Cognitive Flexibility of Large Language Models

Sean M. Kennedy [1 2] **Robert D. Nowak** [1]

Abstract

Cognitive flexibility is a property of cognitive systems which enables success in rapid adaptation to new tasks in quick succession. We investigate the degree of cognitive flexibility exhibited by several Large Language Models by evaluating them on two neuropsychological tests, the Wisconsin Card Sorting Test and the Letter-Number Test. Our findings indicate that some Large Language Models fail to switch tasks within the same context window, despite succeeding at these same tasks in distinct context windows, while others are able to flexibly switch tasks.

1. Introduction

Humans possess the ability to switch between different conceptual representations of objects depending on the context in which they appear. For example, consider three fruits: bananas, apples, and plantains. When asked which two fruits are most similar out of the three, absent any context, most people would respond that plantains and bananas are more similar because they look alike and one might suspect that they must be closely related genetically. If we provide the context of fruit found in a fruit bowl, people would probably select apples and bananas as more similar, as they are more likely to be found in this setting than a plantain. It seems that we create a malleable conceptual manifold that allows us to judge similarity between objects in different contexts.

To some degree, it seems like this ability to adapt to variable task-specific representations is present in transformer models, the prevalent deep learning modeling paradigm for language and vision tasks. For example, Large Language Models (LLMs) possess the ability to learn a new task in-context from few samples. In-context learning has proven to be a useful method to effectively finetune an LLM to perform a variety of tasks (Wei et al., 2021). However, it is unclear to what degree LLMs are able to rapidly switch tasks within a single context window, an ability that is analogous to human *cognitive flexibility*.

Cognitive flexibility is the property of a cognitive system that provides the ability to efficiently switch between different concepts and tasks. The presence of cognitive flexibility in humans is crucial for rapidly switching tasks and allows for adaptation to new or changing tasks (Ionescu, 2012). The capacity for cognitive flexibility in humans can be measured by performing neuropsychological tests [1]. Cognitive flexibility tests typically involve stimuli which need to be processed in different ways depending on a specific task in a set of tasks that change throughout the test. To succeed at the test, the participant needs to infer the correct way to process the stimuli and switch to a different processing strategy if they are first unsuccessful. A single test might involve switching between one or more tasks and the associated strategies to succeed at the task many times.

To investigate the cognitive flexibility of LLMs, we need to employ structured and well-established methods. Therefore, we chose two neuropsychological tests: the Wisconsin Card Sorting Test (WCST) (Grant & Berg, 1948) and the Letter-Number Test (LNT) (Rogers & Monsell, 1993). These tests are commonly used to measure cognitive flexibility in humans and can be adapted as prompts for LLMs. In the following section, we describe our approach to assessing the cognitive flexibility of several state-of-the-art LLMs using these tests.

1.1. Related Works

A significant body of work has assessed LLMs using benchmarks that measure general language understanding and task-specific competencies. Notable among these are the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2019b) and SuperGLUE benchmark (Wang et al., 2019a) which provide a comprehensive suite of tasks for evaluating natural language understanding. More recent benchmarks such as the Massive Multitask

[1] University of Wisconsin-Madison, Wisconsin, USA [2] Warfighter Interactions & Readiness Division, Air Force Research Laboratory, Wright-Patterson AFB, USA. Correspondence to: Sean M. Kennedy <smkennedy5@wisc.edu>.

1st Workshop of Large Language Models and Cognition at the 41st International Conference on Machine Learning , Vienna, Austria.

[1] A standard use case for these tests is to reveal underlying brain injury or neurodegenerative disease, which typically involve damage to the frontal lobe and basal ganglia (Eslinger & Grattan, 1993).

Language Understanding (MMLU) (Hendrycks et al., 2021) and BigBench (Srivastava et al., 2023) benchmarks measure LLM multitask performance across a diverse set of tasks, assessing models' abilities to handle questions from various domains such as history, science, and mathematics, as well as tasks that require complex reasoning and problem-solving skills. While these benchmarks are instrumental in measuring the linguistic and reasoning capabilities of LLMs, they do not explicitly test cognitive flexibility.

There has been growing interest in evaluating LLMs using cognitive tests traditionally used for human subjects. One such study investigates LLMs' performance on Theory of Mind tasks, revealing that certain LLMs can perform at or above human levels in specific ToM tasks such as identifying indirect requests and false beliefs (Strachan et al., 2024).

2. Approach

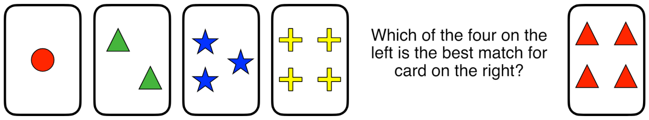

Which of the four on the left is the best match for card on the right?

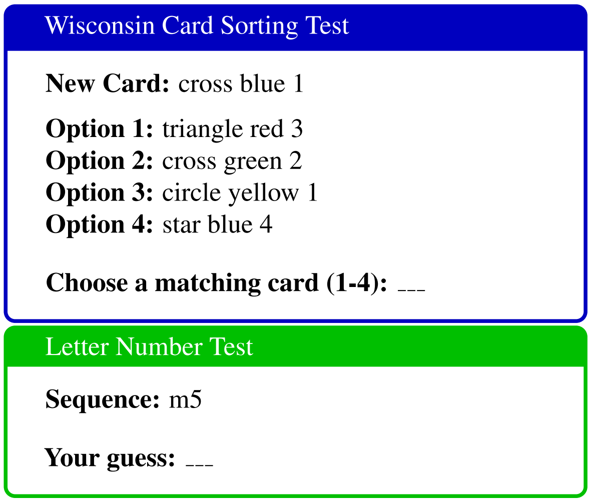
Figure 1. In the WCST, participants identify a matching card without explicit instructions on a matching criteria and receive feedback if there choice was correct or incorrect.

The WCST tasks participants to match a given card containing one or more colored shapes one of four presented cards. The cards can match based on three criteria; shape, color, and number of shapes present. In each trial, there is one card that matches based on each of the three matching rules, and one card that does not match on any criteria. Initially, there will be a matching rule selected which will change at some point during the test. The currently selected rule is not initial know to the participant and must be discovered through exploration. Once discovered, a participant must preserve the knowledge of which matching rule worked previously, and continue to apply that rule. Typically, after a fixed number of successful guesses in succession the rule changes. Once the rule changes, the participant will be incorrect after applying the previous rule, and will need to discover and begin applying a new rule. To succeed at this test, participants need to exhibit cognitive flexibility for switching between different matching strategies.

The LNT is a relatively simpler test that tasks participants to respond to a two-character sequence, which consists of one number followed by one number. If the current task is ”Letter” then the participant needs to respond by indicating if the letter is a consonant or a vowel. If the current task is ”Number” then the participant needs to respond by indicating if the number is odd or even. Similar to the WCST, there is no explicit instructions on which task to use. This test has the fewest number of tasks possible for task switching at two, and the individual tasks are relatively simple and easy.

Adapting Tests for LLMs

The WCST is a visual task, involving physical cards or images of the cards as stimuli. To adapt this task for LLMs, we represent each card as a text description of the card's visual characteristics. For example, the card on the right in Fig. 1 would be representation ”triangle red 4”. The LNT is already a text-based test, so there is no need to modify the stimuli. We provide a task description to the LLMs prior to performing the test (see Appendix A).[2]



Wisconsin Card Sorting Test

New Card: cross blue 1

Option 1: triangle red 3
Option 2: cross green 2
Option 3: circle yellow 1
Option 4: star blue 4

Choose a matching card (1-4): ___

Letter Number Test

Sequence: m5

Your guess: ___

Figure 2. Examples of prompts for the Wisconsin Card Sorting Test (top) and the Letter Number Test (bottom).

3. Results

We evaluated four LLMs including OpenAI's ChatGPT-3.5 Turbo and ChatGPT-4, Google's Gemini 1.5-Pro, and Meta's Llama 70B. Each LLM is evaluated on the two tests for cognitive flexibility, WCST and LNT. We also separate each test into the individual component tasks involved in the test and evaluate some models on these tasks. For example, the Color task in the WCST involves matching cards based on color, and the rule does not change during the test. We do this in the cases where the model performs poorly on the overall test to see if the model is able to achieve success in the individual tasks absent of any task switching.

For each model type, we conducted eight separate evaluations where each model was subjected to twenty-five trials of the designated test. Here a trial is one prompt-response pair, and an evaluation is a sequence of trials. This approach enabled us to gather comprehensive data on the models' abil-

[2] The related code and data is available at github.com/kenneds6/LLM-cognitive-flexibility

Table 1. Average Accuracy and Standard Deviation Across WCST Tasks

Model	Standard Test	Shape	Color	Number
ChatGPT 3.5 Turbo	0.251 (0.1873)	0.377 (0.1212)	0.330 (0.0595)	0.940 (0.0428)
ChatGPT 4	0.8235 (0.1082)	-	-	-
Gemini 1.5-Pro	0.285 (0.2366)	1 (0)	0.9825 (0.0362)	1 (0)
Llama 70B	0.32 (0.1776)	0.925 (0.0943)	0.915 (0.0723)	1 (0)

Table 2. Average Accuracy and Standard Deviation Across LNT Tasks

Model	Standard Test	Letter	Number
ChatGPT 3.5 Turbo	0.15 (0.1242)	0.95 (0.0466)	1 (0)
ChatGPT 4	0.8652 (0.0272)	-	-
Gemini 1.5-Pro	0.225 (0.088)	0.985 (0.0298)	1 (0)
Llama 70B	0.77 (0.0763)	-	-

ity to adapt and switch tasks efficiently. By averaging the performance across these eight runs, we aimed to minimize variability and account for any potential anomalies, providing a clearer picture of each model's cognitive flexibility capabilities.

In both the WCST and LNT, the task switches after so many successful guesses in succession. Repeated success indicates that the participant understands how to perform the current task. The number of successes prior to switching the task is a parameter of the test and in our evaluations we use a value of six, matching previous studies of the WCST and LNT on human populations (Dehaene & Changeux, 1991). It is important to note that this value should be considered when interpreting the performance of a participant. Upon initializing the test the current task is unknown to the participant, and in the case where a task switch occurs, the new task is unknown if there are more than two tasks in the test (such as the WCST). This information can be used to calculate a bound on worst case performance of a participant that is able to switch tasks successfully. In the worst case, a participant might need to try all possible tasks before finding the correct one.

Let t be the number of tasks, and n be the number of successful guesses required before a task switch occurs. The worst-case performance P_w is given by:

$$P_w = \frac{n}{n + (t - 1)}$$

For the WCST, with $t = 3$ and $n = 6$:

$$P_w = \frac{6}{6 + (3 - 1)} = 0.75$$

For the LNT, with $t = 2$ and $n = 6$:

$$P_w = \frac{6}{6 + (2 - 1)} \approx 0.857$$

This means that, in the worst case, a participant who is able to switch tasks successfully should achieve an accuracy of at least approximately 75% on the WCST and 86% on the LNT in identifying a correct response. This formulation helps to set a benchmark for interpreting the performance scores of different models on the WCST and LNT.

Performance on the WCST

In the WCST assessment, ChatGPT-4 outperformed the other models in cognitive flexibility, achieving an average test accuracy of 82.35%. ChatGPT-3.5 Turbo, Gemini 1.5-Pro, and Llama 70B exhibited lower overall performance, with ChatGPT-3.5 Turbo performing the worst.

We further investigate ChatGPT-3.5 Turbo, Gemini 1.5-Pro, and Llama 70B by evaluating them on the component tasks of WCST: Shape, Color and Number. Here we see that ChatGPT-3.5 Turbo is not able to successfully perform the component tasks for Shape and Color, but is successful at Number. This indicates that a lack of exhibited cognitive flexibility is not the reason why this LLM fails to succeed on the WCST. However, Gemini 1.5-Pro is able to succeed at the component tasks, indicating a lack of exhibited cognitive flexibility.

Performance on the LNT

In the LNT, ChatGPT-4 again demonstrated superior performance, achieving an average accuracy of 86.52%. ChatGPT-3.5 Turbo showed significantly lower performance, achieving an average accuracy of 15%. Gemini 1.5-Pro and Llama 70B exhibited poor performance, with Llama 70B performing better than Gemini 1.5-Pro.

Due to a failure to perform the LNT successfully, we also evaluated ChatGPT-3.5 Turbo and Gemini 1.5-Pro on the individual tasks of the LNT: Letter and Number. Here, both LLMs succeed on these individual tasks, indicating that the difficulty of the component tasks was not the reason for the poor performance in the overall LNT, but rather the inability to switch tasks flexibly within the same context window.

4. Conclusion

Other Tasks to Assess Cognitive Flexibility

There are numerous neuropsychological tests that could be adapted for evaluating cognitive flexibility in LLMs. The Stroop Color and Word Test (Stroop, 1935), which requires participants to name either the color of the ink of a word or the word itself, tests the ability to inhibit a more automatic response in favor of a less intuitive one. The Trail Making Test (Reitan, 1958), which involves connecting dots in a sequence that alternates between numbers and letters, can also assess the ability to switch between different types of sequences. Adapting these tests for LLMs could provide a broader assessment of their cognitive flexibility.

Parallels with Inhibited Brain Function

Specific brain regions such as the frontal lobe are crucial for cognitive flexibility in humans, and deficits in this area can lead to difficulties in task switching and strategy adaptation. Similarly, limitations in LLMs' cognitive flexibility could be seen as analogous to impaired brain function. Investigating these parallels could provide insights into improving model architectures to better support adaptive, context-sensitive processing.

Differences in LLM Cognitive Flexibility

Comparing the LLMs in our study, we observe significant variation in the degree of exhibited cognitive flexibility. This could be due to several factors, including larger training datasets, more sophisticated model architectures, and enhanced training techniques. In particular, ChatGPT-4 exhibits superior cognitive flexibility compared to other LLMs. However, it is difficult to further investigate the underlying cause for the discrepancies due to the closed nature of commercial LLMs.

Extension to Multi-Modal Models

The insights gained from evaluating and enhancing cognitive flexibility in LLMs can be applied to Vision-Language Models (VLMs) or Multi-modal Large Language Models (MMLLMs). These models, which integrate visual and textual information, face even more complex task-switching scenarios, and could be evaluated on more natural tests (e.g., the WCST as it is depicted in Fig. 1). Training VLMs and MMLLMs on tasks that require simultaneous processing and integration of visual and linguistic data in varied contexts could improve their adaptability and cognitive flexibility.

Improving Cognitive Flexibility in LLMs

Enhancing the cognitive flexibility of LLMs presents a promising avenue for advancing their capabilities across various applications. By fostering adaptability and versatility in these models, we can broaden their utility and effectiveness in addressing complex tasks and scenarios. Several innovative training paradigms can be explored to achieve this goal.

One promising approach is multi-task learning, where models are trained on a diverse set of tasks simultaneously, encouraging them to switch contexts and strategies more fluidly. Another technique involves dynamic task prompts during training, where the context and required responses change frequently, pushing the model to adapt continuously. Furthermore, interactive learning frameworks that incorporate human feedback and guidance can play a crucial role in enhancing cognitive flexibility. By leveraging human input to steer the model's learning process, interactive learning has the potential to enable LLMs to adapt and refine their behavior based on real-time feedback, leading to more agile and responsive performance. Incorporating these training paradigms into the development of LLMs could enhance their cognitive flexibility and empower them to tackle a wide range of tasks with proficiency and adaptability.

Enhancing cognitive flexibility in LLMs is a multifaceted challenge that will require advances in training paradigms, model architectures, and evaluation methods. By drawing parallels with human cognitive functions and incorporating insights from evaluating LLMs on neuropsychological tests, we can develop models that better reflect human adaptability and contextual understanding. Progress in this area will not only improve language processing capabilities but could also extend to multi-modal applications, paving the way for more versatile and intelligent AI systems.

Limitations

This study does not comprehensively investigate how cognitive flexibility scales with the number of learnable parameters of LLMs. While it could constitute resource intensive experiments, evaluating various sizes of LLMs might provide insight into the degree to which cognitive flexibility is influenced by architectural or scale-related factors.

Acknowledgements

The opinions expressed herein are solely those of the authors and do not necessarily represent the opinions of the United

States Government, the U.S. Department of Defense, the Department of the Air Force, or any of their subsidiaries or employees. This research was supported by the U.S. Science, Mathematics, and Research for Transformation (SMART) program, Part of the National Defense Education Program. DISTRIBUTION A: Cleared for public release, distribution unlimited (AFRL-2024-3462).

References

Dehaene, S. and Changeux, J. P. The wisconsin card sorting test: theoretical analysis and modeling in a neuronal network. *Cereb Cortex*, 1(1):62–79, January 1991.

Eslinger, P. J. and Grattan, L. M. Frontal lobe and frontal-striatal substrates for different forms of human cognitive flexibility. *Neuropsychologia*, 31(1):17–28, January 1993.

Grant, D. A. and Berg, E. A. Wisconsin card sorting test. *Journal of Experimental Psychology*, 1948.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Ionescu, T. Exploring the nature of cognitive flexibility. *New Ideas in Psychology*, 30(2):190–200, 2012.

Reitan, R. M. Validity of the trail making test as an indicator of organic brain damage. *Perceptual and Motor Skills*, 8 (3):271–276, 1958. doi: 10.2466/pms.1958.8.3.271.

Rogers, R. D. and Monsell, S. Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology*, 124(2):207–231, June 1993.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., ..., and Wu, Z. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023. URL https://arxiv.org/abs/2206.04615.

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M. S. A., and Becchio, C. Testing theory of mind in large language models and humans. *Nat Hum Behav*, May 2024.

Stroop, J. Studies of interference in serial verbal reactions. *Jounral of Experimental Psychology*, 18:643–662, 1935.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, 2019a. URL https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2019 International Conference on Learning Representations (ICLR)*, 2019b. URL https://openreview.net/forum?id=rJ4km2R5t7.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are Zero-Shot learners. September 2021.

A. Task Description Prompts

For the WCST, the following is given as a prompt to describe the task.

```
You are performing the Wisconsin Card Sorting Task.
In this task, you need to match a card to one of four cards presented to you.
The card you have either matches the number of shapes on each card, the type of shape,
or the color of the shape.
Respond only with the number of a card.
I will provide feedback if your choice was right or wrong.
If your match was not correct, you need to try a different rule for matching your card.
If your match is correct, stick with that rule until the rule changes.
Acknowledge the feedback and keep performing the task until the end of the test.
```

For the LNT, the following is given as a prompt to describe the task.

```
In this test, a sequence consisting of a letter and a number is presented to you.
If you are performing the letter task,
you should respond with ’vowel’ if the letter is a vowel,
and ’consonant’ if the letter is a consonant.
If you are performing the number task,
you should respond with ’even’ if the number is even,
or ’odd’ if the number is odd.
You must pick one task, do not respond with both in mind.
I will provide feedback if your choice was right or wrong.
If you are correct, you correctly identified the task and gave the right answer.
If you are incorrect, it could be because you are not performing the correct task,
or you are not correctly identifying if a number is odd or even,
or a letter is a consonant or a vowel.
Acknowledge the feedback and keep performing the task until the end of the test.
```