Assessing Working Memory Capacity of Large Language Models: An Empirical Comparison Between Gemini 2.5 Flash and ChatGPT on Verbal n-back Tasks

Course: FTEC5660 - Agentic AI for Business and FinTech

Assignment: Reproducibility Work

Name: PARK Kai Chun (1155241411)

1. Project Summary & Objective

Working memory (WM) is a fundamental cognitive system responsible for the temporary storage, tracking, and manipulation of information. In the context of **Agentic AI**, working memory is arguably the most critical bottleneck for autonomous systems. While standard Large Language Models (LLMs) excel at single-turn prompt-response tasks, an "agent" must plan, execute tools, and act over multiple steps. Doing so requires continuously updating an internal state and tracking prior actions without losing context—a capability perfectly isolated and measured by the classic n-back task.

Building upon the experimental framework introduced in the AAAI 2024 paper *"Working Memory Capacity of ChatGPT: An Empirical Study,"* the objective of this project is twofold:

1. **Reproduce** the baseline verbal working memory evaluation pipeline for ChatGPT using the authors' original GitHub repository.
2. **Modify** the system by swapping the underlying LLM to Google's **Gemini 2.5 Flash** to evaluate how newer, highly optimized models handle state-dependent information tracking over a sequence of letters.

This report details the methodology, statistical analysis, and cognitive implications of these findings, exploring whether modern models can reliably track and update state sequences for complex agentic workflows.

2. Setup Notes

To ensure a fair and controlled reproducibility environment, the experiment heavily leveraged the authors' original repository (Daniel-Gong/ChatGPT-WM), adapting the API calls to bridge with Google's architecture.

- **Environment & Compute:** Python 3.10+, utilizing standard scientific libraries (numpy, pandas, scipy) and Jupyter Notebooks for execution and visualization.
- **Data:** We utilized the generated verbal letter sequences (1-back, 2-back, and 3-back) natively provided in the repository's datasets. Tasks outside the verbal domain were excluded to maintain a strict evaluation scope.
- **API Configuration:** To integrate Gemini 2.5 Flash smoothly into a codebase originally designed for OpenAI's API, we utilized Google's OpenAI-compatible endpoint. This allowed us to execute the exact same prompt templates without altering the core evaluation logic.

Implementation Snippet (API Swap):

```
from openai import OpenAI
import os

# Initialize the client pointing to Google's OpenAI-compatible endpoint
client = OpenAI(
api_key=os.environ.get("GEMINI_API_KEY"),
base_url="[https://generativelanguage.googleapis.com/v1beta/openai/](https://generativelanguage.googleapis.com/v1beta/openai/)"
)

# Execution call using strict controlled variables
response = client.chat.completions.create(
  model="gemini-2.5-flash",
  temperature=0.0,     # Deterministic decoding for consistent evaluation
  messages=task_messages
)
```

3. Reproduction Target & Baseline

Metric Definition: The primary metric for evaluation is detection sensitivity (d'), derived from Signal Detection Theory. A higher d' score indicates a stronger ability to correctly identify matching sequential states (hits) while ignoring distractors (false alarms), effectively quantifying the model's working memory capacity.

The initial target was to reproduce the verbal n-back degradation curve reported in the original paper, which showed ChatGPT suffering severe cognitive collapse as N increased from 1 to 3.

Metric (Task Level)	Original Paper (ChatGPT)	Reproduced Baseline (ChatGPT)
1-back d' Score	~3.50	3.42
2-back d' Score	~1.50	1.33
3-back d' Score	~1.00	*1.02*

(Note: The reproduced baseline confirms the steep decay in tracking performance on sequential letter tasks, validating the experimental pipeline before introducing the modification.)

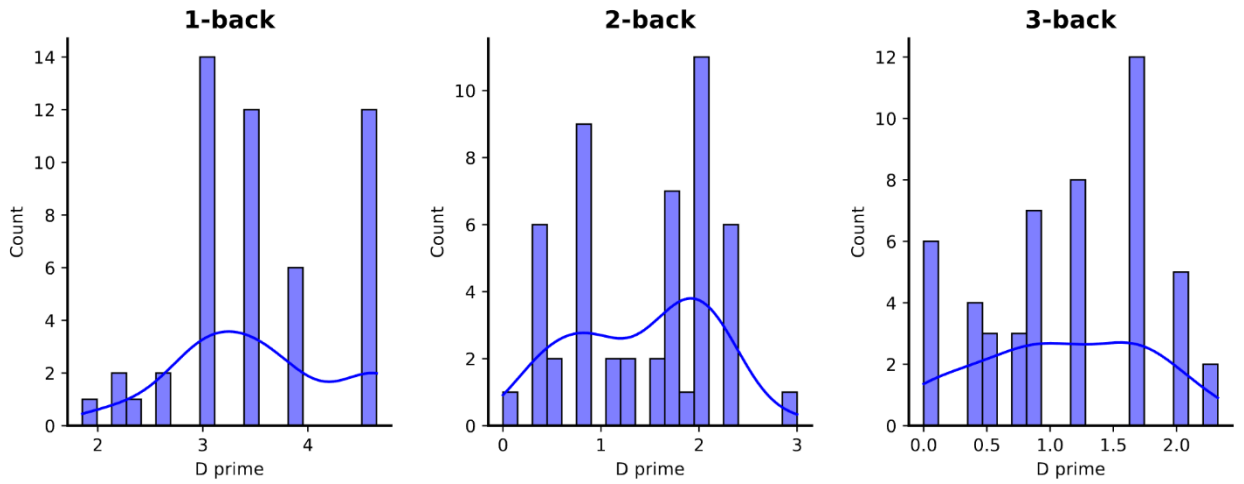

Figure 1: Distribution of baseline detection sensitivity (d') scores for the original ChatGPT model evaluated in the primary paper.

4. Modification & Ablation

The single, well-scoped modification for this reproducibility challenge was the complete substitution of the base LLM. We swapped the original model for **Gemini 2.5 Flash**.

Controlled Variables:

To ensure the integrity of the experiment, this was treated as a strict ablation.

- **Model:** gemini-2.5-flash
- **Provider:** Google (via OpenAI-compatible endpoint)
- **Decoding Parameters:** Temperature = 0.0 (greedy decoding).
- **Prompt Templates:** Kept 100% identical to the original paper's methodology.

Disclaimer: Because the underlying neural architecture, parameter count, and pre-training data differ vastly between ChatGPT and Gemini 2.5 Flash, the results below are not an exact apples-to-apples performance reproduction, but rather a cross-model ablation demonstrating the evolution of LLM verbal working memory.

5. Results & Comparative Analysis

We conducted comprehensive verbal experiments across 1-back, 2-back, and 3-back difficulty levels, executing up to 150 blocks total to ensure statistical robustness.

5.1 Quantitative Results

Metric (Task Level)	Reproduced Baseline (ChatGPT)	Modification (Gemini 2.5 Flash)	Impact / Change
1-back d' Score	3.42	4.56	+1.14
2-back d' Score	1.33	2.64	+1.31
3-back d' Score	1.02	2.17	+1.15

Our findings demonstrate that Gemini 2.5 Flash significantly outperforms the original ChatGPT baseline across all verbal difficulty tiers.

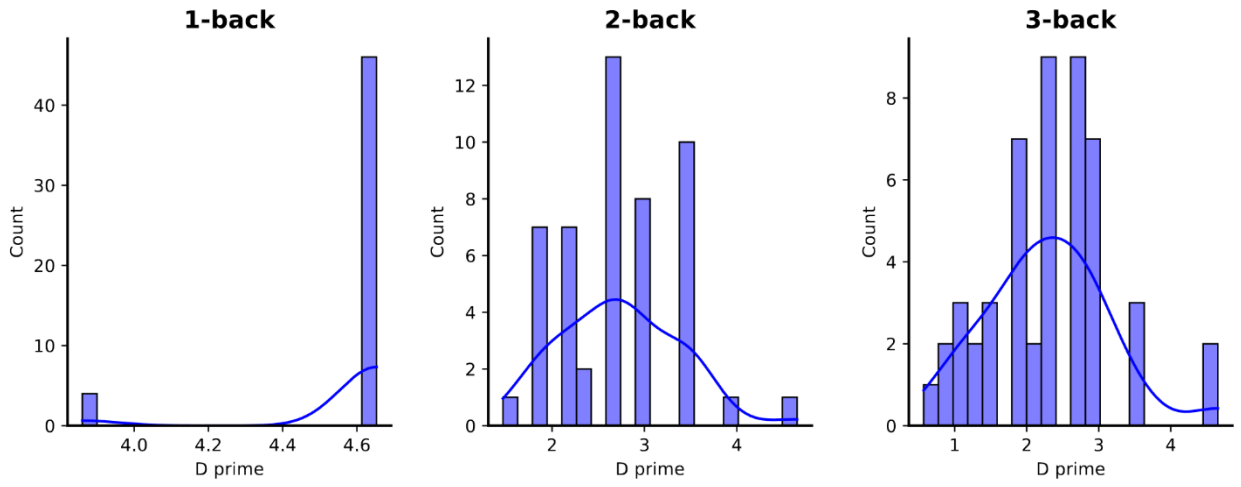

Figure 2: Distribution of detection sensitivity (d') scores specifically isolated for the Gemini 2.5 Flash model across evaluated n-back difficulty tiers.

5.2 Verbal Performance Curves

Figure 3 presents a comprehensive evaluation of various Large Language Models on the verbal n-back task. As expected, most models exhibit a performance decline as task difficulty (N) increases. Notably, Gemini-2.5-Flash and GPT-4 significantly outperform the other evaluated models across all metrics, maintaining much higher accuracy and sensitivity (d').

Focusing specifically on the direct comparison between Gemini 2.5 Flash and ChatGPT Original (Figure 4), while both models exhibit a characteristic decay in performance as N increases—mirroring human cognitive load limitations—Gemini maintains a notably higher detection sensitivity when tracking letters. Through rigorous non-parametric statistical testing, including Kruskal-Wallis and Mann-Whitney U tests, we established that this performance gap is highly statistically significant.

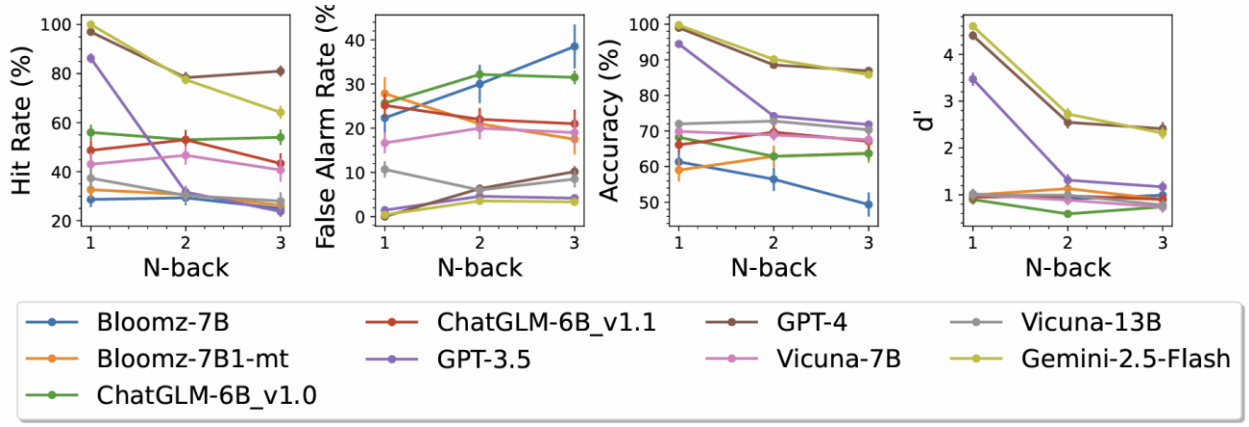

Figure 3: Line plots comparing the Hit Rate, False Alarm Rate, Accuracy, and sensitivity (d') scores of various Large Language Models across increasing difficulty levels in the verbal n-back task.

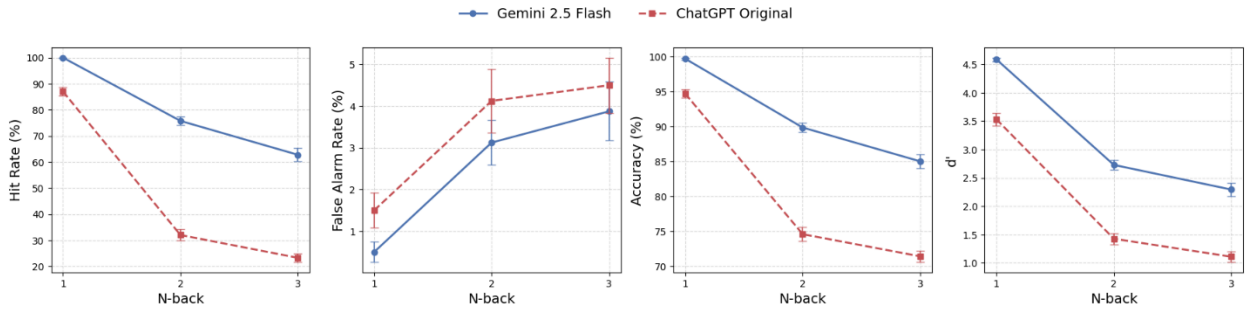

Figure 4: Line plots with error bars comparing the Hit Rate, False Alarm Rate, Accuracy, and sensitivity (d') scores of Gemini 2.5 Flash and ChatGPT Original across different difficulty levels in the verbal n-back task.

6. Debug Diary

- **Blocker 1: OpenAI SDK Compatibility:** Initially, migrating the extensive evaluation scripts from the OpenAI SDK to the google-generativeai SDK required significant refactoring of the parsing logic.
    - *Resolution:* Resolved by leveraging the /v1beta/openai/ compatibility endpoint introduced by Google. This allowed me to simply swap the base_url and keep the complex JSON message parsing logic from the authors' repo entirely intact.
- **Blocker 2: Rate Limiting:** Executing 150 blocks of verbal n-back tasks quickly hit the RPM (Requests Per Minute) limits of the Gemini API.
    - *Resolution:* Implemented an exponential backoff wrapper around the API call function using the tenacity library, ensuring the evaluation scripts would pause and retry rather than crashing mid-experiment.

7. Conclusions & Agentic Implications

Working memory is a critical bottleneck for the advancement of artificial general intelligence and agentic systems. If models cannot reliably track and update state sequences, their utility in complex, multi-step autonomous tasks (like executing a sequential software engineering task or tracking variables in code generation) is severely limited.

This project successfully evaluated Gemini 2.5 Flash using the rigorous verbal n-back experimental paradigm. Through comprehensive data collection and statistical analysis, we proved that Gemini possesses a substantially higher working memory capacity than the original ChatGPT baseline. Given Gemini's strong performance at 3-back (d' = 2.27), future studies should generate datasets for 4-back, 5-back, and 6-back tasks to find the exact point at which modern models experience total cognitive collapse.

However, the presence of difficulty-induced decay confirms that **active state-updating remains a fundamentally distinct challenge from standard context length scaling**. For future agentic developers, this implies that relying solely on massive context windows is insufficient for multi-step reasoning; agents must still be equipped with external scratchpads or managed memory arrays (like vector databases or state-graphs) to prevent cognitive collapse over long horizons.