

False Memories in LLMs: Gist-Based Recognition Despite Verbatim Context

Robin Orthey and Guillaume Pelat and Yuri Nakao

Fujitsu Research

{orthey.robin, pelat.guillaume, nakao.yuri} @fujitsu.com

Abstract

Large language models (LLMs) operate over information held in their context window, but it is unclear whether they use that information by referencing it literally or by relying on a more general semantic gist. We address this question by transferring the Deese-Roediger-McDermott (DRM) false-memory paradigm, a standard method for eliciting gist-based memory errors in humans, to four LLMs spanning two architectures and two parameter scales. We test recognition of studied words, unrelated words, and critical lures across context lengths from 1k to 32k tokens, and measure responses both as discrete judgments and as a continuous probability that captures effects below the threshold of an overt response. We find that a gist-consistent false-memory signal is present in every model, even where overt false recognitions are rare, but that its expression as overt false memory is model-specific. Where false memories occur, they tend to increase with context length, and they do so even though the studied words remain available verbatim in the context, indicating that recognition is driven by gist rather than verification against the literal input. The DRM paradigm thus offers a controlled, psychologically grounded probe of gist-driven errors in LLMs.

1 Introduction

Large Language Models (LLMs) have shown great promise in solving complex, context-dependent tasks. A key enabler is their ability to hold and operate on information within their context window. As context windows grow, it remains unclear how this expanding body of information is actually used: does the model reference it verbatim, or does it rely on a more general gist carried forward? This is, at heart, a question about whether the model engages the literal input or a looser impression of it. In this paper, we examine it by transferring a well-established memory research paradigm from psychology to LLMs.

1.1 Human memory

Human memory is reconstructive rather than a faithful recording (Bartlett, 1995; Loftus and Palmer, 1974; Schacter, 1999), as demonstrated by the Deese-Roediger-McDermott (DRM) paradigm (Roediger and McDermott, 1995). Essentially, the DRM assesses the human ability to memorize and recognize words from studied word lists. What makes the DRM paradigm special is how the studied word lists are constructed. Each word list contains a series of words that are all semantically related to another word that itself is not part of the list, known as the lure word. For example the words: “bed”, “pillow”, “night”, could form a list around the lure word “sleep”. During the recognition test examinees are confronted with old (studied) words, new words, and lure words. The ability to discriminate between old and new words represents standard memory performance, but notably many examinees report recognizing the lure words, which are generally referred to as “false memories”.

False memories in humans can be explained by fuzzy-trace theory (FTT; Reyna and Brainerd, 1995; Brainerd and Reyna, 2002). According to FTT, a memory comprises a “verbatim trace”, an item-specific representation of the event, and a “gist trace”, its semantic content. In a recognition task, both traces support endorsement. Novel words match neither trace and are therefore rejected, studied words match both and are recognized, and lure words share the gist but lack a verbatim trace, so they may nonetheless be falsely recognized. FTT further holds that verbatim traces decay over time while gist traces persist, so that false memories tend to strengthen as delay increases. Thus, the DRM paradigm can distinguish verbatim from gist traces in human research subjects and may be well suited to study how LLMs use their context window as it can quantify the influence of gist alone.

1.2 Transfer to LLMs

Behavioural similarities exist between human memory and LLM’s context window, despite structural divergence. In humans, the information learned during encoding passes through the working memory to the long-term memory and is retrieved again during the recognition test. In contrast, LLMs maintain the literal word lists in their context window, theoretically giving them perfect access to the requested information. Despite that, a preprint by [Cao et al. \(2025\)](#) reported that GPT-4 experienced false memories in a DRM task, while other 8b open LLMs did not.

According to FTT, memory decays with time, and verbatim traces do so faster than gist traces ([Reyna and Brainerd, 1995](#)). LLMs have no concept of time, but we propose that context length serves as an operational analogue for the retention interval. Context window length and time differ conceptually, but there is evidence to suggest that they have similar effects on memory performance. For example, the “Lost in the middle” phenomenon ([Liu et al., 2024](#); [Salvatore et al., 2025](#)) describes the observation that LLM performance in long context situations tends to be best when the relevant information is located near the beginning or the end of the context window, a finding known in psychological memory research as “serial effect” or “primacy and recency effect” ([Deese and Kaufman, 1957](#); [Murdock Jr, 1962](#)). But there is more evidence linking LLM context window length to performance. Evidence on RAG retrieval suggests that as context windows reach certain lengths, well within their theoretical possible lengths, performance can decline ([Cuconasu et al., 2024](#); [Jin et al., 2024](#)). [Du et al. \(2025\)](#) advanced this finding by demonstrating that even under perfect retrieval, performance declines for long context windows. We are hesitant to draw a direct analogy to psychology, as this effect could be explained by multiple factors. Taken together, while there are structural differences between human memory and LLMs, enough behavioral similarities exist to warrant examining LLM capabilities with psychological research tools.

1.3 This work

We adapt the classical DRM paradigm for use with four models spanning two families: Llama 3.1 ([Grattafiori et al., 2024](#)) and Qwen 2.5 ([Yang et al., 2024](#)), and two parameter scales (approximately 7b

and 70b). To examine how context length affects false memory, we manipulate the length of a buffer task placed between the studied lists and the recognition test, with lengths ranging from 1k to 32k tokens. For each test word we record both the discrete recognition response and a continuous measure, the renormalized probability of responding “yes”, which captures graded recognition strength below the threshold of an overt response. Familiar lures are evaluated against an unfamiliar-word baseline and an unfamiliar-lure control, allowing us to separate gist-driven false recognition from a general tendency to respond “yes”.

To our knowledge, the DRM paradigm has been applied to LLMs only once before, leaving its central questions largely open. We address three of them. Our contributions are as follows:

- We adapt the DRM false-memory paradigm to LLMs and introduce a continuous, logit-derived recognition measure (the renormalized probability of “yes”) that reveals gist-driven false-memory signals invisible to discrete yes/no responses.
- We separate gist-driven false recognition from a general response bias by evaluating familiar lures against an unfamiliar-word baseline and an unfamiliar-lure control.
- Across four models spanning two families and two scales, we find that a gist-consistent false-memory signal is present in every model, whereas its expression as overt false recognition, and its tendency to grow with context length, are model-specific and not predicted by parameter count.

2 Method

2.1 The DRM Paradigm (classic psychology)

The original DRM paradigm procedure ([Roediger and McDermott, 1995](#)) consists of an encoding phase and a recognition phase. During the encoding phase, the human research subject is presented with the study material consisting of 16 word lists, with 15 words each. Each word is presented orally with a 1.5-second delay between words. After a list is presented, participants either perform a free recall task or an unrelated buffer task, e.g., solving math problems, before advancing to the next list. After the encoding phase is completed and a 5-minute break, the recognition phase begins. Participants are presented with a test set of words. For

each word, participants must indicate if the word is OLD, i.e., it was presented during the encoding phase, or NEW, i.e., it was not presented during the encoding phase.

An important aspect of the DRM paradigm is the construction of the word lists. [Roediger and McDermott \(1995\)](#) used 24 lists of 15 words each. The lists were constructed in a way so that all 15 words in a list are associated with another semantically related word that does not appear on the list itself, referred to as the LURE. For example, the following list: “mad; fear; hate; rage; temper; fury; ire; wrath; happy; fight; hatred; mean; calm; emotion; enrage” is associated with the word: anger (LURE word). Furthermore, words appearing earlier in the list are semantically more related to the LURE word than those towards the end of the list.

The test set contained 96 words drawn from the 24 DRM lists, 16 studied during encoding and 8 novel lists. It contained 48 OLD words, three from each of the 16 studied word lists, always choosing the first, eighths and tenths word. The remaining 48 words were novel. 24 words were drawn from the 8 novel word lists, again choosing the first, eighths, and tenths word from the original lists. Lastly, we included the 24 LURE words associated with the 16 studied and 8 novel word lists. Thus, the final test set contained 96 words: 48 OLD words, 24 NEW words, 16 OLD LURE words, and 8 NEW LURE words.

The key finding that [Roediger and McDermott \(1995\)](#) demonstrate is that human subjects report remembering familiar LURE words with similar recognition accuracy to familiar OLD words. The false “recognition” of familiar LURE words is seen as the existence of false memories and the ease of use and robust replication of this finding has established the DRM paradigm as one of the major false memory research methodologies.

2.2 Adaptation to LLMs

Our objective was to implement the DRM paradigm as faithfully as possible in LLMs. Therefore, we used the same DRM materials as [Roediger and McDermott \(1995\)](#). DRM instructions were provided to the LLMs as if they were a sapient being.

However, due to the structural differences we had to adapt the DRM paradigm in two ways. Human participants must memorize each word in a given amount of time, e.g., 1.5 seconds in the case of [Roediger and McDermott \(1995\)](#). The available amount of time per item during encoding can

affect memory performance ([McDermott and Watson, 2001](#); [Smith and Kimball, 2012](#)), as can the retention period ([Seamon et al., 2002](#)), i.e., the time between encoding and recognition test. As there is no concept of time for LLM input, we instead considered the context window length as an analogous variable. Regarding the retention interval, we replaced the original math tasks from [Roediger and McDermott \(1995\)](#) with a text inference task. If context window length has analogous effects to the passing of time, we attribute the length of the text for this buffer task to the retention period. To separate the effect of context window length and the DRM effect we tested our experiment with multiple context window lengths up to a maximum of 32k tokens, and examined whether performance was contingent on context window length.

With regard to encoding the studied lists, we omitted implementing any buffer token to simulate time to encode the item, i.e. the 1.5 seconds retention period humans receive after hearing each word. Furthermore, we changed the mode of presentation. Humans are presented with the items one by one with the understanding that each word is a single item. We were concerned that submitting the DRM word list elements in the same manner to the LLMs might confuse them. Therefore, we submitted the studied words as a single list of strings.

2.3 Multi-turn prompt sequence

We adopt the same multi-turn prompt structure for all LLMs in our experiments. We will describe each turn conceptually here and the exact prompt wording can be found in Appendix A.

Each conversation started with a system prompt. We used a generic template for a helpful assistant that covers typical use cases a LLM agent might encounter. The system prompt was followed by the DRM encoding phase. The LLM receives the instructions to memorize the subsequent list of words, and a list of DRM word lists was appended in vector form. After the encoding phase, came the buffer task. The buffer task contained three elements. Firstly, instructions to analyze the subsequent text and the warning that it must answer questions about it later. Secondly, an excerpt from the WHO global patient safety report 2024 ([World Health Organization, 2024](#)). Lastly, the LLM was asked to recall a key statistic from the executive summary at the beginning of the excerpt. Again, for each LLM the answer was produced once and added to the prompt. The purpose of the buffer task is to simulate the re-

tention period and we tested several token lengths to account for the general finding that performance degrades the longer the context window becomes (Du et al., 2025). We use the performance on the buffer task as a safeguard to exclude conditions in which the length of the buffer task leads to a total collapse in performance.

The final part of the prompt was the DRM recognition phase. The LLM was informed that it would be presented with individual words and its task was to judge whether the word was part of the list of words memorized during the encoding phase. It was also instructed to only reply with Yes or No. Then, one of the words from the test set was appended to the end. Notably, we tested each word from the test set with the same prompt structure and recorded the LLMs answer. So, each word from the test set was tested independently.

2.4 Materials

We used the original 24 DRM lists from Roediger and McDermott (1995). We randomly selected the following lists indicated by their corresponding lure words: girl, thief, rough, bread, high, chair, fruit, sweet, anger, mountain, needle, black, spider, soft, foot, cold. The encoding vector also preserved the sequence of items within each list, because words earlier in the list are more similar to the corresponding LURE word.

For the buffer task, we used the WHO global patient safety report 2024 (World Health Organization, 2024). We extracted a text-only segment starting with the executive summary (page 28). For this experiment, we created buffer tasks of varying token lengths of 1k, 2k, 4k, 8k, 16k, and 32k. Again, token lengths were determined using the Llama 3.2 3b tokenizer. Furthermore, once the respective token limit was reached, we continued to include the following text until the end of the paragraph was reached. This results in slightly larger token counts, but avoids potential confusion downstream due to the text being cut off mid-sentence.

2.5 Design & Measures

We evaluated four open-weight models at two scales: smaller models (Llama 3.1 8b, Qwen 2.5 7b) and larger models (Llama 3.1 70b, Qwen 2.5 72b). We used the unquantized Instruct variant for all models and set the temperature at 1. Llama and Qwen as model families were chosen because they are open weights with logit access, there were two comparable parameter counts in the low and high

range per family, and both families are widely used in research and practice. For each test word, we recorded the model’s decision with a single token (YES/ NO). Additionally, we recorded the model’s top 20 output token choices and their corresponding probabilities.

Dependent variables We define four item types, each evaluated under both the discrete and the continuous measure described below. For the discrete measure, the dependent variable is the proportion of "Yes" responses to items of that type; for the continuous measure, it is the mean $p(\text{YES})$ over those items. In both cases the quantity is the rate or strength at which the model judged a word to have appeared during encoding. Whether a "Yes" is correct or an error depends on the item type:

- Familiar words (FW): endorsement of studied words. "Yes" is correct, so FW indexes veridical recognition.
- Unfamiliar words (UW): endorsement of unstudied, unrelated words. "Yes" is an error, so UW indexes the general tendency to endorse any word and serves as a response-bias baseline.
- Familiar lures (FL): endorsement of lures of studied lists. "Yes" is an error indexing false memory.
- Unfamiliar lures (UL): endorsement of lures of unstudied lists. "Yes" is an error; UL serves as a control.

Because UW captures the model’s baseline tendency to respond "Yes," false memory (FL) is meaningful only relative to it: an effect is present when FL exceeds UW. UL is a complementary control, since lures of lists that were never studied should not be endorsed above baseline if false recognition depends on the studied list having been presented.

Discrete measure For each item type, we took the model’s greedy-decoded answer to every word of that type. A response counted as "Yes" only if the decoded token was a "Yes" token (see below), and any other output counted as "No." The dependent variable is then the proportion of "Yes" responses among items of that type, yielding one "Yes" rate per item type per model.

Continuous measure For each item type, to obtain a graded estimate, we computed the probability of a "Yes" response for each word from the model's output distribution obtained at temperature = 1. We first normalized the probabilities of the top 20 tokens to sum to one, then summed the normalized mass of all "Yes" tokens; the remaining mass, comprising "No" and any other tokens, formed the complement. This yields a probability of endorsement, $p(\text{YES})$, in $[0, 1]$ for each word, equal to zero when no "Yes" token appeared among the top 20. The dependent variable is the mean $p(\text{YES})$ over items of that type, yielding one value per item type per model.

Defining "Yes" tokens To identify which tokens counted as "Yes," we first compiled the set of all unique tokens appearing in the top 20 outputs across every model and recognition trial. We then used a language model, Claude Sonnet 4.6 (Anthropic, 2026), to classify which of these tokens corresponded to a "Yes" response. The resulting "Yes"-token set was verified by the principal investigator and used for both discrete and continuous measures.

2.6 Analyses

For the first claim, that false memories occur, we compared each model's endorsement of familiar lures against its endorsement of unfamiliar words (unstudied and unrelated), pooled across buffer lengths. The unfamiliar-word rate serves as a response-bias baseline (familiar lure $n = 96$, baseline $n = 144$). On the binary response, we used a two-sided Fisher's exact test, and report the odds ratio (OR) with its 95% confidence interval (CI), and the p value. As a control, we applied the same test to unfamiliar lures, which should not be endorsed above baseline if false recognition depends on the studied list. On the continuous measure, we used a Mann-Whitney U test of whether $p(\text{YES})$ for familiar lures ranked above baseline. We report the rank-biserial correlation r as the effect size, alongside the raw average probabilities so that the magnitude of the difference can also be judged. For the second claim, that false memory increases with context length, we tested whether endorsement of familiar lures rose across the six buffer lengths (1k to 32k tokens) within each model. On the binary response, we fitted a logistic regression of the YES response on $\log_2(\text{buffer length})$, reporting the slope, its 95% confidence interval, and the p value. On

the continuous measure, we computed a Spearman rank correlation between $p(\text{YES})$ and buffer length, reporting ρ and p .

3 Results

3.1 Buffer task performance

We verified the integrity of the experimental prompt through the performance on the buffer task. All models, over all context lengths, correctly answered the inference question about the content of the buffer task.

3.2 Recognition performance

We first assessed basic memory performance, that is, the ability to correctly recognize studied words and reject novel ones, using the discrete (greedy-decoded) responses. Recognition of previously studied words was at or near ceiling in every model and at every context length. Both Llama models recognized all studied words across all buffer lengths (48/48 at each length), while the two Qwen models did so at 96 to 100%, dropping no lower than 46/48 at the longest context. To assess whether high endorsement instead reflected a general YES-bias, that is, a tendency to accept any word, we examined endorsement of unrelated unfamiliar words, which were never studied and should be rejected. This baseline remained low across most conditions (0 to 4% for Llama 8b, Llama 70b, and Qwen 72b, and 4% for Qwen 7b up to 16k) and rose only moderately at the longest context, reaching 8%, 12%, and 21% at 32k for Llama 8b, Llama 70b, and Qwen 7b, respectively. Models therefore distinguished studied from non-studied words reliably, showing only a slight liberal drift at 32k rather than a pervasive YES-bias.

3.3 False memories

False memories are indexed by endorsement of familiar lure words, that is, words semantically related to studied items but not themselves studied, at a rate exceeding the model's baseline endorsement of unrelated unfamiliar words. On the binary response, this effect was statistically reliable in two of the four models: Llama 70b endorsed familiar lures well above baseline (14% vs. 2%; OR = 7.36, 95% CI [1.92 41.14] $p < .001$) and Qwen 7b likewise (18% vs. 7%; OR = 2.88, 95% CI [1.17 7.39], $p = .012$). The effect did not reach significance for Llama 8b (5% vs. 3%; OR = 1.92, 95% CI [0.4 9.93], $p = .49$) or Qwen 72b (3% vs. 1%;

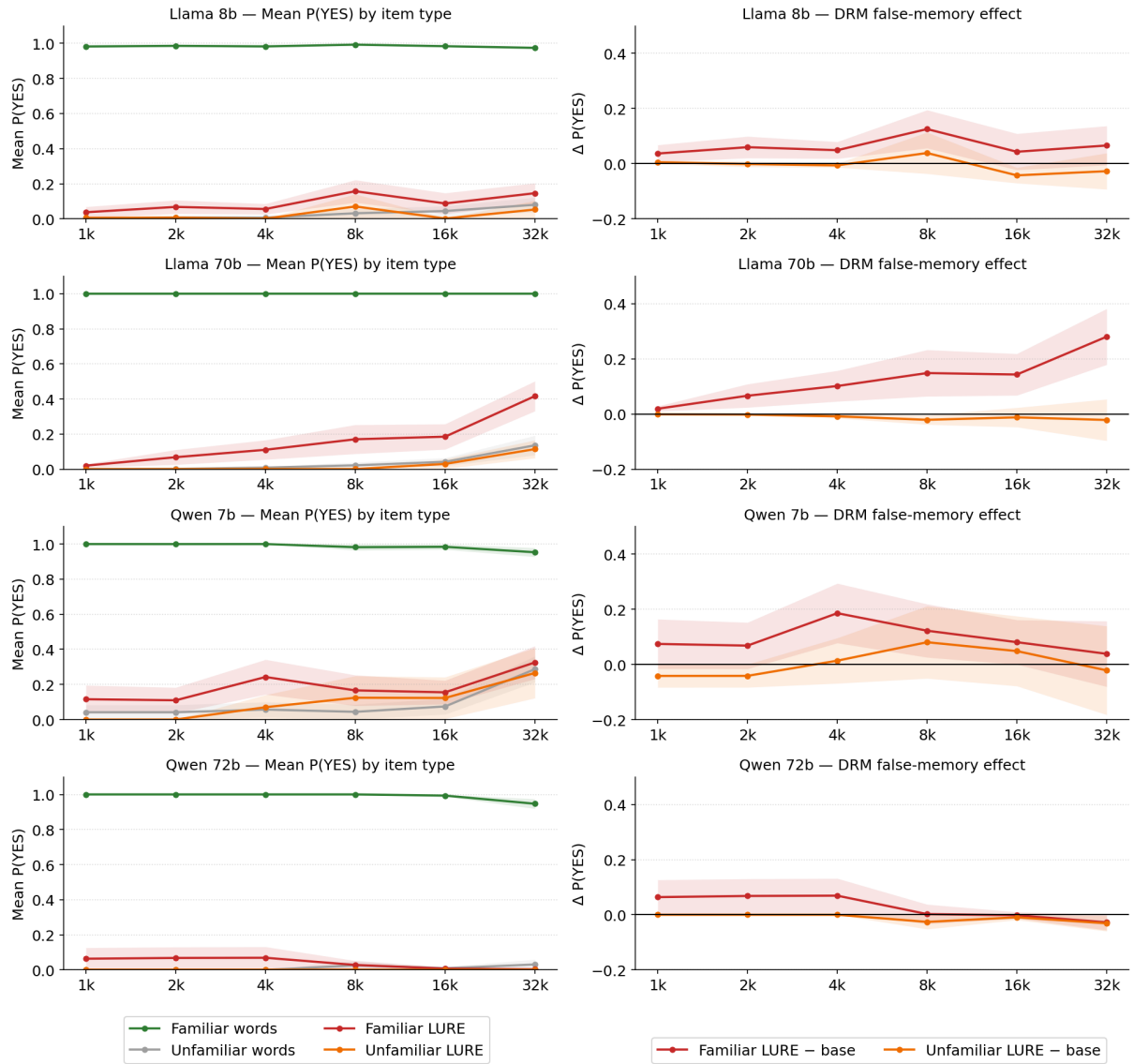


Figure 1: Continuous recognition strength across context length, by model. Left panels: mean $p(\text{YES})$ for each item type — familiar/studied words, unfamiliar/unrelated words (the response-bias baseline), familiar lures (semantic associates of studied lists), and unfamiliar lures (associates of non-studied lists; control). Right panels: the DRM false-memory effect, computed as familiar-lure minus baseline $p(\text{YES})$, with the unfamiliar-lure minus baseline difference shown as a control. Bands are ± 1 SE of the mean. x-axis: buffer length (1k–32k tokens), log2 scale.

OR = 2.29, 95% CI [0.26 27.81], $p = .39$). The unfamiliar-lure control was non-significant in every model, confirming that elevated endorsement was specific to lures of studied lists rather than an intrinsic attractiveness of lure words.

The continuous measure showed that a gist-consistent signal was nonetheless present in all four models (Figure 1, left panels): familiar lures reliably out-ranked baseline words in $p(\text{YES})$ in Llama 8b ($r = 0.54$, $p < .001$, lure mean = 0.09, baseline mean = 0.03), Llama 70b ($r = 0.62$, $p < .001$, lure mean = 0.16, baseline mean = 0.04), Qwen 7b ($r = 0.39$, $p < .001$, lure mean = 0.19, baseline mean

= 0.09), and Qwen 72b ($r = 0.69$, $p < .001$, lure mean = 0.04, baseline mean = 0.01). The control comparison (unfamiliar lures vs. baseline) was non-significant with near-zero effect size in every model ($|r| \leq 0.22$), and the bias-corrected difference between lure and baseline endorsement is shown in the right panels of Figure 1. Critically, this continuous signal corresponded to overt false recognition only in some models. In Llama 70b and Qwen 7b, the rank effect, higher average probability for lure words than baseline words, was corroborated by elevated YES rates (14% and 18%), indicating false memories that the model acted upon. In

Llama 8b and Qwen 72b, by contrast, the rank effect was comparably consistent (indeed, the largest in Qwen 72b) yet was negligible in magnitude: median $p(\text{YES})$ was at or near zero in both conditions, and overt false recognitions were no more frequent than baseline. We therefore interpret these latter models as registering the lures’ semantic relatedness internally without translating it into false recognition. Across the set, the gist signal was thus near-universal, while its conversion into an overt false memory was model-specific.

3.4 Memory performance across context window length

Veridical recognition of studied words remained at or near ceiling across all context lengths in every model, showing no decline as the context window grew. The novel effect appeared instead in false memory. Restricting to familiar-lure trials, endorsement increased with context length most clearly in Llama 70b, where both measures showed a reliable positive trend (logistic slope on $\log_2$ buffer = 0.75, 95% CI [0.26, 1.24], $p = .003$; Spearman $\rho = 0.52$, $p < .001$; see Figure 1, Llama 70b panels). Qwen 7b ($\rho = 0.42$, $p < .001$) and Llama 8b ($\rho = 0.37$, $p < .001$) showed the same positive association on the continuous measure but not on the sparser binary response (Qwen 7b slope = 0.11, 95% CI [-0.20, 0.42], $p = .482$; Llama 8b slope = 0.26, 95% CI [-0.30, 0.82], $p = .356$). Qwen 72b showed no positive trend on either measure, and a numerically negative one (binary slope = -0.65, 95% CI [-1.58, 0.27], $p = .165$; Spearman $\rho = -0.05$, $p = .634$). Thus, where false memory occurred, it tended to intensify with longer context, but this dependence on context length was strongly model-specific.

4 Conclusion

We transferred the DRM paradigm, a well-studied psychological method for evoking false memories in humans, to the study of how LLMs use their context window. Consistent with [Cao et al. \(2025\)](#), we observed DRM-style false memories in both the Llama and Qwen families, but our results refine that picture. A gist-consistent signal, in which semantically related lures attracted more endorsement probability than unrelated words, was present in all four models, yet it produced overt false recognitions reliably in only two of them. The graded probability measure thus revealed an internal tendency toward false memory even in models whose

discrete responses showed little or none.

By contrast, veridical recognition of studied words remained at or near ceiling across all context lengths. The human-like rise in false memory emerged not through decay of the verbatim trace, as fuzzy-trace theory describes in humans, but through a strengthening of gist-based responding while the verbatim information remained intact. This bears directly on how LLMs use the information in their context window. The models endorsed lures as previously seen despite the lists remaining available verbatim in the input. Instead of verifying against the exact context, the models answered based on gist. Our findings should raise practical concern as well. That the correct information is present in the context window is no guarantee that the model uses it correctly. In RAG applications with a large volume of input, or in long-context reasoning, the model may follow the gist rather than the exact information, producing errors. The impact in practice and possible solutions remain to be developed, but the DRM paradigm is well suited to revealing this: it offers a controlled, psychologically grounded way to elicit and measure gist-driven errors in LLMs.

Limitations

Several limitations should be acknowledged. First, in replicating the original DRM paradigm we inherited its imbalance across item categories: the recognition test contained substantially more familiar words (48) than familiar lures (16) whereas a balanced design would include equal numbers per category. Balancing would have required generating additional DRM lists whose lure effectiveness has not been independently validated, so we retained the established lists.

Second, our test of whether false memories occur pooled trials across buffer lengths to increase statistical power and simplify the analysis. This introduces non-independence: each item appears once per buffer length, and, more importantly, items are nested within the 24 DRM lists, which differ in how strongly their lures pull. Our tests treat trials as independent and therefore likely yield more optimistic p -values than a list-aware analysis would. A mixed-effects model that accounts for between-list variance would be more conservative, and confirming our effects under such a model is left to future work; we note that the largest effects are substantial enough that they are unlikely to be artifacts of this

clustering.

Third, our four models span only two families and two parameter scales, so model family and scale are confounded. We can therefore describe the differences between models but cannot attribute them specifically to architecture or to size.

Finally, our continuous measure relies on the answer being expressible as a small, identifiable set of tokens, which is possible here because the recognition test elicits a single "Yes" or "No" response. The approach therefore generalizes to tasks with localized, enumerable answer tokens but does not directly extend to free-form generation.

References

- Anthropic. 2026. [Claude sonnet 4.6 \(large language model\)](#).
- Frederic Charles Bartlett. 1995. *Remembering: A study in experimental and social psychology*. Cambridge university press.
- Charles J Brainerd and Valerie F Reyna. 2002. Fuzzy-trace theory and false memory. *Current directions in psychological science*, 11(5):164–169.
- Zhaoyang Cao, Lael Schooler, and Reza Zafarani. 2025. Analyzing memory effects in large language models through the lens of cognitive psychology. *arXiv preprint arXiv:2509.17138*.
- Florin Cuconasu, Giovanni Trappolini, F. Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. [The power of noise: Redefining retrieval for rag systems](#). *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- James Deese and Roger A Kaufman. 1957. Serial effects in recall of unorganized and sequentially organized verbal material. *Journal of experimental psychology*, 54(3):180.
- Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Babu Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A Huerta, and Hao Peng. 2025. [Context length alone hurts LLM performance despite perfect retrieval](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23281–23298, Suzhou, China. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan Ö. Arik. 2024. [Long-context llms meet rag: Overcoming challenges for long inputs in rag](#). *ArXiv*, abs/2410.05983.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paran-jape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Elizabeth F. Loftus and John C. Palmer. 1974. [Reconstruction of automobile destruction: An example of the interaction between language and memory](#). *Journal of Verbal Learning and Verbal Behavior*, 13(5):585–589.
- Kathleen B. McDermott and Jason M. Watson. 2001. [The rise and fall of false recall: The impact of presentation duration](#). *Journal of Memory and Language*, 45(1):160–176.
- Bennet B Murdock Jr. 1962. The serial position effect of free recall. *Journal of experimental psychology*, 64(5):482.
- Valerie F Reyna and Charles J Brainerd. 1995. Fuzzy-trace theory: An interim synthesis. *Learning and individual Differences*, 7(1):1–75.
- Henry L Roediger and Kathleen B McDermott. 1995. Creating false memories: Remembering words not presented in lists. *Journal of experimental psychology: Learning, Memory, and Cognition*, 21(4):803.
- Nikolaus Salvatore, Hao Wang, and Qiong Zhang. 2025. [Lost in the middle: An emergent property from information retrieval demands in llms](#). *Preprint*, arXiv:2510.10276.
- Daniel L Schacter. 1999. The seven sins of memory: Insights from psychology and cognitive neuroscience. *American psychologist*, 54(3):182.
- John G. Seamon, Chun R. Luo, Jonathan J. Kopecky, Catherine A. Price, Leeatt Rothschild, Nicholas S. Fung, and Michael A. Schwartz. 2002. [Are false memories more difficult to forget than accurate memories? the effect of retention interval on recall and recognition](#). *Memory & Cognition*, 30(7):1054–1064.
- Troy A. Smith and Daniel R. Kimball. 2012. [Revisiting the rise and fall of false recall: Presentation rate effects depend on retention interval](#). *Memory*, 20(6):535–553.
- World Health Organization. 2024. [Global patient safety report 2024](#).
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

A Appendix

This section reports the exact wording of the prompts described in section 2 for reproducibility.

A.1 System prompt

You are a helpful, honest AI assistant. Assist users with questions, explanations, writing, analysis, coding, and discussion.

Principles

- Be helpful: Take requests seriously and give substantive answers.
- Be honest: Do not fabricate information. State when you are uncertain or do not know something.
- Be clear: Use plain language. Avoid unnecessary hedging or filler.
- Reason carefully for complex questions before answering.

Handling uncertainty

- Ask a clarifying question if a request is ambiguous.
- Do not invent facts you are unsure of. Label inferences and opinions as such.
- Your training data has a cutoff; you may not know recent events.

Behavior

- Follow the user's instructions on format and length.
- Prioritize accuracy for factual questions and quality for creative tasks.
- Show reasoning for code, math, and logic where it helps.
- Decline requests that would cause clear and serious harm. Engage thoughtfully with sensitive but legitimate topics rather than refusing reflexively.

A.2 DRM encoding instruction

For this cognitive task you will be studying a lists of words. Please pay close attention to the words because you will be tested on them later. This is the list of words to memorize: {encoding_list}

A.3 Buffer task instruction

Next study the following report. You will be asked a question about the content.

{Buffer text}

According to the WHO report you just read, state in one short sentence what proportion of patients experience harm from care.

A.4 DRM recognition instruction

Earlier you were asked to memorize a list of words. I will now present you with a word and your task is to decide whether it was on that list. Answer with only YES or NO. Was the following word among the words you studied? {recognition_word}