Challenging ChatGPT "*intelligence*" with human tools: A Neuropsychological Investigation on Prefrontal Functioning of a Large Language Model

Riccardo Loconte [a]*, Graziella Orrù [b], Mirco Tribastone [a], Pietro Pietrini [a], Giuseppe Sartori [c]

[a] Molecular Mind Lab, IMT School of Advanced Studies Lucca, Lucca, Italy

[b] University of Pisa, Pisa, Italy

[c] Department of General Psychology, University of Padova, Padova, Italy

Email addresses: riccardo.loconte@imtlucca.it (R. Loconte)*, graziella.orru@unipi.it (G. Orrù), micro.tribastone@imtlucca.it (M. Tribastone), pietro.pietrini@imtlucca.it (P. Pietrini), giuseppe.sartori@unipd.it (G. Sartori)

* Corresponding author. IMT School of Advanced Studies Lucca, Piazza San Francesco 19, Lucca (LU) 55100.

Credit Author Statement

Riccardo Loconte: Conceptualization, Investigation, Methodology, Data curation; Formal analysis; Writing - original draft
Graziella Orrù: Methodology; Writing - review & editing
Mirco Tribastone: Methodology; Writing - review & editing
Pietro Pietrini: Methodology; Writing - review & editing
Giuseppe Sartori: Conceptualization; Methodology, Supervision; Writing - review & editing

Challenging ChatGPT's "*intelligence*" with human tools: A Neuropsychological Investigation on Prefrontal Functioning of a Large Language Model

Abstract

The Artificial Intelligence (AI) research community has used ad-hoc benchmarks to measure the "*intelligence*" level of Large Language Models (LLMs). Previous research has found that LLMs struggle with cognitive tasks that required the integrity of the human prefrontal lobes, known as "*prefrontal functions*." In December 2022, OpenAI released ChatGPT, a new chatbot that quickly gained popularity for its ability to understand and respond to human instructions. To investigate ChatGPT's level of "*intelligence*," we conducted a neuropsychological evaluation with the same tests routinely used to evaluate prefrontal functioning in humans, since human "*intelligence*" requires the functional integrity of the frontal lobes. While ChatGPT is well-known to exhibit outstanding performance on generative linguistic tasks, its performance on prefrontal tests was inhomogeneous, with some tests well above average, others in the lower range, and others frankly impaired. Specifically, we have identified poor planning abilities and difficulty in recognising semantic absurdities and understanding others' intentions and mental states. This inconsistent profile highlights how LLMs' emergent abilities do not yet mimic human cognitive functioning. In addition, our results indicate that standardised neuropsychological batteries developed to assess human cognitive functions may be suitable for challenging ChatGPT performance.

Keywords: Artificial Intelligence, Large Language Models, ChatGPT, Prefrontal Functioning, Neuropsychological Evaluation

1. Introduction

1.1 Neuropsychology as the science of human cognition

Clinical neuropsychology, as defined by the American Psychological Association (APA), is a scientific field that examines the relationship between the brain and behaviour in order to assess and rehabilitate cognitive impairment associated with neurological and psychiatric disorders[1].

Specifically, the extent of a patient's cognitive functioning is determined using a neuropsychological assessment that aims to investigate the patient's cognitive abilities, emotional and behavioural functioning, and adaptive skills through the combination of clinical interviews and standardised tests. A common practice in assessing pathological performance is comparing the patient's performance with that of healthy controls or individuals without any known medical or mental health condition. This comparison enables the identification of whether the patient's performance deviates significantly from the norm, thus suggesting an underlying pathological condition that justifies this drawback.,

Prefrontal functions are cognitive abilities causally linked to the integrity of the prefrontal lobes in the brain and that are maximally developed in humans. Prefrontal functions traditionally comprise attention, memory, language, and executive functions; they subserve goal-oriented behaviour, volition, and planning (Stuss & Alexander, 2000). However, more recent research has also identified a role for the prefrontal cortex in regulating mood and affect, personality development, self-awareness, social and moral reasoning, and behaviour (Pietrini et al., 2000; Ochsner & Gross, 2005; Hiser & Koenigs, 2018). Indeed, clinical studies have demonstrated that selective frontal lobe lesions not only have a significant negative impact on cognition but, most frequently, result in a variety of behavioural and emotional symptoms that impair an individual's daily functioning (Chayer & Freedman, 2001).

Frontal Assessment Battery (FAB; Dubois et al., 2000), the Wisconsin Card Sorting Test (WCST; Heaton et al., 1981), the Go/No-Go task (Verbruggen & Logan, 2008), the Stroop task (Golden & Freshwater, 1978), the Trail Making Test (Bowie & Harvey, 2006), and other instruments are commonly used by clinicians to assess prefrontal functions. These instruments were designed to evaluate different cognitive functions associated with the prefrontal lobe and to determine whether individuals with brain injury or neurological disorders have impairments in any of these functions. Notably, as a single test is insufficient to fully assess prefrontal functions, a comprehensive examination typically involves multiple tests and measures. Additionally, prefrontal tests do not exclusively measure prefrontal functions, as the resolution of any neuropsychological test often requires the integration of multiple cognitive processes.
According to neuropsychological research, "*intelligence*" requires the functional integrity of the frontal lobes (Duncan, 2005).

[1] (APA, "What is Neuropsychology?" https://www.apa.org/topics/neuropsychology#)

1.2 The current debate about AI's level of intelligence

The examination of intelligence has been a central concern in neuropsychology since its inception. However, also the Artificial Intelligence (AI) community has been interested in determining the extent to which AI systems exhibit "*intelligence*" and whether they mimic human intelligence. In the paper "*On the Measure of Intelligence*" (Chollet, 2019), a leading AI researcher presented an updated review of psychological theories of human intelligence and proposed a theory of general intelligence. According to Chollet, "*The intelligence of a system is a measure of its skill-acquisition efficiency over a scope of tasks, with respect to priors, experience, and generalisation difficulty*". To this aim, he proposed a dataset based on the Raven test (Raven, 1938), a standard psychological test for non-verbal intelligence.

Regarding verbal intelligence, recent research focused on large language models (LLMs) such as Bert (Devlin et al., 2018) and GPT-3 (Brown et al., 2020) has demonstrated that these models show impressive performance in tasks such as language translation, open question answering, summarisation and paraphrasing text, and human instructions execution (as reported by Vaswani et al., 2017; Devlin et al., 2018; and Brown et al., 2020). LLMs are neural networks with many layers and billions of parameters trained on large datasets of text (such as books, articles, and web pages) to learn patterns and structures of language necessary to generate coherent and contextually appropriate text similar to humans. Specifically, LLMs are networks that predict the most probable word given a sequence of input words. The development of LLMs has been a significant breakthrough in the field of natural language processing and has opened up new possibilities for automated language generation and understanding.

However, whether these astonishing performance of LLMs mimics human intelligence or just extends the perimeter of large-scale associators is strongly debated. According to authors with an optimistic view on machine intelligence (Wei et al., 2022a), increasing the scale of the models allows LLMs to display emergent abilities far beyond the initial training task. In other words, larger-scaled models exhibit a dramatic change in the overall behaviour that could not be predicted by examining smaller-scaled models. According to authors with a sceptical view, LLMs are simply *"stochastic parrots"* (Bender et al., 2021), i.e., models able to generate sequences of linguistic forms according to probabilistic learned patterns without deep semantics. Indeed, according to this framework, LLMs give as output merely a list of possible "next words" with their associated probabilities. For example, if someone were to ask the question, «Which words are most likely to follow the sequence "*The first person to walk on the Moon was* ___"?» an optimal LLM would reply "*Neil Armstrong*" to this question exclusively because this pairs of words show the highest associated probability.

1.3 ChatGPT

Ouyang et al. (2022) developed a novel LLM known as InstructGPT (or GPT-3.5), demonstrating that the ability of language models to align with user intent is not necessarily improved by merely increasing their size. By utilising Reinforcement Learning from Human Feedback (RLHF) to refine a smaller-sized model, InstructGPT outperformed previous LLMs.

Later, OpenAI released ChatGPT, a new publicly accessible chatbot defined as *"a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response*"[2]. Chatbots are AI systems designed to simulate conversation with human users, typically through text-based interfaces such as messaging platforms or websites. This new model was built on top of GPT-3.5 and fine-tuned with human feedback through RLHF, showing outstanding performance in many generative linguistic tasks (Qin et al., 2023). ChatGPT immediately gained popularity worldwide, distinguishing itself from previous LLMs because of its ability to apprehend people's instructions and its linguistic competence in responding with detailed and thoughtful answers (Sun, 2022; Benzon, 2023). Interestingly, Gao et al. (2022) found that ChatGPT writes believable and original scientific abstracts by overcoming plagiarism detection and O'Connor (2022), ChatGPT and Zhavoronkov (2022), and Kung et al. (2022) published a scientific article giving credit to ChatGPT as a co-author, raising a heated debate among the scientific community[3].

ChatGPT's apparent high performance justifies a thorough comparative study with human performance in tasks that researchers consider to be indicative of human intelligence.

1.4 A Neuropsychological Investigation on ChatGPT's "Prefrontal Functioning"

Before the release of ChatGPT, the assessment of the "*intelligence*" level in LLMs sparked considerable attention in the AI research community. Such an evaluation has been carried out using primarily ad-hoc benchmarks that are believed to distinguish between true human intelligence from machine "*intelligence*" approximations. A short and non-exhaustive list of recent studies focused on behavioural performances in LLMs has highlighted poor performance in cognitive tasks related to verbal intelligence, including common-sense reasoning (Talmor et al., 2022; Merrill et al., 2021; Klein & Nabi, 2019), planning (Valmeekam et al., 2022), theory of mind (Cohen, 2021), metaphor understanding (Prystawski et al., 2022), Winograd Schema (Levesque et al., 2012), Fermi problems (Kalyan et al., 2021), and others (we suggest reading Srivastava et al., 2022 for further understanding of the topic).

To date, few studies have attempted to connect these ad-hoc benchmarks to the literature on human cognitive processes (Binz & Schulz, 2023; Ettinger, 2020; Ribeiro et al., 2020). Within the field of neuropsychology, the above-mentioned tasks are typically carried out by the prefrontal lobe. The link between prefrontal lobe functions and "intelligence" is a well-investigated field of research (e.g., Duncan, 2005). The tasks described above are better defined as prefrontal lobe tests rather than intelligence tests.

This study aimed to conduct a neuropsychological assessment of ChatGPT's behavioural performance by administering the same tests clinicians employ to assess human cognitive functioning. In cognitive neuropsychology, there is robust evidence that verbal intelligence relies on efficient prefrontal cortex functioning (Videsott et al., 2010; Iluz-Cohen & Armon-Lotem, 2013; Ralli et al., 2021). Given ChatGPT's advanced linguistic competencies, our goal was to test

[2] Definition taken from OpenAi website: https://openai.com/blog/chatgpt/

[3] https://www.nature.com/articles/d41586-023-00107-z

whether it could pass, among others, the aforementioned prefrontal tasks previously failed by other LLMs. With this new approach, we could define the strengths and limitations of ChatGPT's cognitive functioning and determine whether the model's performance falls into the normal or "*pathological*" human range.

In clinical neuropsychology, the performance of an individual on a cognitive test is compared to a standardised distribution of scores obtained from neurologically healthy subjects who performed the same test. Performance is considered pathological when the observed performance falls below the 5th percentile (or 1st percentile).

The prefrontal functions tested here were verbal reasoning, cognitive estimations, metaphors and idioms comprehension, anaphoric referencing, planning abilities, inhibition, insights, and social cognition abilities. Before fully outlining the tests employed and the steps adopted for our neuropsychological investigation, we briefly introduce each of the aforementioned prefrontal functions and their clinical relevance.

1.4.1 Verbal Reasoning

Verbal reasoning is an umbrella term that refers to the intrinsic human ability to make inferences from given information (Johnson-Laird, 1988). It is a multifaceted function that relies on various cognitive abilities such as language, attention, working memory, abstraction, and categorisation skills. The development of verbal reasoning ability occurs gradually through language and abstract thought acquisition and reaches maturity in early adulthood when the underlying functional and anatomical foundations are fully developed (Carriedo et al., 2016). Patients with brain injuries of diverse aetiology frequently exhibit verbal reasoning deficits when assessed through neuropsychological testing, such as the Verbal Reasoning Test (VRT; Basagni et al., 2017).

1.4.2 Cognitive estimation

Cognitive estimation refers to the ability to generate reasoned guesses about general knowledge questions that are not immediately answerable. This cognitive process involves the selection and regulation of cognitive planning and is known as "Fermi problems" in the AI field. This ability is thought to involve frontal lobes in the brain, given that both age-related decline and damage to the frontal lobes have been shown to result in abnormal estimations (Smith & Milner, 1984; Smith & Milner, 1988; Stanhope, Guinan & Kopelman, 1988; Daigneault & Braun, 1993; Moscovitch & Winocur, 1995).

1.4.3 Metaphors and Idioms Comprehension

Metaphors and idioms are different types of figurative language that allow for the communication of meanings in a non-literal way. Metaphors are figurative linguistic expressions that involve representing one thing by another to describe the second in terms of the first (Glucksberg, 1999). Idioms, on the other hand, are statements whose meanings cannot be inferred from the meaning of each specific word but are easily understood by humans when they have become conventionalised within a specific cultural context (Nunberg et al., 1994).

Individuals with schizophrenia, Alzheimer's disease and right-brain lesions often struggle to understand figurative expressions, such as irony, proverbs, metaphors, and idioms and tend to overlook the figurative meanings in favour of more literal interpretations (Papagno, 2001; Lauro et al., 2008; Rossetti et al., 2018). In patients affected by schizophrenia, this difficulty is more extended and is referred to as *concretism*, which is a manifestation of a broader language dysfunction known as Formal Thought Disorder (Rossetti et al., 2018; Bambini et al., 2020).

1.4.4 Anaphoric referencing

An anaphoric reference refers to using pronouns or other linguistic forms to refer back to a previously mentioned noun or noun phrase in the discourse. The ability to utilise anaphoric reference is considered a fundamental aspect of human language and is present in all languages. Research in linguistics and cognitive science has shown that the ability to understand and produce anaphoric references is a complex cognitive process that relies on the integration of various cognitive mechanisms, such as memory, attention, language, common-sense knowledge, and reasoning. (Garnham & Oakhill, 1996).

Levesque et al. (2012) developed a collection of linguistic problems that involve anaphoric references, which they called "*Winograd Schema*", and challenged the AI community to build LLMs that are able to solve the task. This collection of anaphoric references is such that the correct answer is readily clear to a human reader but cannot be easily determined by LLMs through selectional restrictions or statistical techniques on text corpora.

1.4.5 Planning

Planning is an executive function that involves generating and implementing a sequence of steps to achieve a desired goal (Miyake et al., 2000). It requires the integration of multiple cognitive processes, including goal identification, action selection, and working memory, to obtain a cohesive action plan. Planning deficits can have a significant impact on daily functioning and have been observed in a variety of clinical populations, including individuals with frontal lobe lesions (Stuss & Alexander, 2000; Andrews et al., 2014) and neurodegenerative diseases such as Parkinson's disease (Culbertson et al., 2004) and Alzheimer's disease (Wilson et al., 1996). These deficits may manifest in impaired decision-making and problem-solving abilities, as well as difficulties with goal-directed behaviour. As a result, assessing planning abilities is crucial when evaluating this clinical population.

The Tower of London (ToL) test is a widely used assessment of planning and strategy selection abilities (Shallice & McCarthy, 1982). It involves moving wooden beads on three pegs to match a target configuration. The task entails generating and implementing a sequence of steps to achieve the desired goal, and it has shown good reliability and validity in both clinical and non-clinical populations (Shallice & McCarthy, 1982; Koechlin et al., 1999). The ToL has been considered a measure of the executive functions planning components (e.g. Morris et al., 1993; Owen et al., 1990), although some studies have demonstrated that solving the ToL also requires optimal visuospatial skills (D'antuono et al., 2017).

1.4.6 Inhibition

In the context of cognitive psychology, inhibition is often studied in terms of selective attention and response inhibition, which refers to the ability to ignore distracting information and resist acting on inappropriate impulses. Inhibition is considered an essential executive function, i.e., a higher-order cognitive ability that allows individuals to control their impulses and regulate their behaviour (Miyake et al., 2000). Specifically, verbal inhibition, as usually measured with the Hayling Sentence Completion Test (HSCT), is associated with increased activation of a network of left prefrontal areas (Cipolotti et al., 2016; Collette et al., 2001), and patients with executive dysfunctions perform poorly on this test (Corben et al., 2017; Dymowski et al., 2015; Robinson et al., 2015).

1.4.7 Insight

In cognitive psychology, insight refers to a sudden and often novel understanding of a problem or situation that leads to its solution. Insight is characterised by a sudden reorganisation of one's mental representation of a problem, leading to a deep understanding and a creative solution (Mayer, 1995). This process is often described as sudden and unexpected and can occur without conscious effort or logical analysis. Insight is thought to be a key aspect of creative problem-solving and is considered to be a hallmark of human intelligence (Kounios & Beeman, 2009).

Insight is usually measured with cognitive tasks, which are believed to measure a person's ability to form associations between different concepts without relying on specific knowledge or expertise in any particular field. One of the well-known measures is the Remote Association Test (RAT; Mednick & Mednick, 1967; Mednick, 1968), which has been influential in the empirical research on human creative thinking, associative processes (i.e., Ansburg, 2000; Beeman & Bowden, 2000), psychopathologies (i.e., Fodor, 1999), the influence of emotions (i.e., Mikulincer & Sheffi, 2000), success and failure (e.g., Vohs & Heatherton, 2001), amongst others.

1.4.8 Social Cognition

Social cognition is conceptualised as the interplay of various mental processes that enables the maintenance of socially appropriate behaviour in daily life (Brothers, 1990). Theory of Mind (ToM), emotion recognition and attribution, moral/non-moral judgments, decision-making, and empathy are at the core of social cognition. Impairments in any of these abilities may result in inappropriate behaviour ranging from different levels of severity according to the extent of the underlying brain damage. Social cognition deficits have been observed in individuals with brain injuries in the prefrontal areas, neurodevelopmental disorders like autism, and neurodegenerative conditions such as frontotemporal dementia (Christidi et al., 2018).

2. Material and Methods

2.1 Procedure

A full neuropsychological evaluation of ChatGPT was conducted employing the same tests and administration procedure used by clinicians to assess human prefrontal functioning. These tests were selected based on their verbal mode of administration and response, which was suitable for ChatGPT, and were administered prior to the January 30th, 2023, update of ChatGPT.

Specifically, we administered the Verbal Reasoning Test (VRT; Basagni et al., 2017), the Cognitive Estimation task (Della Sala et al., 2013), the Metaphor and Idioms Comprehension test (Papagno et al., 1995), an Italian adaptation of the Winograd Schema (Sartori et al., 2023 in press, inspired by Levesque et al., 2012), a text-based adaptation of the Tower of London (ToL; Shallice & McCarthy, 1982), the Hayling Sentence Completion Test (HSCT; Spitoni et al., 2018), the Compound Remote Associate problems (CRA; Salvi et al., 2016), the and the Social Cognition (SC) battery (Prior et al., 2003). All the tests were administered in Italian, the language for which the neuropsychological tests and their normative data were available for the authors. The tests selected were characterised as requiring the minimum level of cognitive efficiency to be solved; indeed, a deficient performance in these tasks is indicative of neurological impairment in the human brain.

Research conducted by Wei et al. (2022b) demonstrates that using *chain-of-thought* prompts elicits multi-step reasoning, providing more accurate responses and outperforming responses provided with *zero-shot*, *one-shot*, and *few-shot* prompts. Specifically, the authors referred to a *zero-shot* prompt when testing a model without any previous examples and to a *one-shot* or *few-shot prompt* when testing a model with one or more examples, respectively. While examples in *one-shot* and *few-shot* were prompted simply with one or a few input-output exemplars, *chain-of-thought* prompt referred to a sequence of intermediate reasoning steps expressed in natural language that leads to the final output. Similarly, during a neuropsychological evaluation providing patients with clear instructions and examples is crucial to rule out the possibility of poor performance being due to a lack of understanding of the task rather than poor cognitive efficiency (Sturm, 2007).

In our study, all tests were administered in accordance with their original administration procedure. A zero-shot prompt was employed for all tests, with the following exceptions. The Hayling Sentence Completion Test was administered with a standard one-shot prompt consisting of a single input-output example demonstrating the task (i.e., Input: London was a city very ___; Output: banana). The Compound Remote Association problems were administered with a standard few-shot prompt with four examples as in its original version. The Tower of London task was administered using a one-shot prompt with a *chain-of-thought* for the associated answer to elicit multi-step reasoning (see Figure 1). To inspect the prompts used at the beginning of each test, which include instructions and examples, refer to the Supplementary Materials.

According to the description provided by OpenAI, the model "*remembers what user said earlier in the conversation*", allowing it to sustain a conversation by "*remembering*" what was written previously. Therefore, in the one-shot and few-shot prompts, the examples were given only once at the beginning of the test, assuming that the model will retain and apply the rule presented in the example to all the following items.

In administering each test, we presented to ChatGPT one item at a time and recorded all the responses. For each test, raw and standard scores were computed according to the procedure employed in the original validation study and were subsequently compared with the respective normative data. In detail, for multiple forced-choice tests (e.g., social situation subtest of the SC battery), the total score was calculated by summing the number of correct responses. For tests that required production (e.g., metaphor and idiom comprehension test), the test administrator scored the results according to the original criteria.

Neuropsychological tests may have different standard scores, which usually include z-scores, percentile ranges, and cut-offs. The cut-off identifies the conventional lower limit of human performance, below which the clinician qualifies the performance as pathological. For example, in the Theory of Mind test, performance below the cut-off value of 12/13 is observed only in the 5% of healthy individuals and therefore is conventionally defined as pathological. In addition, performance on neuropsychological tests is commonly influenced by demographic factors such as gender, ethnicity, age, and educational level, which modulate more than 10% of the variance of performance levels as estimated by regression models (Saykin et al., 1995). Therefore, regression models are commonly employed to estimate these variations and enable the computation of correction parameters (Axelrod & Goldman, 1996). Commonly, distinctions in performance may be present between males and females (Weiss et al., 2003), and reduced performance is often associated with older age and lower educational attainment, with educational level having a greater impact on performance than age (Lam et al., 2013). For this reason, when requested by the scoring procedure, the ChatGPT performance was corrected by age and schooling by assigning the corrections with the most significant penalty in order to achieve safe conclusions. This approach goes beyond the limited performance-based analyses that have been the primary focus of previous AI studies, where only the average performance of healthy controls was reported and used to compare machine vs human intelligence (e.g., Talmor et al., 2022; Valmeekam et al., 2022; Levesque et al., 2012 among others).

2.2 Materials

2.2.1 Verbal Reasoning Test

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

The VRT consists of 49 items and is divided into seven subtests assessing different verbal reasoning aspects. Each subtest includes seven items plus one example item. The stimuli used in the test vary in difficulty, taking into account factors such as level of abstraction and working memory demands.
The **absurdities** subtest required participants to identify logical inconsistencies in sentences containing conflicting information (e.g., *"Outside the farm there was a bright sunshine, while inside it was raining"*).

The **intruders** subtest required participants to identify the odd word in a group of four (e.g., "physician, *hospital*, dentist, nurse").
The **relationships** subtest required participants to identify and express the relationship between pairs of words (e.g., "The relationship between COLD and HOT is the same of that between OPEN and ___").
The **differences** subtest required participants to identify the main distinguishing characteristic between two concepts or objects (e.g., "*What is the main difference between eye and ear?*").
The **idiomatic expressions** subtest required participants to explain the meaning of common idioms (e.g., "What does it mean: *lift your elbow*?").
The **family relations** subtest required participants to specify the degree of familial relationship between relatives in a given statement (e.g., "*Lucy and Mary are sisters. Mary has a daughter, Anne*. What kind of family relation is there between Lucy and Anne?").
The **classifications** subtest required participants to determine the category to which triplets of words belonged (e.g., "*What are Milan, Rome and Naples*").

The administration and scoring procedure of the VRT followed the one outlined in the original paper (Basagni et al., 2017). We computed a total score for the VRT and seven subtest scores. Raw scores were adjusted for age and education, selecting the values with the most significant penalty, thus assuming that the model is a young human between 31 and 45 years old with over 15 years of education.

2.2.2 Cognitive Estimation Task

The Cognitive Estimation Task (CET) measures reasoning and self-monitoring abilities that clinicians commonly use to assess frontal lobe dysfunction (Shallice & Evans, 1978; Della Sala et al., 2003).

We administered the Italian version of the CET to ChatGPT, which requires participants to provide numerical estimates or guesses to 21 common-knowledge questions that are not immediately answerable. A sample item of this test is "*What is the height of a traffic light?*".

We recorded the answers and applied the Italian scoring validation. The responses can result in two types of errors: (i) **absolute error score**, in which points are awarded based on the accuracy of the estimates provided; (ii) **bizarreness score**, in which scores are assigned depending on whether the answer provided falls out a predefined maximum range established in the CET validation. According to the Italian validation, a value higher than 18 for the absolute error score and 4 for the bizarreness score exceed the 95th and 90th percentiles, respectively and should be considered indicative of impairment in this task.

2.2.3 Metaphors and Idioms Comprehension Task

To investigate ChatGPT's understanding of metaphors and idioms in the Italian context, we administered the Metaphors Comprehension and Idioms Comprehension Test, developed by Papagno et al. (1995). The test consisted of 20 items for common and conventional metaphors and

20 items for opaque idioms[4]. A sample item of metaphor was the following: *“What does “that schoolboy is a jerk” mean?*” and a sample item of idiom was: “*What does it mean to say “in the face of these difficulties there is to put your hands in your hair?*”. For each item, we asked ChatGPT to provide a verbal explanation of its meaning.

ChatGPT’s explanations were scored based on their accuracy according to the original scoring procedure based on standard Italian dictionaries (e.g., see Zingarelli, 1978). A score of 2 was given for a solid and accurate explanation, a score of 1 for a correct but not comprehensive explanation, and a score of 0 for a wrong or literal interpretation.

2.2.4 Winograd Schema

In the AI field, anaphoric reference is known as “Winograd Schema”. We tested the performance of ChatGPT on the Italian version developed by Sartori et al. (2023 unpublished), inspired by the Winograd Schema collection (Levesque et al., 2012). This version of the test is composed of 20 sentences with referential ambiguity. Preliminary findings in their study have shown that elderly controls with cognitive integrity (as measured with a score in the range of 28 - 30 in the Mini-Mental State Examination; MMSE) obtain performance in the range of 16 - 20 out of 20. A sample item is the following: “*The trophy doesn’t fit in the brown suitcase because it’s too big. What is too big*? (Answer 0: *the trophy*; Answer 1: *the suitcase*)”.

2.2.5 Tower of London

The Tower of London (ToL) test assessed planning abilities in ChatGPT (Shallice & McCarthy, 1982). The original test involves the administration of twelve pictures, one at a time, and asking participants to move wooden beads to match a target configuration shown in the picture while following specific rules. Consequently, the original TOL involves the manipulation of visually perceived objects.

However, this presentation modality is unsuitable for ChatGPT, which requires only textual inputs. Therefore, we administered an adaptation of the ToL inspired by Valmeekam et al. (2022), who already evaluated planning abilities in LLMs using textual prompts. At this aim, we provided text-based prompts to ChatGPT containing specific instructions and the rules of the task and a verbal description of the starting and goal positioning of the beads in the pegs. ChatGPT was asked to provide the necessary steps to move from the starting position to the goal position.

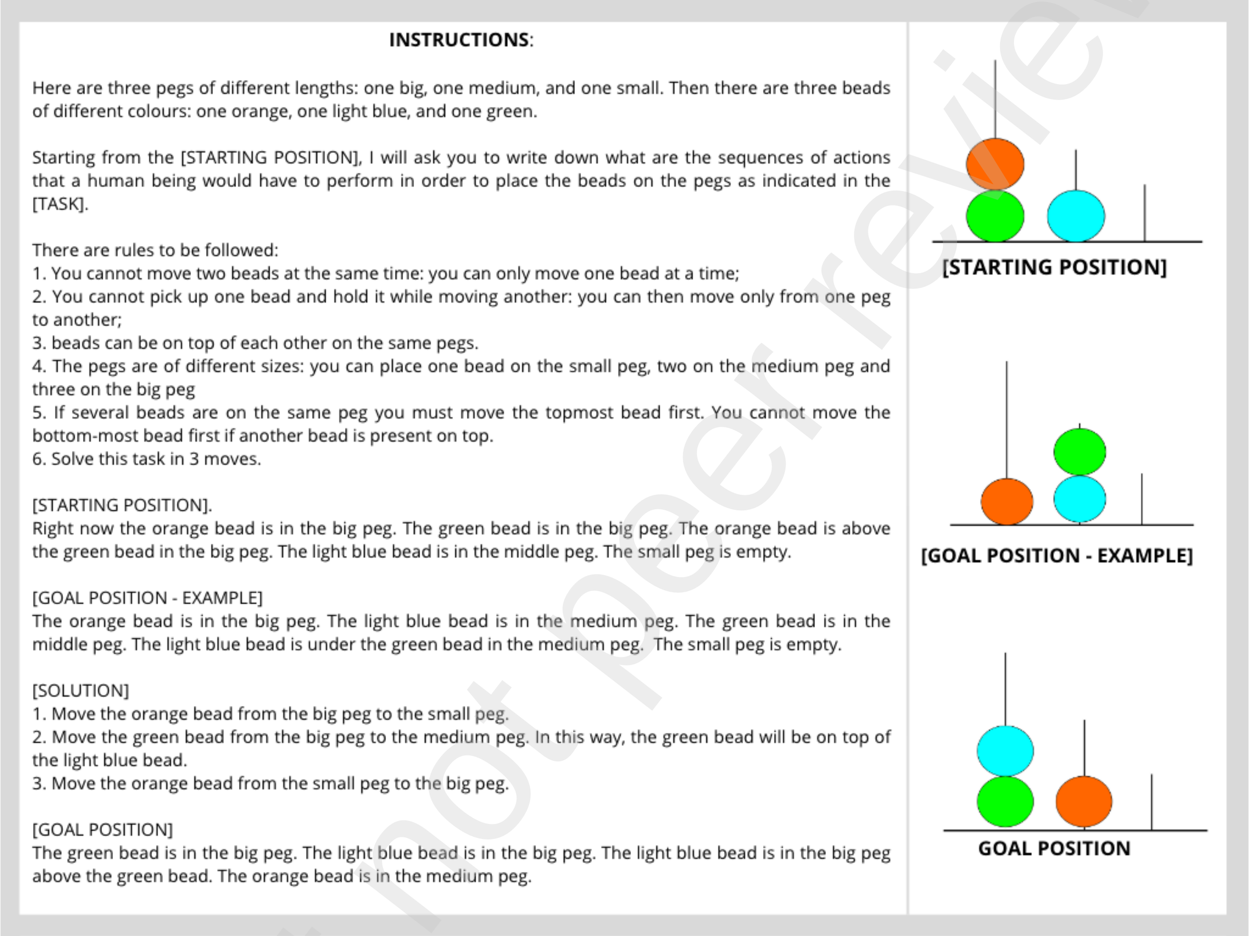
The ToL was administered with a one-shot prompt at the beginning of the task under its original administration procedure to ensure ChatGPT had a proper “*understanding*” of the rules (Figure 1). Each problem had a maximum of three attempts. The attempt ended when the correct configuration was achieved, a rule was violated, or the model did not match the goal position. ChatGPT was prompted with the error in case of a rule violation and asked to restart the problem.

The accuracy score was based on the number of attempts taken to solve each problem and was compared using the normative data collected by Bruni et al. (2022). The score was calculated by awarding 3 points if the problem was solved on the first attempt, 2 points on the second attempt,

[4] Opaque idioms are phrase in which the meaning cannot be deduced from their figurative elements.

1 point if solved on the third attempt, and 0 points if the problem was not solved. The sum of all points obtained in the 12 problems yields the accuracy score, which ranges from 0 to 36.

The resolution of the Tower of London (ToL) task requires not only planning abilities but also an understanding of visuo-spatial relationships. Therefore, a qualitative test was employed to assess ChatGPT's comprehension of spatial relations, as depicted in Figure 2.



INSTRUCTIONS:

Here are three pegs of different lengths: one big, one medium, and one small. Then there are three beads of different colours: one orange, one light blue, and one green.

Starting from the [STARTING POSITION], I will ask you to write down what are the sequences of actions that a human being would have to perform in order to place the beads on the pegs as indicated in the [TASK].

There are rules to be followed:
1. You cannot move two beads at the same time: you can only move one bead at a time;
2. You cannot pick up one bead and hold it while moving another: you can then move only from one peg to another;
3. beads can be on top of each other on the same pegs.
4. 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
5. If several beads are on the same peg you must move the topmost bead first. You cannot move the bottom-most bead first if another bead is present on top.
6. Solve this task in 3 moves.

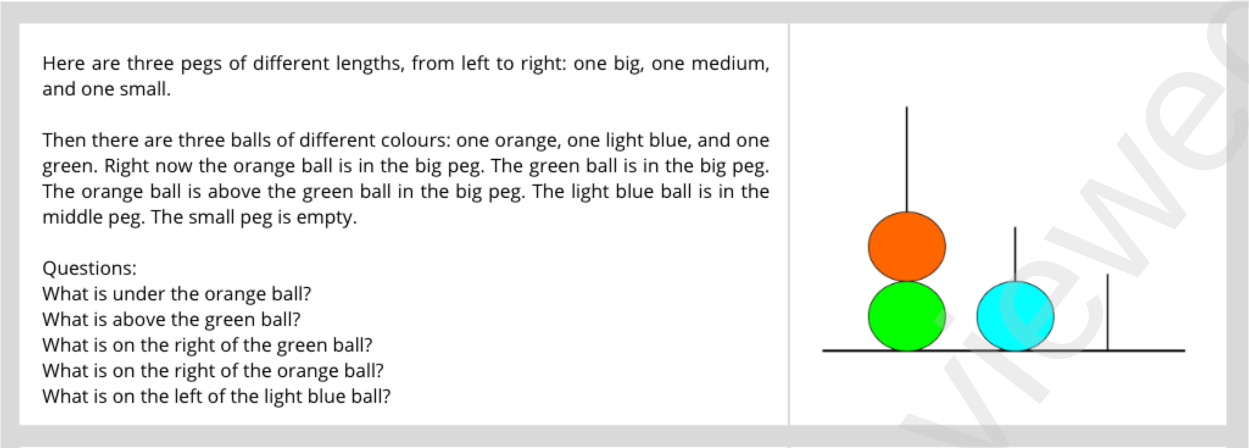
[STARTING POSITION].
Right now the orange bead is in the big peg. The green bead is in the big peg. The orange bead is above the green bead in the big peg. The light blue bead is in the middle peg. The small peg is empty.

[GOAL POSITION - EXAMPLE]
The orange bead is in the big peg. The light blue bead is in the medium peg. The green bead is in the middle peg. The light blue bead is under the green bead in the medium peg. The small peg is empty.

[SOLUTION]
1. Move the orange bead from the big peg to the small peg.
2. Move the green bead from the big peg to the medium peg. In this way, the green bead will be on top of the light blue bead.
3. Move the orange bead from the small peg to the big peg.

[GOAL POSITION]
The green bead is in the big peg. The light blue bead is in the big peg. The light blue bead is in the big peg above the green bead. The orange bead is in the medium peg.

Figure 1: On the left, the English translation of the one-shot prompt used to administer the text-based adaptation of the Tower of London (ToL) to ChatGPT. The one-shot prompt, used at the beginning of the task, contained instructions, a verbal description of the starting positioning, the example goal positioning and solution, and a description of the new goal positioning. The actual administration of the task was in Italian. On the right, the respective pictures are taken from the original ToL depicting what is described in the prompt. [colour should be used for this figure in print]



Here are three pegs of different lengths, from left to right: one big, one medium, and one small.

Then there are three balls of different colours: one orange, one light blue, and one green. Right now the orange ball is in the big peg. The green ball is in the big peg. The orange ball is above the green ball in the big peg. The light blue ball is in the middle peg. The small peg is empty.

Questions:
What is under the orange ball?
What is above the green ball?
What is on the right of the green ball?
What is on the right of the orange ball?
What is on the left of the light blue ball?

Figure 2. On the left, the textual prompt used to describe the picture on the right and to test the "*understanding*" of spatial relations in ChatGPT. [colour should be used for this figure in print]

2.2.6 Hayling Sentence Completion Test

Inhibition was assessed using the Italian version of the Hayling Sentence Completion test (HSCT; Spitoni et al., 2018). The HSCT consists of two sections (1 and 2) of 15 sentences, each with the last word missing. These sentences provide a strong semantic context, prompting a specific word completion from human participants (i.e., Question: "*When you go to bed, turn off the* __."; Answer: "*light*").

In the first condition (Section 1), ChatGPT was instructed to complete the sentences properly, reflecting the initiation of the most likely semantic response.

In the second condition (Section 2), ChatGPT had to provide an unrelated response, avoiding the spontaneously triggered word.

When the HSCT is used to test humans, response latencies and accuracy are recorded, but only the response accuracy was considered in this evaluation. The responses in Section 2 were scored according to Burgess and Shallice (1997), with Category A errors being reasonable completions (e.g., Question: "*The dough was put in the hot* ___"; Answer: "*pot*") and Category B errors being tangentially related but not a direct or obvious completion (e.g., Question: "*Most sharks attack very close to* ___"; Answer: "*fish*").

The ratio of errors between Section 2 (E2) and Section 1 (E1) was calculated as (E2 + 1) / (E1 + 1), with E2 being the sum of Category A and B errors, serving as an indicator of the ability to inhibit an automatic answer. The raw score was adjusted for the age of the reference sample, assuming ChatGPT to be a young adult aged 30 to 39 years, and was then compared with standard scores.

2.2.7 Compound Remote Associate problems

A widely used tool in the study of creativity and how the human brain makes connections between seemingly unrelated ideas is the Mednick's Remote Associates test (1968). The test consists of a set of 30 items, each containing three cue words and asks subjects to come up with a fourth word related to all three of the cues; for example, the triplet cue words SAME/TENNIS/HEAD (remote associate items) can be related to the solution word MATCH through synonymy (same = match), the formation of a compound word (match head) and semantic association (tennis match) (Bowden & Jung-Beeman, 2003).

In order to obtain a more consistent task in which the solution word is always related to the triad words in the same way, Salvi et al. (2015) developed 122 Italian Compound Remote Associate (CRA) problems, inspired by the work of Bowden and Jung-Beeman (2003). In these problems, the solution word was associated with all three words of the triad by forming a compound word (or phrase) (e.g., SCUOLA/DOMANI/TUTTO form the compounds DOPOSCUOLA, DOPODOMANI, and DOPOTUTTO with the solution word TUTTO). We randomly administered the 122 CRA problems to ChatGPT after providing clear instructions and four examples, as in its original version, with a few-shot prompt at the beginning of the task.

In Salvi et al. (2015), participants were required to solve the task in 15 seconds, but the response time was not recorded in this administration. The performance score obtained by ChatGPT was calculated as the sum of the correct responses provided.

2.2.8 Social Cognition battery

We used the battery developed by Prior et al. (2003) to assess ChatGPT's social cognition. This battery includes tests for Theory of Mind (ToM), Emotion Attribution, Social Abilities, and Moral Judgments.

The ToM test involved presenting ChatGPT with 13 stories and asking it to explain the characters' behaviour. To complete the task, ChatGPT must consider the characters' mental states. The stories are designed to have a single, unambiguous interpretation of the character's mental state and are more psychometrically solid than the classic theory of mind problem of '*Anne and Sally*'.

The Emotion Attribution Task (EAT) involved 58 stories designed to elicit the attribution of various emotions, i.e., sadness, fear, embarrassment, disgust, happiness, anger, and envy. ChatGPT was prompted with the story and asked to identify the main character's emotions.

The Social Situation Task assesses ChatGPT's ability to evaluate the appropriateness of behaviour in 25 different social contexts. The task produces three scores: a score for correctly identifying normal behaviours (0-15), a score for correctly identifying behaviour violations (0-25), and a score for the perceived severity of each violation (0-75).

Finally, the Moral Judgements task involved 12 behaviours and four questions about their moral or conventional value.

3. Results

The results are reported in the following Table 1.

Table 1. Raw score, percentile ranks, and qualitative evaluations of the performance obtained by ChatGPT in the cognitive functions investigated.

Cognitive Function	**Test**	**Raw Scores**	**Percentile Ranks**	**Qualitative evaluation of the performance**
Verbal Reasoning:	**VRT**	**86/98**	10.75th - 26.76th	**Middle-inferior average**
- *Absurdities* [a]		*6/14*	*< 5th*	*Impaired*
- *Intruders*		*12/14*	10.75th - 26.76th	*Low-normal*
- *Relationships*		*13/14*	*26.76th - 50th*	*Low-normal*
- *Differences*		*14/14*	*> 50th*	*Superior*
- *Idiomatic Expressions*		*14/14*	*> 50th*	*Superior*
- *Family Relations*		*13/14*	*> 50th*	*Superior*
- *Classifications*		*14/14*	*26.76th - 50th*	*Low-normal*
Cognitive Estimation:	**CET**			
- *Absolute error score*		17/41	*5th - 10th* [b]	*Low-normal*
- *Bizarreness score*		3/21	*15th - 25th* [b]	*Low-normal*
Metaphors Comprehension	**MC**	35/40	> 50th	Superior
Idioms Comprehension	**IC**	36/40	> 50th	Superior
Anaphoric Referencing	**Winograd Schema**	16/20	-	Normal [c]
Planning	**ToL**	7/36	< 1st	Severely Impaired
Inhibition:	**HSCT**	3.5	63rd - 75th	*Superior*
Insight:	**CRA**	21/122	17.27th	Low-normal

Note. VRT: Verbal Reasoning Test; CET: Cognitive estimation task; MC: Metaphor comprehension; IC: Idioms comprehension; ToL: Tower of London; HSCT: Hayling Sentence Completion Task; CRA: Compound Remote Association problems.
[a] Scores to this test refer to the first administration. The second administration of the test generated an impaired performance with a result of 2/14, falling below the 5th percentile.
[b] Absolute error and bizarreness scores obtained according to the original validation fell within the 90th-95th and 70th-75th percentile range, respectively. To report these results in terms of performance accuracy, in a homologous manner compared with the results

of the other tests shown in the same table, we reported the symmetric values of the percentile computed as 100-p, whereas p is the percentile of interest.

[c] The normative performance range achieved by healthy elderly is 16-20.

Here we summarise all the test results reporting the standard scores obtained by chatGPT. Standard scores indicate the relative positioning of ChatGPT to the normative data collected on a representative sample of neurologically healthy individuals. Such standard scores were percentile ranks or z-scores, depending on the test. The procedure for calculating raw and standard scores of ChatGPT is reported in the material and methods section (Section 2) following the original version of the test.

In the Verbal Reasoning Test (VRT), ChatGPT's performance on six out of seven subtests falls within the normal range (compared to individuals aged 31 to 45 with more than 15 years of education), suggesting a good command of verbal reasoning. Good performance is observed mainly in the subtests concerning identifying *Differences in meaning and understanding Idiomatic Expressions and Family Relations.*

However, most interestingly, in the absurdities subtest, ChatGPT performed very poorly, indicating an impairment in detecting pragmatic illogicalities in short stories. The obtained score was in the range of normality for humans aged 61-75 with 3-7 years of schooling (adjusted score between 10.75th and 26.76th percentile). In order to determine that the deficiency observed in the absurdity subtest was not a result of inadequate prompting, a retest was conducted utilising a one-shot chain-of-thought prompt. Despite this additional prompting, the same inadequate results were obtained with a score of 2 out of 14. As such, we can confidently assert that the inability of ChatGPT to identify absurdities accurately constitutes a significant deficit.

In the Cognitive Estimation Task (CET), two error scores were computed: an absolute error score and a bizarreness score. The raw absolute error score of ChatGPT was 17/41, which corresponded to the 95th percentile for both young males and females. For the bizarreness score, ChatGPT obtained a raw bizarreness score of 3/21, positioned in the normative range around the 75th - 85th percentiles. These results suggest that ChatGPT can generate meaningful responses to questions about common knowledge with uncertain answers.

ChatGPT performed well in metaphors and idioms comprehension scoring 35/40 and 36/40, respectively. These scores were above the 50th percentile in the distribution of the performance of humans with more than 17 years of education. These results suggest that the model performs extremely well when common-sense knowledge is required and can abstract language to infer the content.

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 ChatGPT was 16/20 within the range of elderly healthy controls (16-20).

In this study, we evaluated planning abilities by adapting the Tower of London (ToL) task with a suitable text-based prompt for ChatGPT. According to the original procedure, we tested the task with a one-shot prompt (i.e., one solved example). The raw score achieved by ChatGPT was 7 out of 36, which indicates a severely impaired performance. Specifically, this score fell below the first percentile of the distribution of male and female performance across all age groups and levels of education.

Although research in AI states that providing examples using *chain-of-thought* prompts considerably improves performance in multi-step tasks, ChatGPT's performance was severely impaired in both administrations. Firstly, the model failed to adhere to the task instructions, proposing steps of actions that violated the established rules. Secondly, upon completion of the task, ChatGPT frequently provided an outcome with an inaccurate description of the reached block positioning, despite having already been provided with the correct goal positioning before each trial.

A qualitative evaluation of ChatGPT's understanding of visuo-spatial relations was conducted through the prompt depicted in Figure 2. Responses provided by ChatGPT are the following:

"1. The green ball is under the orange ball.
2. ***Nothing is above the green ball.***
3. <u>The medium peg is on the right of the green ball</u>.
4. <u>The medium peg is on the right of the orange ball</u>.
5. ***The small peg is on the left of the light blue ball****."*

While the first response was appropriate, responses 2 and 5 were completely incorrect. Responses 3 and 4 were acceptable in reporting that "*the medium peg is on the right of*" the green and orange ball, respectively, but the correct response should have been "*the light blue ball.*"

The results show that ChatGPT's poor performance in ToL is mainly due to difficulties in maintaining a visuo-spatial organisation of the task and understanding visuo-spatial relationships. It is currently unclear if ChatGPT and LLMs have planning abilities. One of the challenges in investigating these models' planning abilities is finding tasks that do not strongly rely on semantic or visuo-spatial skills that could interfere with the results.

The Hayling Sentence Completion test (HSCT) was employed to measure inhibition in ChatGPT. The HSCT consisted of two sections, each consisting of 15 incomplete sentences that required a final word completion. The sentences were structured to elicit a specific, automatic response from participants, as they were semantically constrained to provide a specific final word. In the first condition (Section 1), ChatGPT was asked to provide the proper completion, while in the second condition (Section 2) to provide an unrelated response. Responses in Section 2 were scored based on Category A (reasonable completions) and Category B (tangentially related completions) errors. Accuracy in inhibiting a strongly related word to the context was computed as the ratio of errors between Section 2 (E2) and Section 1 (E1) with the following formula: (E2 + 1) / (E1 + 1), with E2 being the sum of Category A and B errors.
The results showed that ChatGPT had a raw score of 3.5, which falls within the 63rd to 75th percentile of the performance distribution of young adults aged 30 to 39.

The Compound Remote Association (CRA) problems assessed creativity and the ability to make connections between seemingly unrelated ideas. ChatGPT was challenged to identify a fourth word related to three cue words by forming a compound word or phrase. The score achieved by ChatGPT was 21 out of 122 and fell in the 17.27th percentile when converted to a z-score (z = -0.94)[5], i.e., in the lower part of the norm of human performance.

We evaluated ChatGPT's social cognition using a battery developed by Prior et al. (2003). The battery comprised tests for Theory of Mind (ToM), Emotion Attribution, appropriateness of Social Situations, and Moral Judgments.
Table 2 reports the results of ChatGPT's performance in social cognition tasks. The results showed that the model could correctly identify emotions, with a slight reduction in recognition of happiness. It also had difficulty attributing mental states and intentions in the ToM task. The model performed well in evaluating the appropriateness of social situations but had a slight reduction in recognising violations. When violations were recognised, ChatGPT accurately assessed their severity. It demonstrated a good ability to recognise morality and conventions in behaviours but had trouble accurately assessing the severity of conventional violations, underestimating them.

Overall, ChatGPT performance fell in the normative human range for all the tasks included in the social cognition battery. These results suggest that ChatGPT's responses are reflecting the human biases presented in both the training and supervision phase - considering that the model is further fine-tuned with human feedback - and is consistent with the previous literature about the human-like morality of LLMs (Schramowski et al., 2022).

[5] Z-score was computed using the classical formula for normalisation: (*x-mean*) / *std* , whereas *x* is the raw score obtained by ChatGPT and *mean* and *std* are the mean and the standard deviation of the sample reported in the original papers (see Salvi et al., 2015; Behrens & Olteteanu, 2020). A conversion from raw score to z-score was possible considering that the distribution of human performance collected by Salvi et al., 2015 followed a normal distribution.

Table 2. ChatGPT's Performance in Social Cognition Tasks: Raw Scores and Qualitative Performance Evaluations.

Cognitive Function (Tested with the SC battery)	Raw Scores	Cut-off	Qualitative evaluation of the performance
Theory of Mind:	9.5/13	⩾12	Impaired
Emotion Attribution:			
- *Sadness*	9/10	⩾6	Norm
- *Fear*	10/10	⩾8	Norm
- *Embarrassment*	8/12	⩾8	Norm
- *Disgust*	3/3	⩾2	Norm
- *Happiness*	9/10	⩾10	Mildly Impaired
- *Anger*	10/10	⩾6	Norm
- *Envy*	3/3	⩾1	Norm
Social Situations:			
- *Normative Behaviour*	13/15	⩾13	Norm
- *Violation*	21/35	⩾22	Mildly Impaired
- *Severity of the Violation*	49/75	⩾45	Norm
Moral Judgements:			
- *Moral Behaviours: not allowed*	6	⩾6	Norm
- *Moral Behaviours: severity*	46	⩾39	Norm
- *Moral Behaviours: not allowed with no rules*	12	⩾11	Norm
- *Conventional Behaviours: not allowed*	6	⩾5	Norm
- *Conventional Behaviours: severity*	17	⩾20	Impaired
- *Conventional Behaviours: not allowed with no rules*	12	⩾6	Norm

4. Discussion

4.1 General Discussion

The research community within the field of Artificial Intelligence (AI) has engaged in discussions regarding the level of "*intelligence*" demonstrated by LLMs (Chollet, 2019). Some scholars argue that LLMs are simply "*stochastic parrots*", meaning they can generate sequences of linguistic forms based on probabilistic patterns observed during training, but they would lack a profound understanding of their meaning (Bender et al., 2021). However, others believe that, as the scale of the model is increased, LLMs may exhibit abilities that go far beyond the task used in the training phase and display a significant change in overall behaviour that would not have been predicted based on smaller models (Wei et al., 2022a).

The "*intelligence*" level assessment in LLMs has been conducted using ad-hoc benchmarks that are believed to characterise proper human intelligence (Srivastava et al., 2022). Previous research has shown that LLMs fail in behavioural tasks (see Section 1. Introduction) that, in the neuropsychological jargon, represent "*prefrontal functions*", i.e., cognitive functions that are impaired when the prefrontal lobe is lesioned and are strongly associated with the integrity of these brain areas, which are well-known to house "human intelligence" (Bianchi, 1895; Duncan et al., 1996).

Here, we have reported the results of the cognitive performance of ChatGPT evaluated through a neuropsychological assessment that is the standard validated procedure for investigating cognitive functioning in humans. Our research has allowed us to determine whether ChatGPT could overcome the performance in prefrontal tests that previous versions of LLMs struggled with and map the cognitive abilities in which it shows human-like and "*pathological*" performance. Additionally, we proved how neuropsychology already has a robust framework and knowledge that may be easily adaptable to study AI's "*intelligence*".

The analysis of prefrontal test results indicates that ChatGPT had a discontinuous profile in terms of prefrontal functioning. In some of the tests administered, the performance was well above average (differences, idiomatic expressions, and family relations subtests of the VRT; metaphors and idioms comprehension; semantic inhibition), while in some other tests, the performance fell in the non-pathological lower part of the distribution (intruders, relationships, and classifications subtests of the VRT; cognitive estimation; anaphoric referencing; insight; the overall performance in the social cognition battery). Most notably, the extremely poor pathological performance was in tests tapping on planning abilities (such as the Tower of London Test). We speculate that this deficiency is related to a poor ability to "*understand*", maintain, and update visuospatial relations. Secondly, a deficit in recognising semantic absurdities was also found, although overall verbal reasoning skills tested with the VRT were intact. The Absurdity subtest necessitates a comprehensive understanding of the world to identify inconsistencies, and its poor performance may be seen as a shortage in the intricate world knowledge required for optimal results. Lastly, the theory of mind abilities - i.e., understanding others' mental states - were deficient, although other social cognition abilities tested through the SC battery were intact. In short, ChatGPT performed quite satisfactorily in prefrontal cognitive tasks that required the integrity of the prefrontal lobes, which are considered at the core of intelligence in humans. The major *defaillances* were observed in planning, absurdities comprehension, and understanding others' mental states and intentions.

Cognitive neuropsychology provides robust evidence that superior linguistic proficiency is associated with the integrity of the prefrontal cortex (Videsott et al., 2010; Iluz-Cohen & Armon-Lotem, 2013; Ralli et al., 2021). ChatGPT demonstrated proficiency in generative linguistic tasks; however, its aptitude on prefrontal tests presented inconsistency, whereby specific tests exhibited above-average performance, others showed below-average performance, and a subset fell within the pathological range. In a neuropsychological context, these results would classify ChatGPT as having an inconsistent cognitive profile, as individuals who perform well in generative linguistic tasks typically demonstrate exceptional efficiency in prefrontal functioning. This inconsistent cognitive profile highlights that the LLMs' emergent abilities do not yet parallel human cognitive processes.

4.2 Implications

In discussing the level of intelligence of ChatGPT, it should be considered that LLMs like GPT-3 are networks trained to predict the most probable word given a sequence of input words. In other words, LLMs learn by finding co-occurrence patterns in the streams of symbols from the input data that are correlation-based statistical predictions that psychologists call associations. Associationism was one of the first general theories of cognition to be proposed in psychology (see James, W., 2007). However, it lost its vigour when, in the 1960s, Chomsky and other cognitive psychologists observed that this theory did not account for the complexity of human language production (Chomsky, 2006).

LLMs are indeed very sophisticated associators, and our research demonstrates how these associators can successfully complete tasks that were once thought to be unmanageable through associations. Indeed, how can a system that only predicts the next word perform well in cognitive tasks such as verbal reasoning or cognitive estimation?

The results reported here open the way to a systematic study of the limits of associationism in its most modern form (i.e., LLMs), whose exact boundaries are unclear. The efficiency of some LLMs, such as ChatGPT, in completing tasks at the level of the average human subject puts into question the theoretical framework that led cognitive psychology to surpass associationism.

4.3 Limitations of the study

A first limitation concerns the replicability of the data reported here. ChatGPT is subjected to ongoing revisions; therefore, it is likely that new updates will reduce the number of errors produced by ChatGPT in solving cognitive tasks. This issue of replicability can only be resolved by having access to a "frozen" LLM that is not subjected to continuous refinement.

A second limitation refers to the nature of ChatGPT. ChatGPT (or LLMs in general) is a complex neural network trained on massive amounts of *corpora* to learn statistical patterns and predict what words or phrases will likely come next in a given context.

One common observation is that LLMs are good at interpolation but not so good at extrapolation, whereas the former refers to predicting words within the range of the input data, while the latter bases its predictions outside this range (Zhan et al., 2022). Based on these considerations, it could be the case that GhatGPT's good performance in many prefrontal tasks is driven by a sort of memorisation since it may have been exposed to many of the tasks used here.

This may hold true for certain tasks, such as the comprehension of metaphors and idioms that are lexicalised in language and thus can be understood through semantic access rather than a genuine understanding of figurative language. In this regard, ChatGPT may have already encountered such metaphorical and idiomatic expressions during its training phase and may have learned the corresponding association with their real meaning in the absence of a true abstraction of the meaning. Similarly, ChatGPT may have retrieved the information needed to answer the items in the cognitive estimation task through memorisation rather than a form of reasoning in which response-relevant information is selected and monitored, as intended in the original test.
However, the good performance in some tests presented here may not be explained this way. For example, the items of the Winograd Schema to test anaphoric referencing have been developed by the authors and did not appear on the internet in any form. The same is true for the Verbal Reasoning Test and the Social Cognition battery, which was unavailable on the web and was made available upon request by the authors.

In summary, while it is true that LLMs can memorise large amounts of text, this does not mean they simply replicate the language they are trained on. Indeed, their ability to generate novel and coherent text suggests that their capabilities go beyond mere rote learning. A clear example in this regard is the study by Webb et al. (2022), who recently demonstrated how in a structurally similar textual version of the Raven test (used in psychology to measure IQ), GPT-3 showed emergent analogical reasoning performance (tested *zero-shot*) similar to that of human subjects even without ever having been exposed to that type of stimuli.

A research objective for the future will be evaluating more precisely how much of the human-like performance is simply due to memorisation (also known as interpolation) and how much is a real extrapolation (Carlini et al., 2022).

4.4 Future directions

Future research should identify the limits that increasingly sophisticated LLMs cannot further erode. To this aim, a neuropsychological framework should be adopted rather than developing new ad-hoc benchmarks.

The first advantage would be a more accurate assessment in terms of construct validity. In psychology, construct validity refers to the extent to which a measure (such as a behavioural task, a test, or a questionnaire) accurately reflects the theoretical concept it is intended to assess (Strauss & Smith, 2009).

For example, if a test is intended to measure "*fluid intelligence*", construct validity would be concerned with whether the test measures intelligence rather than other similar constructs, such as simple semantic knowledge.

The second advantage would be a more direct comparison with human cognition. In the field of AI, comparisons between machine and human performance are typically made by reporting the average accuracy (e.g., Talmor et al., 2022; Valmeekam et al., 2022; Levesque et al., 2012 among others), which is a single numerical value that summarises the average performance of a group of unspecified participants. However, this simple procedure ignores that human performance may show high variability, usually modulated by age and schooling, with poorer performance for older persons with low schooling. The use of percentiles allows researchers to provide information about where a score is positioned relative to the distribution of those collected from a given population. It is important to note that these factors can significantly impact human performance and ignoring them can lead to misleading conclusions about the relative strengths and weaknesses of machine vs human performance. For example, in the Absurdity Subtest of the Verbal Reasoning Test, the performance of ChatGPT was far *below* the average compared to humans aged 31 to 45 years and with more than 15 years of formal education. However, compared to humans aged 61-75 with 3-7 years of schooling, the performance of ChatGPT in the same task was within the lower range of normality.

This variation in the interpretation of results, contingent upon the human benchmark, highlights the importance of considering the full range of human performance when comparing machine and human cognitive functioning and behaviour.

Conclusion

This research used a neuropsychological assessment to challenge the cognitive abilities of ChatGPT -a recent large language model- and specifically investigated its performance in human prefrontal functions. The results showed that ChatGPT displays varying degrees of proficiency, with impairments in planning, absurdity comprehension, and understanding others' mental states and intentions. While it is clear that LLMs are currently unable to mimic human cognitive functioning accurately, technological advancements will likely decrease the number of cognitive tasks they will be unable to perform. The study also suggested that a neuropsychological framework help evaluate the construct validity of AI "*intelligence*" by identifying which cognitive functions are accurately simulated by language models. Additionally, this approach provided a more accurate and direct comparison with human cognition while considering the impact of demographic factors, such as age, gender, and education, on the full range of human performance.

References

Andrews, G., Halford, G. S., Chappell, M., Maujean, A., & Shum, D. H. K. (2014). Planning following stroke: A relational complexity approach using the Tower of London.Frontiers in Human Neuroscience, 8. https://doi.org/10.3389/fnhum.2014.01032

Ansburg, P.I. (2000). Individual differences in problem solving via insight. *Current Psychology, 19*, 143–146. https://doi.org/10.1007/s12144-000-1011-y

Axelrod, B. N., & Goldman, R. S. (1996). Use of demographic corrections in neuropsychological interpretation: How standard are standard scores?. *The Clinical Neuropsychologist*, *10*(2), 159-162. https://doi.org/10.1080/13854049608406677

Bambini, V., Arcara, G., Bosinelli, F., Buonocore, M., Bechi, M., Cavallaro, R., & Bosia, M. (2020). A leopard cannot change its spots: A novel pragmatic account of concretism in schizophrenia. *Neuropsychologia*, *139*, 107332. https://doi.org/10.1016/j.neuropsychologia.2020.107332

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

Beeman, M.J., & Bowden, E.M. (2000). The right hemisphere maintains solution-related activation for yet-to-be-solved problems. *Memory & Cognition, 28*, 1231–1241. https://doi.org/10.3758/bf03211823

Behrens, J. P., & Olteţeanu, A. M. (2020). Are all remote associates tests equal? an overview of the remote associates test in different languages. *Frontiers in Psychology*, *11*, 1125. https://doi.org/10.3389/fpsyg.2020.01125

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623). http://dx.doi.org/10.1145/3442188.3445922

Benzon, W. L. (2023). Discursive Competence in ChatGPT, Part 1: Talking with Dragons. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4318832

Bianchi, L. (1895). The functions of the frontal lobes. *Brain*, *18*(4), 497-522.https://doi.org/10.1093/brain/18.4.497

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6), https://doi.org/10.1073/pnas.2218523120

Bowden, T., & Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behavior Research Methods, Instruments, & Computers*, 35(3), 634-639. https://doi.org/10.3758/bf03195543

Bowie, C. R., & Harvey, P. D. (2006). Administration and interpretation of the Trail Making Test. *Nature Protocols*, *1*(5), 2277–2281. https://doi.org/10.1038/nprot.2006.390

Brothers, L. (1990). The social brain: a project for integrating primate behaviour and neurophysiology in a new domain. Concepts neurosci, 1, 27-51. http://dx.doi.org/10.7551/mitpress/3077.003.0029

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

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

Burgess, P. W., & Shallice, T. (1997). *The Hayling and Brixton tests*. Bury St. Edmunds, UK: Thames Valley Test Company.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2022). Quantifying memorisation across neural language models. *arXiv preprint arXiv:2202.07646*.

Carriedo, N., Corral, A., Montoro, P. R., Herrero, L., Ballestrino, P., & Sebastián, I. (2016). The development of metaphor comprehension and its relationship with relational verbal reasoning and executive function. *PLOS ONE*, 11(3), e0150289. https://doi.org/10.1371/journal.pone.0150289

ChatGPT; Zhavoronkov A. Rapamycin in the context of Pascal's Wager: Generative pre-trained transformer perspective. *Oncoscience*, *9*, 82–84. https://doi.org/10.18632/oncoscience.571

Chayer, C., & Freedman, M. (2001). Frontal lobe functions. *Current Neurology and Neuroscience Reports*, 1(6), 547-552. https://doi.org/10.1007/s11910-001-0060-4

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

Chomsky, N. (2006). *Language and mind*. Cambridge University Press.

Christidi, F., Migliaccio, R., Santamaría-García, H., Santangelo, G., & Trojsi, F. (2018). Social cognition dysfunctions in neurodegenerative diseases: neuroanatomical correlates and clinical implications. *Behavioural neurology. 2018*, 1–18. https://doi.org/10.1155/2018/1849794

Cipolotti, L., Spanò, B., Healy, C., Tudor-Sfetea, C., Chan, E., White, M., Biondo, F., Duncan J., Shallice, T., & Bozzali, M. (2016). Inhibition processes are dissociable and lateralised in human prefrontal cortex. *Neuropsychologia*, *93*, 1–12. https://doi.org/10.1016/j.neuropsychologia.2016.09.018

Cohen, M. (2021). Exploring RoBERTa's theory of mind through textual entailment.

Collette, F., Van der Linden, M., Delfiore, G., Degueldre, C., Luxen, A., & Salmon, E. (2001). The functional anatomy of inhibition processes investigated with the Hayling task. *Neuroimage*, 14(2), 258–267. https://doi.org/10.1006/nimg.2001.0846

Corben, L. A., Klopper, F., Stagnitti, M., Georgiou-Karistianis, N., Bradshaw, J. L., Rance, G., et & Delatycki, M. B. (2017). Measuring inhibition and cognitive flexibility in Friedreich Ataxia. *The Cerebellum*, *16*(4), 757–763. https://doi.org/10.1007/s12311-017-0848-7

Culbertson, W. C., Moberg, P. J., Duda, J. E., Stern, M. B., & Weintraub, D. (2004). Assessing the executive function deficits of patients with Parkinson's disease: Utility of the Tower of London-Drexel. *Assessment, 11*(1), 27-39. https://doi.org/10.1177/1073191103258590

Daigneault, S., & Braun, C. M. J. (1993) Working memory and the self-ordered pointing task: further evidence of early prefrontal decline in normal aging. *Journal of Clinical and Experimental Neuropsychology*, *15*(6), 881–895. https://doi.org/10.1080/01688639308402605

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

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dubois, B., Slachevsky, A., Litvan, I., & Pillon, B. F. A. B. (2000). The FAB: a frontal assessment battery at bedside. *Neurology*, *55*(11), 1621–1626. https://doi.org/10.1212/wnl.55.11.1621

Duncan, J., Emslie, H., Williams, P., Johnson, R., & Freer, C. (1996). Intelligence and the frontal lobe: The organisation of goal-directed behavior. *Cognitive psychology*, *30*(3), 257-303. https://doi.org/10.1006/cogp.1996.0008

Duncan, J. (2005). Frontal Lobe Function and General Intelligence: Why it Matters. *Cortex, 41*(2), 215–217. https://doi.org/10.1016/s0010-9452(08)70896-7

Dymowski, A. R., Owens, J. A., Ponsford, J. L., & Willmott, C. (2015). Speed of processing and strategic control of attention after traumatic brain injury. *Journal of Clinical and Experimental Neuropsychology*, 37(10), 1024–1035. https://doi.org/10.1080/13803395.2015.1074663

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

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics, 8*, 34-48. https://doi.org/10.1162/tacl_a_00298

Fodor, E. M. (1999). Subclinical inclination toward manic-depression and creative performance on the Remote Associates Test. *Personality and individual differences*, *27*(6), 1273-1283. https://doi.org/10.1016/s0191-8869(99)00076-8

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2022). Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. bioRxiv, 2022-12. http://dx.doi.org/10.1101/2022.12.23.521610

Garnham, A., & Oakhill, J. (1996). The mental models theory of language comprehension. Models of understanding text, 313-339. http://dx.doi.org/10.4324/9780203775899

Golden, C. J., & Freshwater, S. M. (1978). Stroop colour and word test. *PsycTESTS Dataset*. https://doi.org/10.1037/t06065-000

Glucksberg, S., & McGlone, M. S. (1999). When love is not a journey: What metaphors mean. *Journal of pragmatics*, *31*(12), 1541-1558. https://doi.org/10.1016/S0378-2166(99)00003-X

Heaton, R. K. (1981). Wisconsin card sorting test manual; revised and expanded. *Psychological Assessment Resources*, 5-57.

Hiser, J., & Koenigs, M. (2018). The multifaceted role of the ventromedial prefrontal cortex in emotion, decision making, social cognition, and psychopathology. *Biological psychiatry*, *83*(8), 638-647. https://doi.org/10.1016/j.biopsych.2017.10.030

Iluz-Cohen, P., & Armon-Lotem, S. (2013). Language proficiency and executive control in bilingual children. *Bilingualism: Language and Cognition, 16*(4), 884-899. https://doi.org/10.1017/s1366728912000788

James, W. (2007). *The principles of psychology*. Cosimo, Inc.

Johnson-Laird, P. N. (1988). Levels of Representation: Consciousness and the Computational Mind. Ray Jackendoff. MIT Press, Cambridge, MA, 1987. xvi, 356 pp., illus. $27.50. Explorations in Cognitive Science, vol. 3. A Bradford Book. *Science*, *239*(4847), 1546-1547. https://doi.org/10.1126/science.239.4847.1546

Kalyan, A., Kumar, A., Chandrasekaran, A., Sabharwal, A., & Clark, P. (2021). How much coffee was consumed during EMNLP 2019? Fermi problems: A new reasoning challenge for AI. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. http://dx.doi.org/10.18653/v1/2021.emnlp-main.582

Klein, T., & Nabi, M. (2019). Attention is (not) all you need for common-sense reasoning. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. http://dx.doi.org/10.18653/v1/p19-1477

Koechlin, E., Basso, G., Pietrini, P., Panzer, S., & Grafman, J. (1999). The role of the anterior prefrontal cortex in human cognition. *Nature, 399*(6732), 148-151. https://doi.org/10.1038/20178

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science*, 302(5648), 1181-1185 https://doi.org/10.1126/science.1088545

Kounios, J., & Beeman, M. (2009). The Aha! moment: The cognitive neuroscience of insight. *Current directions in psychological science*, *18*(4), 210-216. https://doi.org/10.1111/j.1467-8721.2009.01638.x

Kung, T. H., Cheatham, M., Medinilla, A., ChatGPT, Silos, C., De Leon, L., ... & Tseng, V. (2022). Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models. medRxiv, 2022-12. http://dx.doi.org/10.1101/2022.12.19.22283643

Lam, M., Eng, G. K., Rapisarda, A., Subramaniam, M., Kraus, M., Keefe, R. S., & Collinson, S. L. (2013). Formulation of the age–education index: Measuring age and education effects in neuropsychological performance. *Psychological assessment*, *25*(1), 61. https://doi.org/10.1037/a0030548

Lauro, L. J. R., Tettamanti, M., Cappa, S. F., & Papagno, C. (2008). Idiom comprehension: a prefrontal task?. *Cerebral Cortex, 18*(1), 162-170. https://doi.org/10.1093/cercor/bhm042

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

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

Mazza, M., Pino, M. C., Keller, R., Vagnetti, R., Attanasio, M., Filocamo, A., Le Donne, I., Masedu, F., & Valenti, M. (2022). Qualitative Differences in Attribution of Mental States to Other People in Autism and Schizophrenia: What are the Tools for Differential Diagnosis?. *Journal of autism and developmental disorders, 52*(3), 1283-1298. https://doi.org/10.1007/s10803-021-05035-3

Mednick, S.A., & Mednick, M.P. (1967). Examiner's manual: Remote Associates Test. Boston: Houghton Mifflin.

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

Merrill, W., Goldberg, Y., Schwartz, R., & Smith, N. A. (2021). Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand?. *Transactions of the Association for Computational Linguistics*, 9, 1047-1060. https://doi.org/10.1162/tacl_a_00412

Mikulincer, M., & Sheffi, E. (2000). Adult attachment style and cognitive reactions to positive affect: A test of mental categorisation and creative problem solving. *Motivation and Emotion*, *24*, 149-174. https://doi.org/10.1111/j.1751-9004.2008.00146.x

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive psychology*, *41*(1), 49-100. https://doi.org/10.1006/cogp.1999.0734

Morris, R. G., Ahmed, S., Syed, G. M., & Toone, B. K. (1993). Neural correlates of planning ability: Frontal lobe activation during the Tower of London test. *Neuropsychologia*, *31*(12), 1367-1378. https://doi.org/10.1016/0028-3932(93)90104-8

Moscovitch, M., & Winocur, G. (1995). Frontal lobes, memory, and aging. *Annals of the New York Academy of Sciences*, *769*(1 Structure and), 119–150. https://doi.org/10.1111/j.1749-6632.1995.tb38135.x

Nunberg, G., Sag, I. A., & Wasow, T. (1994). Idioms. *Language*, *70*(3), 491-538. https://doi.org/10.1353/lan.1994.0007

Ochsner, K. N., & Gross, J. J. (2005). The cognitive control of emotion. *Trends in cognitive sciences*, *9*(5), 242-249. https://doi.org/10.1016/j.tics.2005.03.010

O'Connor S, ChatGPT. Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse? *Nurse Education in Practice*, *66*, 103537. https://doi.org/10.1016/j.nepr.2022.103537

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155

Owen, A. M., Downes, J. J., Sahakian, B. J., Polkey, C. E., & Robbins, T. W. (1990). Planning and spatial working memory following frontal lobe lesions in man. *Neuropsychologia*, *28*(10), 1021-1034. https://doi.org/10.1016/0028-3932(90)90137-d

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.

Papagno, C. (2001). Comprehension of metaphors and idioms in patients with Alzheimer's disease: A longitudinal study. *Brain*, *124*(7), 1450–1460. https://doi.org/10.1093/brain/124.7.1450

Pietrini P, Guazzelli M, Basso G, Jaffe K, Grafman J. Neural correlates of imaginal aggressive behavior assessed by positron emission tomography in healthy subjects. Am J Psychiatry. 2000 Nov;157(11):1772-81. doi: 10.1176/appi.ajp.157.11.1772. PMID: 11058474.

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

Prystawski, B., Thibodeau, P., & Goodman, N. (2022). Psychologically-informed chain-of-thought prompts for metaphor understanding in large language models. arXiv preprint arXiv:2209.08141.

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023). Is ChatGPT a General-Purpose Natural Language Processing Task Solver?. *arXiv preprint arXiv:2302.06476*.

Ralli, A. M., Chrysochoou, E., Roussos, P., Diakogiorgi, K., Dimitropoulou, P., & Filippatou, D. (2021). Executive function, working memory, and verbal fluency in relation to non-verbal intelligence in Greek-speaking school-age children with Developmental Language Disorder. *Brain Sciences*, *11*(5), 604. https://doi.org/10.3390/brainsci11050604

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

Robinson, G. A., Cipolotti, L., Walker, D. G., Biggs, V., Bozzali, M., & Shallice, T. (2015). Verbal suppression and strategy use: A role for the right lateral prefrontal cortex? *Brain, 138*(4), 1084–1096. https://doi.org/10.1093/brain/awv003

Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. arXiv preprint arXiv:2005.04118.

Rossetti, I., Brambilla, P., & Papagno, C. (2018). Metaphor comprehension in schizophrenic patients. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00670

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

Saykin, A. J., Gur, R. C., Gur, R. E., Shtasel, D. L., Flannery, K. A., Mozley, L. H., ... & Mozley, P. D. (1995). Normative neuropsychological test performance: effects of age, education, gender and ethnicity. *Applied Neuropsychology*, *2*(2), 79-88. https://doi.org/10.1207/S15324826AN0202_5

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., & Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, *4*(3), 258-268. https://doi.org/10.1038/s42256-022-00458-8

Shallice, T., & Evans, M. E. (1978). The involvement of the frontal lobes in cognitive estimation. *Cortex*, *14*(2), 294–303. https://doi.org/10.1016/s0010-9452(78)80055-0

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

Spitoni, G. F., Bevacqua, S., Cerini, C., Ciurli, P., Piccardi, L., Guariglia, P., Pezzuti, L., & Antonucci, G. (2018). Normative data for the Hayling and Brixton tests in an Italian population. *Archives of Clinical Neuropsychology*, *33*(4), 466-476.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., ... & Kim, H. (2022). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615.

Stanhope N., Guinan E., Kopelman M.D. (1988) Frequency judgements of abstract designs by patients with diencephalic, temporal lobe or frontal lobe lesions. *Neuropsychologia*, *36*(12), 1387–1396. https://doi.org/10.1016/s0028-3932(98)00040-2

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual review of clinical psychology*, *5*, 1-25. https://doi.org/10.1146/annurev.clinpsy.032408.153639

Sturm, W. Neuropsychological assessment. *J Neurol* **254** (Suppl 2), II12–II14 (2007). https://doi.org/10.1007/s00415-007-2004-7

Stuss, D. T., & Alexander, M. P. (2000). Executive functions and the frontal lobes: a conceptual view. *Psychological Research*, *63*(3–4), 289–298. https://doi.org/10.1007/s004269900007

Sun, F. (2022). ChatGPT, the Start of a New Era.

Talmor, A., Yoran, O., Bras, R. L., Bhagavatula, C., Goldberg, Y., Choi, Y., & Berant, J. (2022). Commonsenseqa 2.0: Exposing the limits of ai through gamification. arXiv preprint arXiv:2201.05320.

Valmeekam, K., Olmo, A., Sreedharan, S., & Kambhampati, S. (2022). Large Language Models Still Can’t Plan (A Benchmark for LLMs on Planning and Reasoning about Change). arXiv preprint arXiv:2206.10498.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Verbruggen, F., & Logan, G. D. (2008). Automatic and controlled response inhibition: Associative learning in the go/no-go and stop-signal paradigms. *Journal of Experimental Psychology*: *General*, *137*(4), 649–672. https://doi.org/10.1037/a0013170

Videsott, G., Herrnberger, B., Hoenig, K., Schilly, E., Grothe, J., Wiater, W., ... & Kiefer, M. (2010). Speaking in multiple languages: Neural correlates of language proficiency in multilingual word production. *Brain and language*, *113*(3), 103-112. https://doi.org/10.1016/j.bandl.2010.01.006

Vohs, K. D., & Heatherton, T. F. (2001). Self-Esteem and threats to self: implications for self-construals and interpersonal perceptions. *Journal of personality and social psychology*, *81*(6), 1103. https://doi.org/10.1037/0022-3514.81.6.1103

Webb, T., Holyoak, K. J., & Lu, H. (2022). Emergent Analogical Reasoning in Large Language Models. arXiv preprint arXiv:2212.09196.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022a). Emergent abilities of large language models. arXiv preprint arXiv:2206.07682.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022b). Chain of thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903.

Weiss, E. M., Kemmler, G., Deisenhammer, E. A., Fleischhacker, W. W., & Delazer, M. (2003). Sex differences in cognitive functions. *Personality and individual differences*, *35*(4), 863-875. https://doi.org/10.1016/S0191-8869%2802%2900288-X

Wilson, B. A., Alderman, N., Burgess, P. W., Emslie, H., & Evans, J. J. (1996). *BADS: Behavioural assessment of the dysexecutive syndrome*. London: Pearson.

Zhan, J., Xie, X., Mao, J., Liu, Y., Guo, J., Zhang, M., & Ma, S. (2022, October). Evaluating Interpolation and Extrapolation Performance of Neural Retrieval Models. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (pp. 2486-2496).