# Reproducing a Small-Scale Verbal N-Back Result in Working Memory Capacity of ChatGPT: An Empirical Study

Name: [XIE XINXUAN] Student ID:[1155247025]

## 1 Project Summary and Reproduction Goal

For this reproducibility work, I selected the project associated with the paper Working Memory Capacity of ChatGPT: An Empirical Study(Dongyu Gong et al. 2024). The paper evaluates large language models with verbal and spatial n-back tasks. In the verbal base task, the model is prompted trial by trial and must output `"m"` for a match and `"-"` for a nonmatch. The authors report results for verbal n-back with (n = 1, 2, 3), using `gpt-3.5-turbo` with temperature set to 1.

Drawing on the findings of the original article, this study focuses on a small-scale reproduction of the oral n-back task. The primary claim under investigation is that model performance deteriorates as memory load increases, specifically from the 1-back to the 3-back condition, which represents the central trend reported in the original paper.

## 2 Setup Notes

Building on the authors' repository, I conducted a reproduction of a subset of the oral experiments.

### 2.1 Environment

Python notebook environment (Google Colab / Jupyter-style workflow)

Main libraries: `openai`, `numpy`, `pandas`, `scipy`, `matplotlib`, `tabulate`

### 2.2 Model

In the original paper, the verbal base experiment was conducted using gpt-3.5-turbo with a temperature of 1. In contrast, this reproduction employed deepseek-chat as the underlying model, representing a controlled deviation from the original setup. To maintain reproducibility and methodological transparency, the model name, provider, base URL, and decoding parameters were documented explicitly. Given this change in model, the results obtained here are not directly comparable to those reported in the original study. The model configuration used in this reproduction was: Model: deepseek-chat; Provider: DeepSeek; Base URL: DeepSeek OpenAI-compatible API endpoint; Temperatures tested: 1 and 0.

### 2.3 Task Scope

The original paper implemented the verbal base task across three memory-load conditions (n=1,2,3), with 50 blocks for each condition. The stimulus set was drawn from the alphabet bcdfghjklnpqrstvwxyz, and each block consisted of 24 letters, comprising 8 match trials and 16 non-match trials. To keep the present reproduction computationally manageable, the experimental design was simplified in two ways. First, only the 1-back and 3-back conditions were reproduced. Second, the number of

blocks was reduced to 2 per condition. As a result, the present experiment should be regarded as a small-scale pilot reproduction rather than a full replication of the original paper.

## 3 Reproduction Target and Metrics

The primary objective of this reproduction was to examine whether model performance decreases under higher memory load. To operationalize this question, I compared model performance between the 1-back and 3-back conditions. Performance was evaluated using four metrics: hit rate, false alarm rate, accuracy, and d′ (d-prime). Among these, accuracy served as the principal metric for assessing the reproduction target, whereas hit rate, false alarm rate, and d′ were used as supplementary measures to help interpret the underlying sources of performance variation.

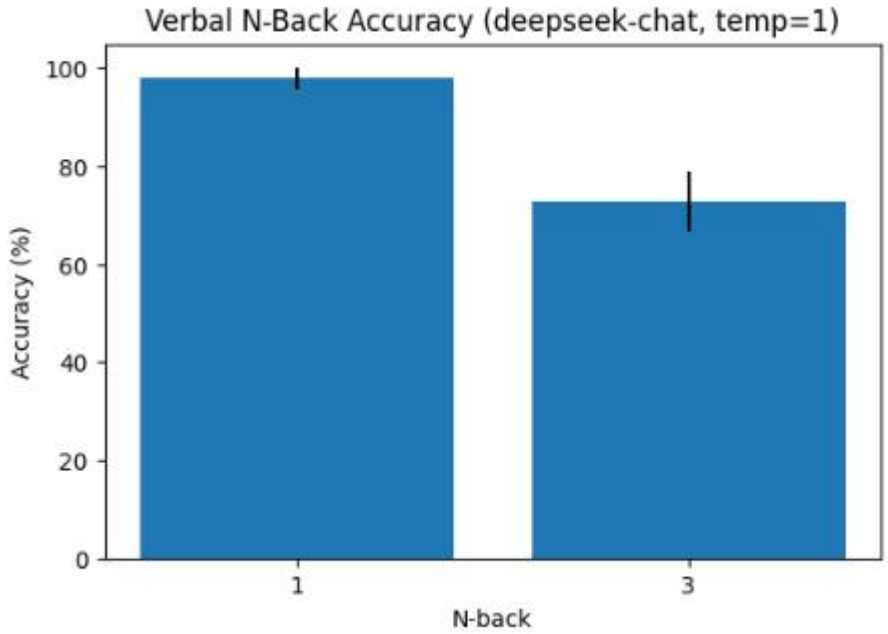
## 4 Reproduction Results

### 4.1 Temperature = 1

For the first run, I used `deepseek-chat` with `temperature = 1`.

| N-back | Hit Rate (%) | False Alarm Rate (%) | Accuracy (%) | D Prime |
|---|---|---|---|---|
| 1-back | 93.75 ± 6.25 | 0.00 ± 0.00 | 97.92 ± 2.08 | 4.06 ± 0.59 |
| 3-back | 31.25 ± 18.75 | 6.25 ± 0.00 | 72.92 ± 6.25 | 0.96 ± 0.58 |

| | N-back | Hit Rate (%) | False Alarm Rate (%) | Accuracy (%) | D Prime |
|---|---|---|---|---|---|
| 0 | 1-back | 93.75 ± 6.25 | 0.00 ± 0.00 | 97.92 ± 2.08 | 4.06 ± 0.59 |
| 1 | 3-back | 31.25 ± 18.75 | 6.25 ± 0.00 | 72.92 ± 6.25 | 0.96 ± 0.58 |



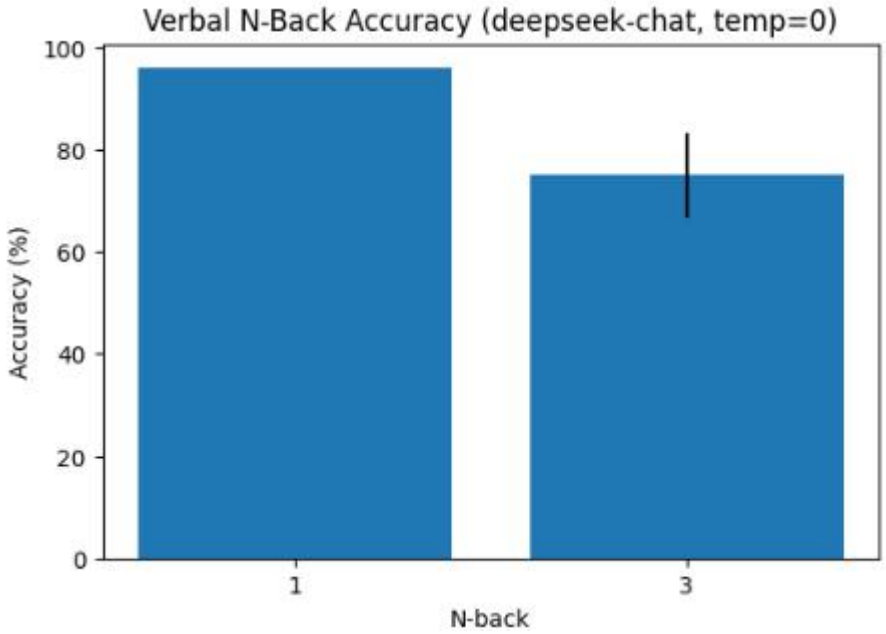
The results reveal the expected trend, with the model performing substantially better in the 1-back condition than in the 3-back condition. This pattern aligns with the main finding of the original paper, which suggests that model performance deteriorates as working-memory demands increase.

**4.2 Temperature = 0**

For the modified run, I changed only one variable: the temperature.

| N-back | Hit Rate (%) | False Alarm Rate (%) | Accuracy (%) | D Prime |
|---|---|---|---|---|
| 1-back | 87.50 ± 0.00 | 0.00 ± 0.00 | 95.83 ± 0.00 | 3.48 ± 0.00 |
| 3-back | 25.00 ± 25.00 | 0.00 ± 0.00 | 75.00 ± 8.33 | 1.16 ± 1.16 |

| | N-back | Hit Rate (%) | False Alarm Rate (%) | Accuracy (%) | D Prime |
|---|---|---|---|---|---|
| 0 | 1-back | 87.50 ± 0.00 | 0.00 ± 0.00 | 95.83 ± 0.00 | 3.48 ± 0.00 |
| 1 | 3-back | 25.00 ± 25.00 | 0.00 ± 0.00 | 75.00 ± 8.33 | 1.16 ± 1.16 |




The second run provides additional support for the same qualitative trend, as accuracy in the 1-back condition remained consistently higher than in the 3-back condition.

**4.3 Interpretation**

Across both runs, the reproduced pattern was stable and qualitatively consistent. Performance in the 1-back condition remained higher than in the 3-back condition, supporting the view that increasing memory load leads to reduced model performance. The performance decline was mainly reflected in a lower hit rate, indicating that the model was more likely to miss true matches in the 3-back condition. Moreover, the substantial decrease in d′ from 1-back to 3-back suggests weaker discrimination under higher memory load. Overall, these findings provide evidence that the core behavioral trend reported in the original paper was successfully reproduced in this small-scale pilot study.

## 5. Modification and Results After Modification

### 5.1 Controlled Parameter Modification

To satisfy the requirement of introducing one small but meaningful modification, I implemented a single controlled parameter change. Specifically, the temperature setting was reduced from 1 to 0, while all other components of the experiment were kept unchanged. These included the verbal n-back task setup, the 1-back and 3-back conditions, the evaluation pipeline, the DeepSeek model, and the output format ("m" or "-"). This design ensured that any observed differences in performance could be more plausibly attributed to the change in decoding temperature.

### 5.2 Comparison Table

| N-back | Temp=1 Accuracy | Temp=0 Accuracy | Accuracy Delta | Temp=1 D′ | Temp=0 D′ |
|---|---|---|---|---|---|
| 1-back | 95.83 | 95.83 | 0 | 3.48 | 3.48 |
| 3-back | 75 | 75 | 0 | 1.36 | 1.16 |

| | N-back | Temp=1 Accuracy | Temp=0 Accuracy | Accuracy Delta | Temp=1 D′ | Temp=0 D′ |
|---|---|---|---|---|---|---|
| 0 | 1-back | 95.83 | 95.83 | 0.0 | 3.48 | 3.48 |
| 1 | 3-back | 75.00 | 75.00 | 0.0 | 1.36 | 1.16 |

The parameter change from temperature 1 to 0 did not affect accuracy, as performance remained identical in both the 1-back and 3-back conditions. The qualitative trend of better performance under lower memory load was therefore preserved. A modest change was observed in d′ for the 3-back condition, which decreased from 1.36 to 1.16, whereas the 1-back condition remained unchanged. This suggests that the temperature manipulation had little effect on overall accuracy, but may have slightly reduced discriminative performance under higher memory load.

### 5.3 Interpretation of Modification

In this pilot experiment, reducing the temperature from 1 to 0 did not produce any observable change in accuracy. Several factors may account for this result. First, the output space was highly constrained, as the model was required to choose only between "m" and "-". Second, the prompt format was highly explicit, which may have made the task largely deterministic regardless of decoding temperature. Third, the sample size was limited to only 2 blocks per condition, reducing the likelihood that subtle performance differences would be reflected clearly in the accuracy metric. Taken together, these observations suggest that the modification was measurable in principle, but its effect on accuracy was effectively negligible in this small-scale pilot reproduction.

## 6 Debug Diary

Several practical challenges were encountered during implementation. The first concerned API compatibility: the original notebook was written using an older OpenAI-style interface and therefore had to be adapted to work with the OpenAI-compatible DeepSeek API. The second involved authentication, as initial runs failed due to invalid or placeholder API key values; this problem was resolved by supplying a valid DeepSeek API key and properly initializing the client. A third issue

stemmed from notebook state management, since some errors occurred when the API client variable was not defined in the active runtime environment; re-executing the setup cell corrected this problem. In addition, because the full experiment described in the original paper was beyond the available time and computational budget, the reproduction scope was reduced to the 1-back and 3-back conditions with 2 blocks each. This simplification made the experiment feasible while preserving the core reproduction objective.

## 7 Conclusion

This report presents a reproduction of a small but meaningful subset of the paper Working Memory Capacity of ChatGPT: An Empirical Study. The primary trend examined in this study was successfully reproduced: performance on the verbal n-back task declined as task difficulty increased, with performance in the 3-back condition consistently lower than that in the 1-back condition. This result is consistent with the paper's broader conclusion that working-memory-related performance deteriorates as memory load increases.

As the required experimental modification, the temperature parameter was changed from 1 to 0. In this pilot reproduction, however, this manipulation produced no observable difference in accuracy. This suggests that decoding temperature may have limited influence in a tightly constrained binary-response task of this kind, at least under a small-scale experimental setting.

At the same time, several limitations should be acknowledged. First, this reproduction used DeepSeek rather than the original gpt-3.5-turbo, and results obtained with a different underlying language model are not directly comparable to those reported in the original paper. Second, only a small-scale subset of the original experimental design was implemented. Accordingly, the findings reported here should be interpreted as evidence of trend-level reproducibility rather than as an exact numerical replication of the original results.

Overall, the authors' repository proved usable after modest technical adjustments, the core verbal n-back trend was successfully reproduced, and the temperature modification had little measurable effect in this pilot setting.