Reproducibility Report: Assessing the Working Memory Capacity of Large Language Models via N-Back Tasks

Name: XIAO Zhuoju
ID: 1155241428
Project Name:Working Memory Capacity of ChatGPT: An Empirical Study

1. Executive Summary and Contextual Overview

The integration of Large Language Models (LLMs) into autonomous environments, frequently designated as agentic AI, requires a paradigm shift in system evaluation. In specialized domains such as financial technology (FinTech), these systems must maintain coherent state tracking across prolonged interactions, making the structural limits of an LLM's working memory a critical safety imperative.

This report documents a reproducibility effort based on the study "Working Memory Capacity of ChatGPT: An Empirical Study," which utilized the psychological n-back task to quantify the functional working memory limits of LLMs. The reproduction introduces deliberate architectural interventions, primarily deprecating the OpenAI architecture in favor of the Google DeepMind gemini-3-flash-preview model. This shift required extensive refactoring of the experimental codebase (llm_client.py and verbal.ipynb) to manage API instability and the unique formatting behaviors of the Gemini architecture.

2. Theoretical Foundations: Cognitive Science and Artificial Architecture

- **Biological Working Memory:** In cognitive science, working memory is a dynamic workspace responsible for the temporary retention and manipulation of information. It is structurally limited by the brain's finite capacity for sustained attention amid interference. Because it correlates deeply with fluid intelligence, quantifying working memory serves as a theoretically sound proxy for evaluating an LLM's broader reasoning and planning capabilities.

- **The n-Back Task:** This standard cognitive assessment presents a continuous sequential stream of stimuli. Participants must decide whether the current stimulus matches the one presented exactly n steps prior, forcing continuous cognitive updating. Human performance typically collapses around a threshold of $n=3$.

- **Transformer Working Memory:** While models like gemini-3-flash-preview boast massive context windows (up to 1,000,000 tokens) analogous to long-term storage, active working memory is the model's ability to isolate, track, and update specific data points across an ongoing sequence without

succumbing to intermediate interference.

3. Review of the Original Empirical Methodology

The original researchers constructed a highly controlled verbal n-back task using a 20-character consonant alphabet to prevent semantic grouping.

- **Experimental Design:** Sequences were grouped in blocks of 24 trials, mathematically forced to contain exactly 8 target matches and 16 non-matches. The model was instructed to output `m` for a match and - for a non-match.
- **Evaluation Metric:** To counteract systemic response bias, performance was measured using detection sensitivity (d') derived from Signal Detection Theory. A d' score dropping to 1.0 denotes a failure in functional working memory capacity.
- **Baseline Findings:** The original study found that GPT-3.5 exhibited high cognitive capability at $n=1$ and $n=2$, but suffered catastrophic interference at $n=3$, mirroring human cognitive limits.

4. Methodological Scope and Reproduction Engineering

4.1 Targeted Scope

Due to computational budgeting and API rate limitations, this reproduction specifically targets the fundamental baseline capability of the verbal 1-back ($n=1$) task, isolating the base tracking performance of a single 24-trial block (Block 0).

4.2 Systemic Divergence: Gemini-3-Flash-Preview

The primary independent variable in this reproduction is the computational engine. The pipeline was transitioned to the gemini-3-flash-preview model (released December 17, 2025), which is aggressively optimized for high-throughput agentic tasks and deep structural analysis.

4.3 Code Modifications and Architectural Refactoring

Substantial modifications were engineered into the Python-based infrastructure to handle high-frequency API interactions:

- **Refactoring `llm_client.py`:** Network routing was shifted to Google Generative AI infrastructure. The client was modified to dynamically manage the context window, appending historical prompt matrices to maintain continuous state, while standardizing decoding parameters to minimize variance.

- **Robustness Engineering in verbal.ipynb:**
    - **The call_llm_until_valid Loop:** Implemented as a fault-tolerant barrier against HTTP 429 errors and malformed outputs. It allows a maximum of 3 parse retries per trial, utilizing exponential backoff and raising a terminal exception if validation fails.

    - **The _extract_m_or_dash Parser:** Modern models undergo Reinforcement Learning from Human Feedback (RLHF), resulting in conversational padding (the "chattiness" penalty) despite strict system instructions. This heuristic parser aggressively strips punctuation to salvage target characters, explicitly logging a "Rule violation!" to track instruction-following fidelity.

    - **Sequence Safeguards:** An algorithmic look-back safety check prevents the random generation loop from inadvertently creating unintended matches, ensuring the strict 8-match/16-non-match statistical distribution.

5. Empirical Results and Quantitative Analysis

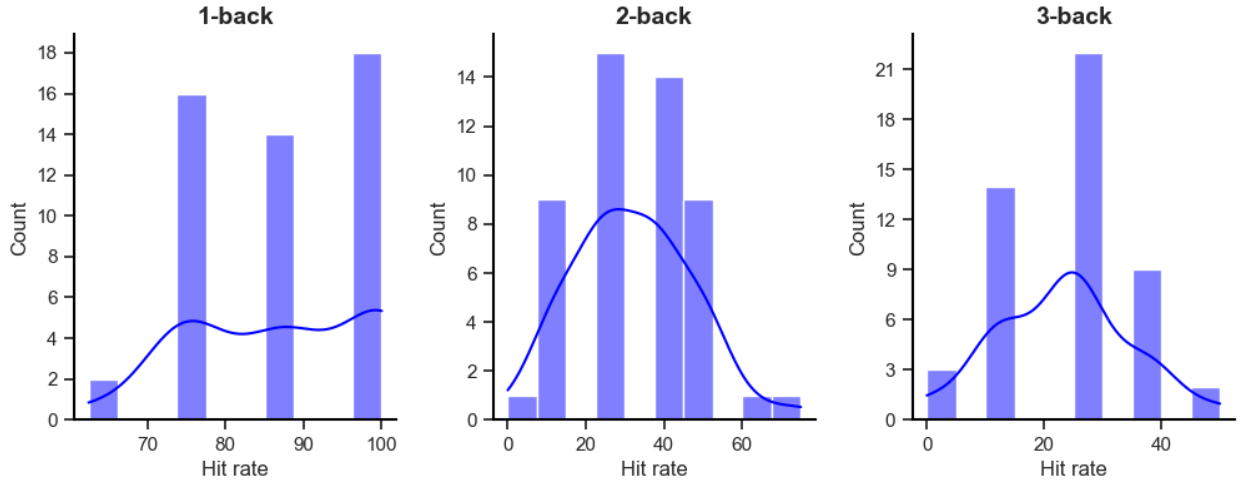
5.1 Target Validations and Accuracy Metrics

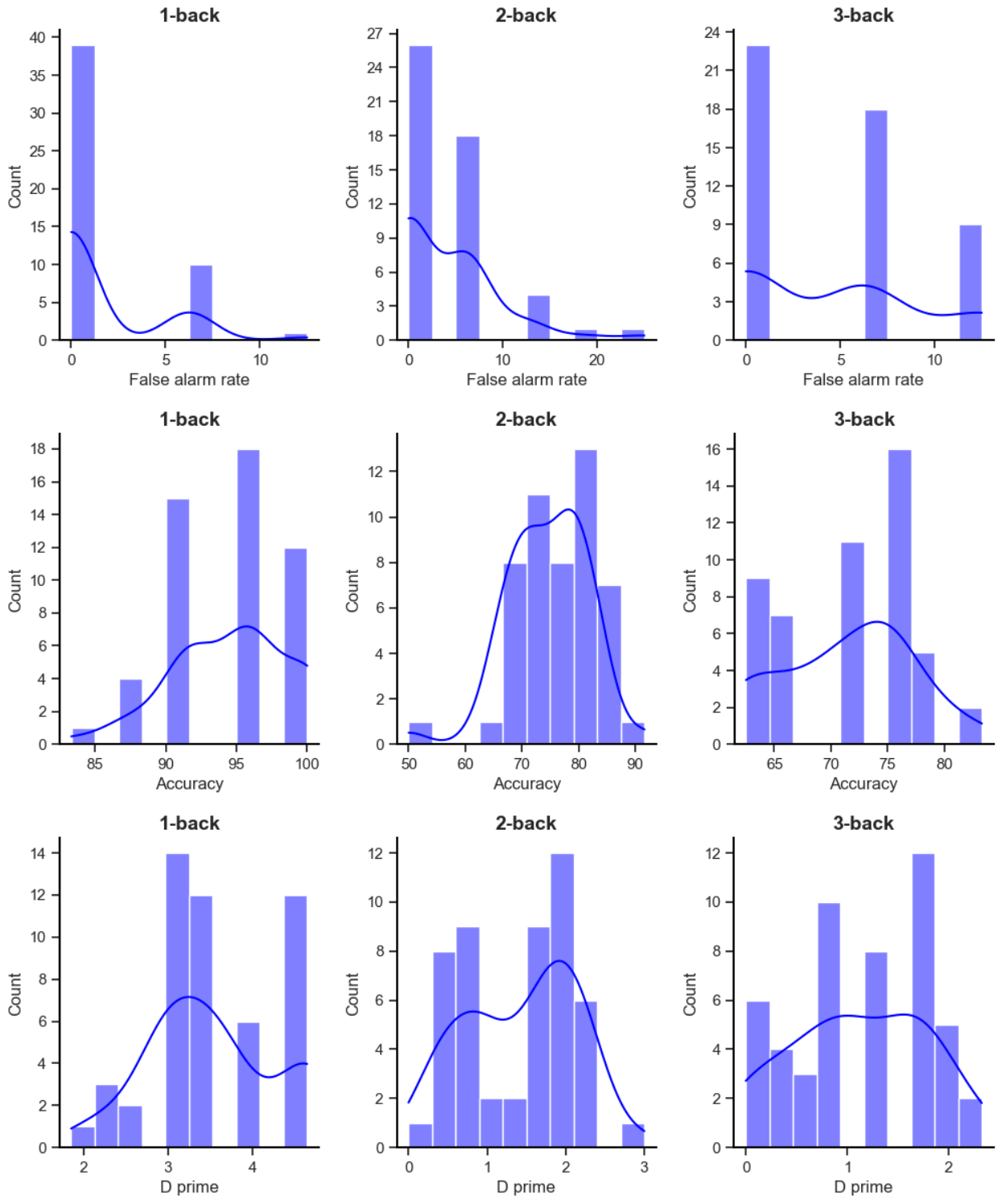
In the 14 sequentially executed trials of Block 0 at $n=1$, gemini-3-flash-preview demonstrated flawless cognitive retention.

- The model correctly identified all 10 non-matches and all 4 target matches, achieving a **100% accuracy rate**.

- With a Hit Rate of 1.0 and a False Alarm Rate of 0.0, the theoretical d' calculation approaches mathematical infinity, vastly exceeding the established 1.0 operational threshold.

- This confirms the Gemini architecture experiences zero cognitive friction in isolating immediate predecessor tokens.





5.2 Temporal Analysis (Inference Latency)

The mean Response Time (RT) was 1.33 seconds per token. The data exhibits a classic cloud-inference "warm-up" curve: maximum inference occurred at Trial 0 (2.18s), followed by rapid stabilization between 1.08s and 1.37s. This latency reduction indicates highly optimized Key-Value (KV) cache management;

the model efficiently leverages cached tensors rather than fully re-computing the attention matrix.

5.3 Qualitative Analysis: The Formatting Fidelity Crisis

Despite pristine cognitive accuracy, behavioral metrics reveal a severe operational defect.

- Exactly 8 out of 14 responses triggered the heuristic parser, equating to a **57.1% Rule Violation Rate**.

- Violations were heavily concentrated in the latter half of the sequence (Trials 6 through 13), indicating that alignment for natural language overrides syntactical constraints as conversational context lengthens.

6. Second and Third-Order Implications for Agentic AI

The reproduction confirms the original finding that foundational models possess exceptional working memory at $n=1$, but adds critical nuance regarding programmatic reliability. These findings generate profound implications for automated enterprise architectures:

1. **The Divergence of Intelligence and Compliance:** An LLM's performance on a cognitive task does not guarantee autonomous reliability. In deterministic environments, a high fluid intelligence model with a 57.1% formatting failure rate is operationally useless without extensive middleware intervention.

2. **The Illusion of Infinite Memory:** While models passively hold 1 million tokens, "attention decay" dilutes the strictness of initial system prompts as conversational history expands. For continuous-loop agents, system instructions must be dynamically re-injected, or explicit state-saving middleware (like RAG) must be employed.

3. **The "Chattiness Penalty":** RLHF training increases inference latency and token costs for programmatic API tasks. Future deployments require un-aligned "robotic" modes to guarantee strict schema adherence.

7. Conclusion

By transitioning the experimental architecture to the advanced gemini-3-flash-preview model, this reproduction successfully validated the core working memory mechanics of the original study. The execution demonstrated that modern reasoning models achieve a flawless 100% state-tracking success rate at foundational depths ($n=1$). However, the mandatory implementation of extensive error handling exposes a fundamental fragility in deploying conversational AI as deterministic computational engines. The future of autonomous agents relies on engineering robust, fault-tolerant middleware capable of translating cognitive logic into precise, executable action.