Università degli Studi di Padova

Analogical Reasoning of the Gemini 1.5 flash Large Language Model

Cognitive, Behavioural and Social Data Project Report

Author: Felice Francario, Matteo Squeri

July 19, 2024

Abstract

This study explores the analogical reasoning capabilities of large language models (LLMs) through the lens of the Relational Luring Effect, a cognitive phenomenon where individuals are more likely to falsely remember or recognize non-presented items related to presented information. Building on the theoretical framework established by Popov et al., we conducted two experiments to assess the Gemini 1.5 flash LLM's performance in associative recognition tasks involving relational lures. In Experiment 1, the LLM demonstrated a tendency towards the Relational Luring Effect similar to that observed in human participants, with a higher tendency to falsely recognize relationally similar pairs as previously seen, although this was not statistically significant. In Experiment 2, the LLM did not exhibit the Relational Luring Effect, likely due to its perfect memory for intact word pairs. These findings suggest that while LLMs can simulate certain aspects of human analogical reasoning, their perfect memory for studied items may lead to differences in behavior compared to human cognition.

1 Introduction

Analogical reasoning stands as one of the foundational pillars of human cognition, enabling individuals to draw upon familiar experiences and patterns to understand novel situations, solve complex problems, and make informed decisions. At its essence, analogical reasoning involves the process of recognizing structural correspondences between different domains. This cognitive process relies on the identification of shared relational patterns, functional similarities, or causal relationships between entities, situations, or concepts, thereby facilitating the transfer of knowledge and understanding across contexts.

Popov et al. argued that semantic relations (e.g., works in: NURSE HOSPITAL) have abstract independent representations in long-term memory and that the same representation is accessed by all exemplars of a specific relation. If relations are represented abstractly in long-term memory, then every time a participant sees a different exemplar of a relation, the activation strength of that relation in LTM should increase and we should observe an effect called the Relational Luring Effect. The Relational Luring Effect refers to a memory phenomenon where individuals are more likely to falsely remember or recognize a non-presented item that is related to presented information. This effect can be verified particularly in associative recognition tasks.

Large language models, exemplified by cutting-edge architectures like OpenAI's GPT (Generative Pre-trained Transformer) series, have revolutionized natural language processing by exhibiting remarkable proficiency in understanding, generating, and manipulating textual data. These models, trained on vast corpora of text, encode intricate linguistic patterns, semantic relationships, and contextual nuances, enabling them to produce coherent and contextually relevant responses to a wide array of prompts and queries. Yet, despite their impressive capabilities, questions persist regarding their ability to engage in higher-order cognitive processes such as analogical reasoning.

2 Objectives

In this report, we embark on a exploration of analogical reasoning within large language models, building upon the theoretical framework established by Popov et al. in their seminal study on the Relational Luring Effect. We will replicate, as much as possible, the 2 experiments introduced in Popov et al.'s paper to verify the relational luring effect on a Large Language Model known as Gemini 1.5 flash with 32 billion parameters.In particular we will only focus on the accuracy and False Alarm rates in these experiments, ignoring all the aspects that are not testable on a LLM like the response times.

3 Setup description

3.1 Experiment 1

In Experiment 1, we aim to test the Gemini LLM's ability to recognize word pairs and its susceptibility to relational lures. The materials for this experiment include a curated set of word pairs where each pair shares a semantic relation (e.g., FLOOR CARPET). Additionally, relationally similar lures (e.g., TABLE CLOTH) and relationally dissimilar lures (e.g., PIPE WATER) are prepared. The participants in this study are instances of the Gemini 1.5 flash large language model. We created 10 different data sets for different instances of the model. For each instance the model was provided with 63 word pairs divided into 3 blocks of 21 word pairs for both, a study phase and a test phase. In between we added a distraction phase in which the model was asked to count backwards from 60 until 0.

During the study phase, the Gemini LLM is presented with a list of 21 word pairs. In the test phase, the Gemini LLM is presented with a mixture of intact pairs (those presented during the study phase), relationally similar lures (new pairs that share the same relational structure), and relationally dissimilar lures (new pairs with different relational structures). The model is asked to classify each pair as "intact" (studied) or "recombined" (lure). Responses are recorded to determine hits (correctly identified old pairs), false alarms for relationally similar lures, and false alarms for relationally dissimilar lures. The focus is on measuring accuracy and false alarm rates to identify the presence of the Relational Luring Effect.

3.1.1 Dataset Building

No proper dataset was provided with the original paper in order to carry out the experiment with perfect replication. However we were able to come closer to the original design by constructing sets of word pairs for both the study and test phase of the experiment using exemplars of word pairs for 35 different relation groups provided in the Appendix A of the paper.

We followed the procedure detailed for constructing multiple sets of 4 pairs, where 2 belongs to a relational group, and 2 to another, we then selected 3 random pairs for each set and permuted 2 pair words in order to construct our training set for one participant. For a single model instance performing the entire task, this meant that starting from 21 sets of such sets of 3 pairs, we constructed sets containing the intact pair, it's corresponding relational lure (the permuted pair of the same relational group of the intact one) and a non relational lure. The resulting pairs order in which they are shown in the model for both the study and testing phase was randomized in each instance.

We decided to perform the experiment on 10 different instances of the model, ideally each simulating a different human participant, where the original study tested 12 students.

Following this design we were able to collect meaningful statistics from the different instances of the experiment, since no model instance was shown the same pairs in the same role, while each block of the trial, corresponding to 21 pairs studied and then classified, was guaranteed to not have duplicate pairs, ensuring that the model wasn't confused by previous section of the experiment.

3.1.2 Prompt Design and Model Behaviour

The core approach we followed in order to design and execute this project was one of trying to get as closer as possible to the original design as was performed on human participants.
We performed multiple iterations on the design of the prompts, concentrating mostly on the aspect of formalism and the specificity of the information provided to explain and execute the task.
In the end we settled to a concise and "human" approach to the prompts style. We did this in order to obtain a comparison as direct as possible for the results to the human counterpart of the study. The responses provided by Gemini 1.5 flash show that the LLM followed the instructions perfectly, never responding with unrequested text added outside the scope requested, be it classification or a counting task on the distraction phase, which was also maintained from the original study.
The only strange phenomena in the responses of the model was the addition of multiple line break symbols $\backslash n$ after the requested, the amount of line breaks increased the further in the experiment the instance was in. We found no explanation for this minor repeated pattern.
The final prompts are listed in appendix A.

3.2 Experiment 2

In Experiment 2, we investigated the impact of varying exemplar frequency on the Gemini LLM's recognition performance. The materials included a comprehensive set of word pairs representing specific relations (e.g., works in: NURSE HOSPITAL) and multiple exemplars for each relation (e.g., works in: DOCTOR CLINIC, TEACHER SCHOOL). Since there were 527 different word pairs in the test phase and it was continuous, the experiment was run once with one instance of the model due to the time-consuming nature of the task and limits of the free-version of the model's API.

During the study phase, the Gemini LLM was presented with a sample of word pairs in order to understand the tasks, with different exemplars for each relation shown a specified number of times (0, 1, 2, 3, or 4 different exemplars). For example, for the relation "works in", the model might see "NURSE HOSPITAL", "DOCTOR CLINIC", and "TEACHER SCHOOL". In the test phase, the Gemini LLM was presented with a series of 527 word pairs that included new unseen pairs (which are all non relational pairs), recombined pairs (which are all relational) and old/intact pairs (which the model has exactly seen before in the experiment) . The model was asked to classify correctly each pair shown as "new ", "recombined or"old". Responses were recorded, focusing sorely on the accuracy on the correct types of responses as the experiment progressed.

3.2.1 Dataset Building

In the case of the second experiment the dataset was also not provided by the original study. We started again from the pairs grouped in relational types provided by the appendix: this provided us with a set of 165 unique word pairs, where each pair was an instance of 1 of 35 unique relations. Each of this relation type was represented by four (10 cases) or five (25 cases) pairs.
The original design indicated that to build a full dataset of 527 pairs to use in an experiment each of the 165 pairs was to be shown as a "recombined" pair, as a "old/intact" pair, and as a "new" pair. Precisely each of the "new" pair was a permutation of a pair with another, always maintaining relative word position inside the pair.
In performing the experiment it's of course needed that each "recombined" pair appears after a permutation of the original pairs that contained the two words was already shown to the model (so two "new" pairs). Furthermore each "old" pair needed to appear after it's "recombined" identical version was already shown previously in the experiment.
Given the original 165 unique word pairs it's indeed possible to build a dataset that can correctly used to perform the experiment, however the original paper also indicated strict conditions for the randomizations of the experiment, which are as follows:

1. each recombined pair is presented at least 20 trials after the presentation of the new pairs that contain the words in that recombined pair;
2. each intact(old) pair is presented at least 20 trials after its previous occurrence;
3. there is a trial lag of at least 20 trials between presentations of different exemplars of the same relation to prevent grouping strategies;
4. no new exemplars of a relation appear between a specific recombined exemplar and its repetition as an intact pair;
5. there are no more than 4 consecutive intact or recombined trial types

We weren't able to design a randomization features that consistently satisfied all those constrains. We were however able to design an iterative algorithm for building the dataset that however provided us a single correct dataset of 527 pairs trials that satisfied all the constrain. For this reason we limited our experiment to just that dataset and it's result to just one complete run performed on one instance of the model, not producing average results.
Finally, the API of Gemini 1.5 flash limited the number of calls we were able to run in the experiment, concluding it short of showing the full 527 pairs to the model.
The final dataset classified by the model in the testing phase included just 416 pairs divided in 155 "new" pairs, 134 "recombined" and 127 "old".

3.2.2 Prompt Design and Model Behaviour

The prompting strategy applied in this experiment was similar to the first one.
Also similarly, the model always provided adequate responses to the pairs to classify and never strayed from the instruction producing hallucinations even after 400 prompted trials.
The prompts used for the second experiment are described in Appendix B.

4 Results

4.1 Experiment 1 Results

The results of Experiment 1, conducted with 10 different instances of the Gemini LLM, are depicted in Figure 1. The graph shows the proportion of false alarms for both relational and non-relational lures, separated by the order in which they were presented relative to intact pairs.

The data indicate that the proportion of false alarms was generally higher for relational lures compared to non-relational lures. Specifically, when relational lures were presented after intact pairs, the false alarm rate was approximately 0.15, while non-relational lures had a slightly lower false alarm rate. When relational lures were presented before intact pairs, the false alarm rate increased to around 0.20. These results suggest that the Gemini LLM exhibits the Relational Luring Effect, with a higher tendency to falsely recognize relationally similar pairs as previously seen.

When comparing these findings to the original experiment by Popov et al., depicted in Figure 2, we observe a similar pattern. In the original human study, the proportion of false alarms for relational lures was also higher than for non-relational lures, with false alarms increasing when relational lures were presented before intact pairs. However, it is important to note the significant standard deviations observed in our LLM results. The large standard deviation could be attributed to the small sample size of test pairs. A single additional error in a limited number of pairs can disproportionately affect the false alarm rate.

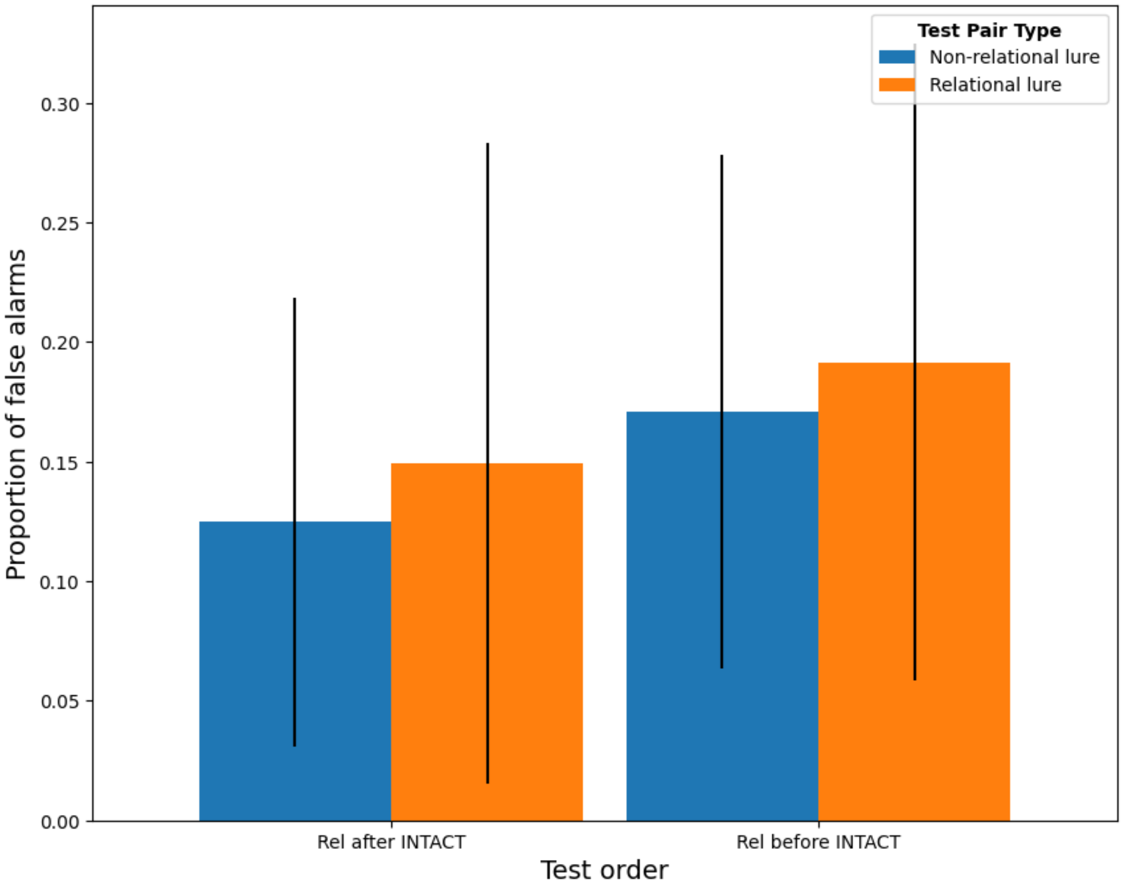

FIGURE 1: Proportion of false alarms for relational and non-relational lures, separated by test order (relative to intact pairs) in Experiment 1. Error bars represent standard errors.

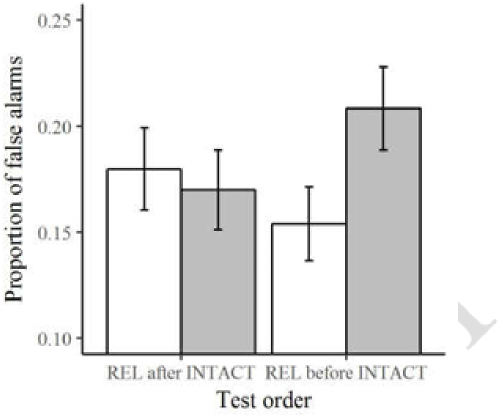

FIGURE 2: Proportion of false alarms for relational and non-relational lures, separated by test order (relative to intact pairs) in the original human experiment by Popov et al.

Despite the variance, the mean results are noteworthy, showing a clear relational luring effect similar to that observed in human participants. This may indicate that the Gemini LLM can simulate certain aspects of human analogical reasoning and memory processing. However the overall accuracy scores between recombined relational lure and recombined non-relational lures dont seem to be statistically significant even though the overall accuracy of recombined relational lure's is lower as expected. One interesting result is that the model ended up having a 100% accuracy for the intact word pairs, indicating that the model completely memorized the previously studied pairs.

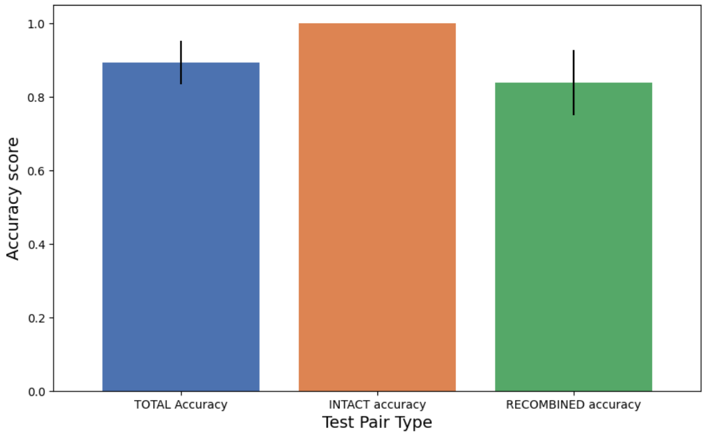

FIGURE 3: Accuracy percentages of intact and recombined word pairs in the first experiment.

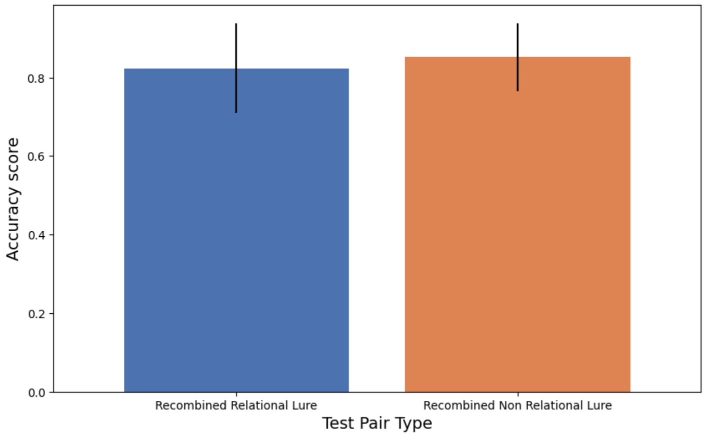

FIGURE 4: Accuracy percentages of recombined word based on if they were relational lures or not.

	Instance 1	Instance 2	Instance 3	Instance 4	Instance 5	Instance 6	Instance 7	Instance 8	Instance 9	Instance 10
First Trial	1.0	0.81	1.00	0.95	0.81	0.95	1.00	1.00	0.81	0.95
Second Trial	0.9	0.52	0.86	1.00	0.90	0.95	0.81	0.86	0.81	0.90
Third Trial	0.9	0.90	0.95	0.86	0.90	0.95	0.86	0.90	0.90	0.81

TABLE 1: Overall accuracy values obtained by each model instance on all three parts (trials of 21 pairs) of the first experiment

4.2 Experiment 2 Results

In Experiment 2, contrary to our hypothesis, we did not find evidence of the Relational Luring Effect. Instead, the obtained results indicated the opposite trend. The false alarm rates for recombined pairs did not increase with the number of different exemplars presented previously during the experiment. In fact, there was a decrease in false alarm rates, meaning an increase in accuracy, for relationally similar lures as the number of exemplars increased, suggesting that the LLM became more accurate in distinguishing between recombined and new pairs with more exposure to relational structures. This result is the opposite of what was found in original experiment done on humans (figure 7), which suggested for the relational luring hypothesis to hold that false alarms to recombined pairs, meaning lower accuracy on this type of pairs, should increase when people have already seen the relation in different exemplars, and that the effect should increase with increasing the number of exemplars. Another significant result, possibly the cause of the earlier discrepancy, can be seen in figure 6. The Model has also perfect accuracy in identifying intact pairs throughout the entirety of the experiment, suggesting very different capabilities compared to humans, which didn't exhibit this amount of recall of identical exemplars for this long of a duration.

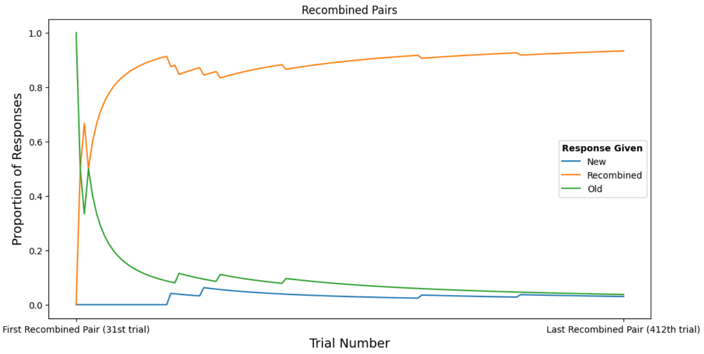

FIGURE 5: Proportion of responses in Experiment 2 for each type of response for recombined pairs as a function of the trial number

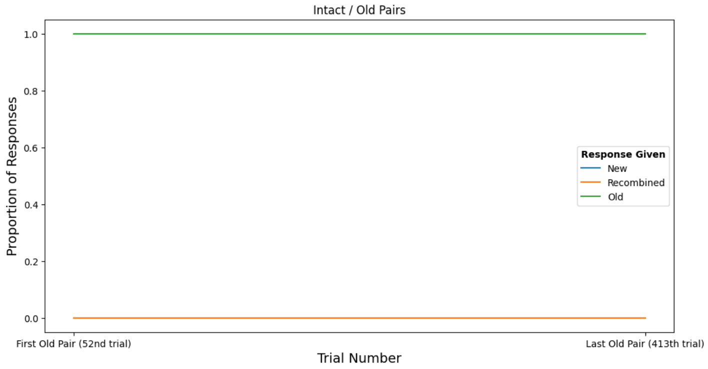

FIGURE 6: Proportion of responses in Experiment 2 for each type of response for intact pairs as a function of the trial number

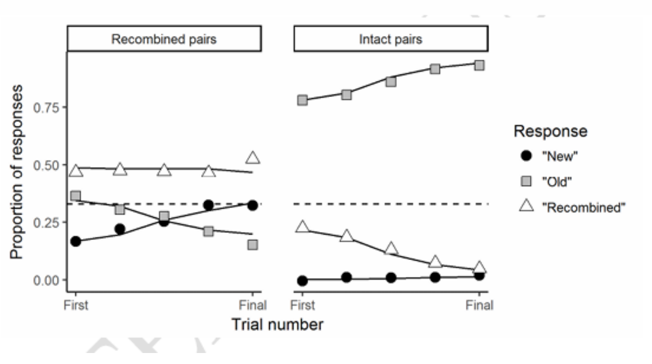

FIGURE 7: Proportion of responses in Popov's original Experiment 2 for each type of response for each type of pair as a function of the trial number

5 Conclusions

The findings from these experiments have significant implications for our understanding of large language models' capabilities in analogical reasoning. The observed Relational Luring Effect in Experiment 1 suggests that these models can simulate certain aspects of human cognitive processing, particularly in recognizing and generalizing relational structures.

However, the absence of the Relational Luring Effect in Experiment 2 indicates a divergence from human cognition under different conditions. Specifically, the results suggest that the Gemini LLM has a perfect memory for intact word pairs, which prevents the typical increase in false alarms seen in human participants when exposed to relationally similar lures. This discrepancy highlights a key difference between human and model memory processes.

To address this, future research should consider increasing the number of word pairs in the study phase. By presenting a larger set of pairs, it may be possible to challenge the model's memory capacity and better simulate the conditions under which the Relational Luring Effect is observed in humans. Additionally, varying the complexity and diversity of the word pairs could provide further insights into the model's relational reasoning abilities.

Overall, these results suggest that while large language models like Gemini can exhibit human-like cognitive processing in certain contexts, their perfect memory for studied items can lead to differences in behavior, particularly in tasks involving relational lures. Understanding these differences is crucial for advancing the development of models capable of more closely mimicking human cognitive processes.

References

[Popov et al., 2017] Popov, V., Hristova, P., & Anders, R. (2017). The relational luring effect: Retrieval of relational information during associative recognition. Journal of experimental psychology. General, 146.

A Prompts used in the first experiment

The first prompt used in the experiment introduced the training phase, providing also the 21 pairs that the model will need to remember for the trial run:

Training Prompt

Good morning, This is a memory experiment. You will be presented with a list 21 word pairs. Each row of the file contains the 2 words that comprise the word pair in the columns Word1 and Word2. Memorize the word pairs without writing them down. Reply with ”Got it!” if you understand the task and have memorized the word pairs. You may begin the experiment! Here are the 21 word pairs:

The second prompt was used to mirror the distraction phase for human participants, before beginning the actual testing phase:

Distraction Prompt

Count backwards from 60 by threes until you reach 0.

The third prompt introduces the proper task for the model and kickstart the testing phase, being then followed by 21 simple messages that just report the pairs to classify:

Testing Prompt

I am going to give you one word pair at a time. Do the following: If you think you have seen the pair of words among the 21 word pairs in the last file that i previously gave you, respond with ”intact”; If you think you have studied the single words in separate pairs in the same last file that i previously gave you respond with ”recombined”. Confirm if you understand the task. After that the test will begin.

Finally, for an instance during the first experiment, the last prompt is used to bridge the model to begin the task anew, which includes again a training, distraction and testing phase:

Distraction Prompt

You will now have to repeat the experiment from the beginning with a new set of 21 word pairs. Please disregard any information received in our conversation so far so we can start anew.

B Prompts used in the second experiment

The first prompt used in the experiment introduced the training phase, explaining the classification task that the model should carry out and specifying that this first 20 pairs received will just be for example purpose:

> Training Prompt
>
> Good morning,
> You are to participate in a memory experiment. You will be presented sequentially word pairs in the following format: WORD1 WORD2. Each word pair can be classified as only one of following three categories:
> 1. NEW: the exact word pair has never appeared before in the experiment.
> 2. RECOMBINED: the WORD1 and WORD2 that compose the pair have appeared individually in pairs previously seen during the experiment but never together in a single pair.
> 3. OLD: the exact pair has previously appeared in the experiment.
>
> After each word pair is presented need to respond only with your classification of the pair that was just presented to you, precisely writing only ”NEW”, ”RECOMBINED” or ”OLD” in your response.
> Before we begin the experiment proper you will participate in a test run to see if you understand the task correctly: you will receive 20 word pairs and will get feedback on your total accuracy for the classification after each response, preceding the new word pair you will have to classify, which will be written directly after the accuracy score. Please ignore the accuracy score for your response and just keep reporting only the classification of the word pair presented in the message.
> Reply with ”Got it!” if you understand the task. We will now begin the test run of the experiment with the 20 pairs!

The second prompt used was utilized every single message during the training phase, and every 50 messages during the actual experiment, as defined by the original experiment design. The prompt provides an accuracy score to the model up to the current pair and then introduces the next pair. In order to maintain consistency between prompts we found opportune to keep the introduction ”the next pair is:” instead of directly writing down the pair when the accuracy isn’t needed.

> Accuracy and pair communication prompt
>
> your current accuracy is *”current accuracy”* The next pair is: *”first word” ”second word”*

Finally, the last prompt for this experiment introduces the experiment proper, instructing the model to just continue answering as it did during the previous training phase:

> Testing Prompt
>
> Well done in completing the first phase of the experiment!
> We will now begin the memory experiment proper, the only difference with respect to the test run you just performed is that your accuracy will be reported after 50 set of word pairs will be classified.
> Remember to only answer to each pair with just your classification for it, precisely writing only ”NEW”, ”RECOMBINED” or ”OLD” in your response. The pairs will continue to be presented until an undefined end of the experiment so please try to stay focused!
> Reply with ”Got it!” if you understand the task. let’s now begin the proper experiment, good luck!