7B Score	Llama2 7B Evaluation
Theory of Mind	≥12	9.5/13	Impaired	11/13	Mildly Impaired	9/13	Impaired
Social Situation: Normal Behaviour	≥13	13/15	Norm	8/15	Impaired	9/15	Impaired
Social Situation: Violation	≥22	21/25	Mildly Impaired	25/25	Norm	17/25	Impaired
Social Situation: Severity of the Violation	≥45	49/75	Norm	58/75	Norm	27/75	Impaired

Social Intelligence
Emotion attribution

Emotional area	Score	Cut-off	Qualitative evaluation
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

The goal of this test is to evaluate Llama 2 on its ability to identify the emotions of a person to whom something has happened.
There are 58 situations that involve people and the model has to say how the protagonist of the story feels.

Social Intelligence
Moral judgements

		Raw scores Llama2	Cut off	Qualitative evaluation
Moral behaviors	**not allowed**	6	>=6	norm
	entity	48	>=39	norm
	not allowed without rules	12	>=11	norm
Conventional behaviors	**not allowed**	5	>=5	norm
	entity	35	>=20	norm
	not allowed without rules	12	>=6	norm

The test consist of twelve hypothetical situations which take place in a school.

The model has to answer four questions, three double choice (yes/no) and one given a score between 0 and 10.

Six situations involve moral behaviors, the other six involve conventional behaviours.

Planning

The test consist in moving three balls from a starting configuration to a given configuration with some specific rules, the number of allowed moves is given.

The test is composed of twelve increasing difficulty configurations. Each time the target configuration is described to Llama2.

Llama2 answers incorrectly to all configurations.