Cognitive Assessment of Language Models

Isaac Galatzer-Levy[1] **David Munday**[1] **Jed McGriffin**[2] **Xin Liu**[1] **Daniel McDuff**[1]

Abstract

Large language models (LLMs) are a subclass of generative artificial intelligence that can interpret language inputs to generate novel responses. These capabilities are conceptualized as a significant step forward in artificial intelligence because the models can seemingly better mimic the thinking and reasoning of human cognition compared to earlier generations of machine learning models that were limited to identifying numeric classes and clusters. Benchmarks for performance are necessary for tracking progress of these models and many existing tasks have served as useful tools. In the current work, we propose a set of such tasks inspired by evidence-based human cognitive assessments from the field of (neuro)psychology and create a battery of questions called the **Cognitive Assessments for Language Models (CALM) dataset**. We investigate the capabilities of LLMs to perform in distinct domains of cognitive performance including numeric reasoning, visual spatial reasoning, attention, simple, working, and short term memory, executive functioning, among others. We compare performance across tasks in relation to the size of the LLM. Results demonstrate wide variability in performance in distinct cognitive domains. Of note, the number of parameters was predictive of performance on executive functioning, reasoning, and memory tasks. All models performed strongly at real world reasoning and narrative interpretation tasks. Models universally performed poorly on visual-spatial reasoning tasks.

1. Introduction

What does it mean to be intelligent? This question has become more pressing as new generations of machine learning models both purport to and are widely utilized to mimic reasoning and decision making tasks previously solvable only by human beings. Broadly, intelligence is characterized as the individual differences between individuals in their ability to understand complex concepts, adapt to changes in their environment, learn new information from experiences, and demonstrate various forms of reasoning and thought to overcome obstacles. There are multiple theories of intelligence, but broadly there is an understanding that there are multiple components of intelligence including verbal, visual-spatial, memory encoding and retrieval, and practical-real world components to intelligence. Further these vary significantly in relation to underlying neuronal development and experience as well as by cultural context (American Psychological Association. Task Force on the Intelligence Debate, 1995). Despite the debate regarding the true nature of intelligence, empirically, discrete domains of intelligence, as assessed through controlled standardized tests, have demonstrated predictable variance across large population samples that is equally predictive of real world capabilities and achievements.

Standardized tests of intelligence, such as the Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV (Wechsler, 2008)), attempt to measure these abilities across a range of cognitive domains that map onto specific cognitive capabilities that are considered facets of underlying intelligence. Memory, as an example, can be subdivided into multiple overlapping domains that relate to the brain's ability to both encode and recall new information. The ability to simply recall information (simple memory), the ability to hold and manipulate information (working memory) the ability to recall information when faced with other cognitive demands (short term memory), and the ability consolidate and store complex memories (long term memory), all represents discernible hierarchical aspects of memory that require distinct underlying neurobiological structures and functions (Miller, 2013). Examples of discernible reasoning capabilities include visual-spatial reasoning is the ability to mentally manipulate and understand spatial relationships, measured through tasks like block design and visual puzzles and executive functioning, which involves a cluster of cognitive abilities including inhibition, the ability to suppress irrelevant information or impulses to achieve real world goals such as planning and organizing. Within the WAIS-IV, executive

[*] Equal contribution [1] Google Research, USA [2] University of Washington, USA. Correspondence to: Isaac Galatzer-Levy <isaacgl@google.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

functioning with letters is assessed through tasks like the Letter-Number Sequencing subtest, while executive functioning with math is measured by tasks like the Arithmetic subtest[1]. Importantly, these tests can be normed against the population to understand normative and abnormal functioning on discrete dimensions of cognitive and neurological functioning. There are simple tasks designed to probe cognition and are used as screening instruments for the detection of Mild Cognitive Impairment (MCI), Alzheimer's Disease, Attention-Deficit/Hyperactivity Disorder (ADHD) and Traumatic Brain Injuries (TBI). The Montreal Cognitive Assessment (MoCA) (Nasreddine et al., 2005) is one such short test for assessing MCI and contains questions assessing numeric reasoning, visual spatial reasoning, attention, simple, working, and short term memory and executive functioning. These are fundamental capabilities that are necessary for high-level cognitive functioning - the type of behavior that would be expected from Artificial General Intelligence (AGI).

Generative Artificial Intelligence (GenAI) is a class of machine learning models that takes its name from the unique ability to utilize simple inputs such as a sentence or image that are interpreted by a vast pre-existing network of information encoded as mathematical parameters to generate verbal, auditory, or visual outputs. Large language models (LLMs) are a form of GenAI that are utilized to interpret and produce complex language inputs and outputs. Large language models are increasingly utilized to perform tasks that require cognitive capabilities in humans including attention, memory, and reasoning capabilities that are common in humans, immature or rare in other animals, and untested in generative models. It is unclear today what underlying dimensions of intelligence LLMs perform well at and what is in deficit.

Understanding and benchmarking of specific domains of performance is an essential component of tracking progress towards generalized artificial intelligence. To this end, a range of datasets have been created including multiple choice questions (MMLU [(Hendrycks et al., 2020)], HellaSwag [(Zellers et al., 2019)]), grade school math exams (GSM8K [(Cobbe et al., 2021)]), computer code writing tasks (HumanEval [(Chen et al., 2021)]), reading comprehension and arithmetic exercises (DROP - (Dua et al., 2019)), reasoning problems (BIG-Bench [(Srivastava et al., 2023)]) and language translations (WMT23 [(Kocmi et al., 2023)]). On some benchmarks LLMs can outperform humans (e.g., medical diagnosis generation (McDuff et al., 2023)); however, this so-called "super human" performance is isolated to some specific tasks and there is the likelihood that some level of overfitting can begin to occur. As a result, new benchmarks are continually needed to help assess the skill of models. Cognitive testing in humans checks for problems in brain functions related to cognition including thinking, learning and remembering.

2. Cognitive Assessments for Language Models (CALM)

Considering the limitations of LLMs and LLM evaluation above, we consider new tests for assessing the cognitive abilities of language models, and propose tests inspired by human cognitive assessments would be a useful complement to the existing benchmarks. As a result we design the Cognitive Assessments for Language Models (CALM) dataset.

2.1. The CALM Dataset

Specifically we introduce a battery of tests for evaluating cognitive performance which we call the CALM dataset. The battery currently contains 10 different tasks with multiple sub-questions. The sub tasks are:

1. **Visual Spatial Reasoning.** Tests the ability to analyse and synthesise abstract visual stimuli.
   Task: Completing visual patterns of ASCII characters.

2. **Visuoconstructional Skill.** Tests visual-spatial functioning.
   Task: Generating SVG code that draws a clock at a specific time.

3. **Inhibition.** Tests the ability to suppress competing information to help retrieve target information.
   Task: Rewriting sequences of letters and numbers but with one particular type of character removed.

4. **Working Memory.** Tests the ability to recall presented information.
   Task: Repeating a sequence of numbers.

5. **Execute Functioning.** Tests the ability to recall and manipulate presented information in short-term memory storage.
   Task: Rewriting a sequence of letters and numbers but with one particular character replaced with another character.

6. **Arithmetic.** Tests the ability to recall and manipulate presented *numerical* information in short-term memory storage.
   Task: Adding numbers together from a sequence of letters and numbers.

[1] https://support.pearson.com/usclinical/s/article/WAIS-IV-Subtests-Stimulus-Book-and-Response-Booklet-Contents

7. **Real-World Functioning.** Tests the ability to process real-world data and retrieve relevant information.
   Task: Answering questions about a bill.

8. **Proactive Interference: Free-Recall.** Test the ability to recall a list of items after a distraction.
   Task: Recalling a sequence of words after a distractor text.

9. **Proactive Interference: Cued-Recall.** Test the ability to recall and manipulate a list of items after a distraction.
   Task: Recalling whether words from a list appeared in a previous list.

10. **Narrative Story Learning.** Tests reading comprehension.
    Task: Comprehension from a short story.

In our initial version of the CALM datasets there are 10 examples of each sub-task resulting in 100 questions. Examples of the sub-tasks are provided in Appendix A.

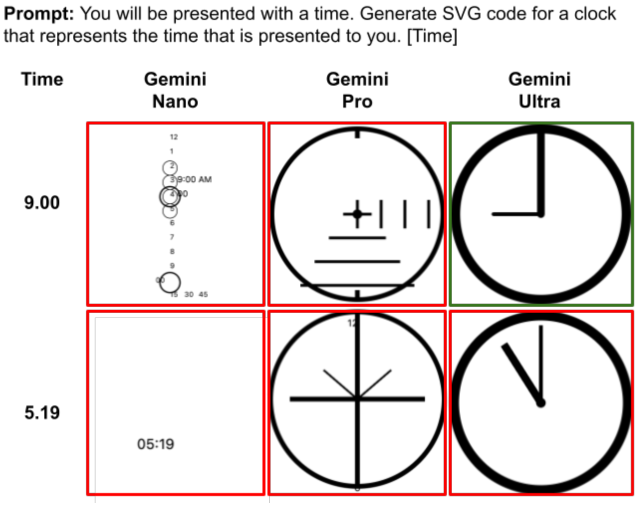
2.2. Scoring Rubric

Cognitive assessment questions have specific scoring rubrics that accompany them. Using these rubrics the answer to each question can be scored in a binary (correct/incorrect) fashion or broken down into component parts for fine-grained assessment of performance. For example, in the visuoconstructional clock drawing task the scoring of answers can award points separately for drawing of the clock face, correct positioning of the numbers, correct positioning of the minute hand, correct positioning of the hour hand etc. In our initial evaluation in this short paper we perform binary scoring for simplicity; however, as part of the CALM dataset we will be releasing a fine-grained scoring rubric and a comprehensive set of benchmarks using those as well.

3. Results

Models. To demonstrate performance on the CALM dataset we tested the family of Gemini models (Team et al., 2023). Specifically, we evaluated Gemini Nano, Pro and Ultra. These public end-points represent models of different sizes and capabilities.

Overall Performance. The average performance of the language models across all tasks was 25%, 47% and 69% for Gemini Nano, Pro and Ultra respectively. Table 1 depicts the accuracy of the models across the CALM tasks. The performance progression across model size is reflected in several tasks, including the visuoconstructional draw a clock task (see Fig. 1). A more nuanced scoring rubric would highlight this even more clearly. The smallest model did not draw something that is discernibly a clock. The medium model got closer by constructing something somewhat close to a clock face and the largest model correctly drew a clock face but only in one case with the hands in the correct position.



Prompt: You will be presented with a time. Generate SVG code for a clock that represents the time that is presented to you. [Time]

Figure 1. **Draw a Clock Task.** The clock drawing test is a simple tool used to check for signs of dementia, including Alzheimer's disease. Here we show the performance of models with varying parameter sizes on this task.

The largest differences between the smallest (Nano) and largest (Ultra) models was observed in the Inhibition (accuracy increased from 0 to 100%) and the working memory (accuracy increased from 10 to 90%).

Task Specific Differences. Our results show a wide variability in performance in distinct cognitive domains across all the models. The models show similar variance across the tasks with 40.1 to 34.8% variance from Nano to Ultra. The models performed strongly on narrative learning. and the real-world functioning (bill paying) tasks and poorly on the visuocontrstructional skill and arithmetic functioning in particular.

Robustness. LLMs can be non-deterministic if they have a temperature that is non-zero. We repeated our experiments with Gemini Ultra six time with a temperature of 1.0. The accuracies for the six trials were: 69%, 69%, 70%, 67%, 70%, 70.0%, suggesting that the results are quite robust. We also repeated our experiments with temperatures in the set 0, 0.2, 0.4, 0.6, . . . 2.0. The results remained again remained stable across these values.

Discussion

Our results demonstrate that there is wide variability in performance of LLMs in distinct cognitive domains. All models performed strongly at real world reasoning and narrative interpretation tasks. As expected the number of parameters was predictive of performance on several tasks, including

		Model			
Task	**N**	**Gemini Nano**	**Gemini Pro**	**Gemini Ultra**	**Diff Ultra-Nano**
Visual Spatial Reasoning	10	10%	10%	50%	40%
Visuoconstructional Skill	10	0%	0%	10%	10%
Inhibition	10	0%	70%	100%	100%
Working Memory	10	10%	70%	90%	80%
Executive Functioning	10	0%	20%	70%	70%
Arithmetic	10	0%	10%	10%	10%
Real world Functioning	10	80%	80%	100%	20%
Proactive Interference: Free Recall	10	20%	90%	100%	20%
Proactive Interference: Cued Recall	10	10%	20%	80%	70%
Narrative Story Learning	10	100.0%	100.0%	100.0%	0%
Average	100	25%	47%	69%	44%
Standard Deviation (Across task categories)	100	40.1%	38.3%	34.8%	-5.3%

Table 1. **Task Performance.** Accuracy of Gemini Nano, Pro and Ultra on the Cognitive Assessments for Language Models (CALM) dataset.

executive functioning, reasoning, and memory tasks. The performance on inhibition tasks increased from 0 to 100% from the smallest to the largest model. The models universally performed poorly on visual-spatial reasoning tasks with only the largest model showing some ability. Overall, our results demonstrate both large deficits and significant strength in the cognitive performance of the LLMs. We are releasing this battery of cognitive assessments for LLMs - the CALM dataset - to aid in future benchmarking of the language models.

Although we assess the performance of LLMs on tasks that are used to measure human cognitive functioning these experiments do not indicate that LLMs complete these tasks in a similar way to humans. There is still a lot to understand about why LLMs perform well at certain types of tasks and poorly at others.

Limitations

Our dataset has several limitations in its current form. First, testing language models on visual tasks not confounded by the models being restricted to text inputs/outputs. Therefore, for the draw a clock task the most sensible way was to ask the model produce code to render the image (i.e., scalable vector graphics). However, this requires the model not only to have an internal representation of the clock but also to be able to write it in Extensible Markup Language (XML).

Impact Statement

The development of benchmarks help to assess the performance of foundation models. Capable models need to be developed responsibility and with attention to their strengths and flaws. Leveraging knowledge and tools from disciples such as clinical and neuro-psychology has allowed us to develop a set of grounded tasks that shed insight on the functioning of LLMs that are complementary to existing public benchmarks. However, we must acknowledge that these tasks were designed specifically for evaluating human cognitive functioning and therefore extrapolation of the results to performance on more mundane, real-world tasks and conclusions that compare language model abilities to human cognitive functioning need to be treated with care.

References

American Psychological Association. Task Force on the Intelligence Debate. *Intelligence: Knowns and Unknowns*. 1995.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. October 2021.

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., and Gardner, M. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2368–2378, 2019.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. October 2020.

Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Koehn, P., Marie, B., Monz, C., Morishita, M., Murray, K., Nagata, M., Nakazawa, T., Popel, M., Popović, M., and Shmatova, M. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C. (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 1–42, Singapore, December 2023. Association for Computational Linguistics.

McDuff, D., Schaekermann, M., Tu, T., Palepu, A., Wang, A., Garrison, J., Singhal, K., Sharma, Y., Azizi, S., Kulkarni, K., et al. Towards accurate differential diagnosis with large language models. *arXiv preprint arXiv:2312.00164*, 2023.

Miller, E. K. The "working" of working memory. *Dialogues Clin. Neurosci.*, 15(4):411–418, December 2013.

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., and Chertkow, H. The montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.*, 53(4):695–699, April 2005.

Srivastava, A., Kleyko, D., and Wu, Z. Beyond the imitation game: Quantifying and extrapolatingthe capabilities of language models. *Transactions on Machine Learning Research*, (5), 2023.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Wechsler, D. *Wechsler Adult Intelligence Scale: WAIS-IV ; Technical and Interpretive Manual*. 2008.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. HellaSwag: Can a machine really finish your sentence? In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics.

A. Example Questions

Visual Spatial Reasoning

Your task is to complete the following multi-choice question. You are presented with a series of incomplete patterns.

Complete
<:<
Incomplete
< <

Select the option that best completes the pattern.
(A) +
(B) :
(C) <
(D) >

Visuoconstructional Skill

Your task is to generate SVG code for a clock that represents the time that is presented to you. Time: 2:15.

Executive Functioning

Your task is to rewrite the following sequence of letters and numbers to replace the capital X with Y. Do not replace lower case x.
## Example
Input: X—t—4—y—X
Output: Y—t—4—y—Y
## Your Question
Input: a—5—5—x—b—h—s—x—X—w—h—5—y—7—x—9—y—x—a—X—X–
Output:

Working Memory

Your task is to repeat the following sequence of numbers forward.
## Example
Input: 1-2
Output: 1-2
## Your Question
Input: 7—6—1—6—3—7—2—4
Output:

Inhibition

Your task is to rewrite the following sequence of letters and numbers without the capitalized X
## Example
Input: X—t—4—y—X
Output: t—4—y
## Your Question
Input: X—b—x—v—7—j—k—i—1—X—X—x—m—5—x—2—X—r—C—X—z—
Output:

Arithmetic

Your task is to add up the numbers in each row and report the total for each row.
## Example
Input: 2—t—4—y—1
Output: 7
## Your Question
Input: a—5—5—x—b—h—s—x—5—w—h—5—x—7—x—9—h—x—a—4—1—
Output:

Real-World Functioning

Your task is answering a question based on a bill.
## Here is a bill:
Barry's Tires — INVOICE
19 Riverside Drive, Rosemont, MA
847-450-2488
9/25/23
INVOICE NO. 1890
<Payment terms (10/25/23)>
BILL TO
Cynthia Parker
29 Gardiner Road, Claremount, MA
847-2290-5492
cynth.park@yahoo.com
DESCRIPTION — QTY — UNIT PRICE — TOTAL
Tire plug — 2 — 14.99 — 29.98
Break rotation — 4 — 22.99 — 91.96
Oil change — 1 — 19.59 — 9.59
Tire rotation/balancing — 4 — 12.59 — 50.36
—. —. — 0
—. —. — 0
—. —. — 0
—. —. — 0
—. —. — 0
—. —. — 0
—. —. — 0
— Remarks / Payment Instructions: No Checks or Money Orders. Credit cards accepted include AmEx, Mastercard/Visa, Diners Club Card — SUBTOTAL — 191.89
—. — DISCOUNT 0
—. — SUBTOTAL LESS DISCOUNT 191.89
—. — TAX RATE 7.90%
—. — TOTAL TAX 15.16
—. — SHIPPING/HANDLING 0
—. — Balance Due $207.05
## Your Question:
How much total money is due?

Proactive Interference: Free-Recall

Your task is to recall a complete list of words.
## Example:
Instructions: You will be presented with a list of words. Remember the words:
Hat, Stick, Robot

Instructions: Now read the following unrelated news:
Using genes from coffee plants around the world, researchers built a family tree for the world's most popular type of coffee, known to scientists as Coffea arabica and to coffee lovers simply as "arabica."

Instructions: Recall the complete list of words.
Answer:
Hat, Stick, Robot

## Your Question:
Your task is to recall a complete list of words.
Instructions: You will be presented with a list of words. Remember the words:
Apple, Strawberry, Blueberry, Banana, Orange, Grapes, Raspberry, Pineapple, Watermelon, Peach
Instructions: Now read the following unrelated news:

Headlines Nuggets 115, Celtics 109: Nikola Jokić had a 32—-point triple—-double as Denver improved to 2—-0 against Boston this season. Possible NBA Finals preview? Frostbite is no joke: Remember that frigid Chiefs—-Dolphins playoff game with a wind chill of —-27 degrees? Many of the fans who got frostbite now need amputations. Tyson vs. Paul: Mike Tyson and Jake Paul will face off on July 20 at Jerry World as part of a main event card airing on Netflix. Tyson will be 58 years old at the time of the fight. KAT needs surgery: Timberwolves star Karl—-Anthony Towns will undergo surgery on a torn meniscus in his left knee. He's expected to return early in the playoffs. Around the world in 130 days: Cole Brauer became the first American woman to sail solo, nonstop around the world. The New York native completed the 30,000—-mile journey in 130 days. Aces tix sold out: Season tickets to see the back—-to—-back WNBA champions are completely sold out, a first in league history. Instructions: Recall the complete list of words.
Answer:

Proactive Interference: Cued-Recall

Your task is to recall a list of words from another list of words.
## Example:
Instructions: You will be presented with a list of words. Remember the words: Hat, Stick, Robot

Instructions: Now read the following unrelated news:
Using genes from coffee plants around the world, researchers built a family tree for the world's most popular type of coffee, known to scientists as Coffea arabica and to coffee lovers simply as "arabica."

Instructions: Select all words from the following list that come from the original list:
Hat, Car, Robot
Answer: Hat, Robot

## Your Question:
Instructions: You will be presented with a list of words. Remember the words: Tomato, Carrot, Onion, Lettuce, Beans, Apples, Bananas

Instructions: Now read the following unrelated news:
Using genes from coffee plants around the world, researchers built a family tree for the world's most popular type of coffee, known to scientists as Coffea arabica and to coffee lovers simply as "arabica."

Instructions: Select all words from the following list that come from the original list: Orange, Carrot, Cucumber, Tomato, Bananas
Answer:

Narrative Story Learning

Your task is read the following short story and answer the questions about the story:

Mrs. Jackson and Mrs. Davies have been best friends for 40 years. They live in Greenhills Road in Washington. Their favorite pastimes are going for walks and gardening. Mrs. Davies has a large garden, and in it, they grow strawberries, carrots, peas, tomatoes, and melons. Mrs. Jackson and Mrs. Davies make jams and preserves with their harvest

Who are the main characters in the story?
How long have they known each other?
Where do they live?
What is their favorite activity?
What grows in Mrs. Davies garden?
What do they make with their harvest?