Step-by-step analogical reasoning in humans and neural networks

***Jacob Russin (jacob_russin@brown.edu)**
Department of Computer Science
Department of Cognitive and Psychological Sciences
Brown University, Providence, RI 02912 USA

***Joonhwa Kim (joonhwa_kim@brown.edu)**
Department of Neuroscience, Brown University, Providence, RI 02912 USA

Ellie Pavlick (ellie_pavlick@brown.edu)
Department of Computer Science, Brown University, Providence, RI 02912 USA

Michael J Frank (michael_frank@brown.edu)
Department of Cognitive and Psychological Sciences, Brown University, Providence, RI 02912 USA

Abstract

Both humans and large language models (LLMs) perform better on some reasoning tasks when encouraged to think step by step. However, it is unclear whether these performance gains are based on similar principles. Testing both humans and LLMs on a novel word analogy task, we find that interference caused by semantic similarity hurts performance in both and drives humans to engage in a sequential reasoning process. These findings pave the way for investigation into the mechanisms that underlie the benefit of chain-of-thought and the decision process behind sequential thinking.

Introduction

Recent work has investigated the principles underlying the effectiveness of CoT prompting in LLMs (Merrill & Sabharwal, 2023). In computational cognitive neuroscience, sequential processing is thought to resolve interference by ensuring that conflicting representations are processed one at a time (Musslick & Cohen, 2020). This interference principle may be operative in vision-language models (Campbell et al., 2024), and could explain some of the performance gains observed with CoT prompting. Here, we investigate this hypothesis by studying whether CoT or sequential processing can mitigate these interference effects in a novel analogical reasoning task in humans and LLMs.

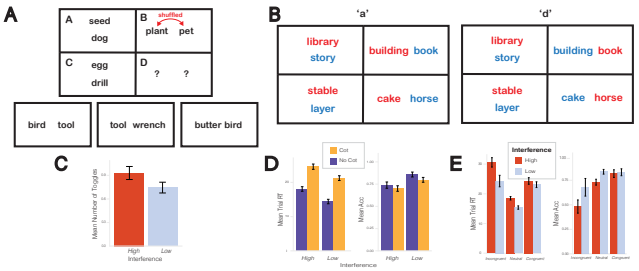

Figure 1: Human experiments.

Method

We designed a word analogy task modeled after Raven's Progressive Matrices (Raven & Raven, 2003), where we manipulated interference by controlling the semantic similarity between the words present in a given matrix. 2x2 matrices were constructed from two analogy problems, each with the form A : B :: C : D (Fig. 1A). The top-left panel of the matrix contained the two A words, the top-right contained the two B words, the bottom-left contained the two C words, and the goal was to fill in the last D panel with the two words that would complete both analogies.

While the two analogies in each matrix were independent, the two words in Panel B were randomly shuffled so that it was unclear which of the two B words belonged to which of the two analogies. This meant that each matrix had two possible *readings* — a correct reading ([A1 : B1 :: C1 : ?]; [A2 : B2 :: C2 : ?]), where the B words were correctly aligned with the A words, and an incorrect reading ([A1 : B2 :: C1 : ?]; [A2 : B1 :: C2 : ?]), where the B words were interpreted as being paired with the wrong A words. Before participants could complete a matrix, they had to infer which of these two readings was consistent with the available answers.

We hypothesized that semantic similarity would cause interference when incorrect A-B pairs were highly similar. We therefore manipulated the semantic similarity between the two sets of A-B pairs. In the low interference condition, the similarity between the correct A-B pairs was higher than that of the incorrect pairs, providing a veridical cue to the correct reading. In the high interference condition, this relationship was reversed: the similarity between the incorrect A-B pairs was higher than that of the correct pairs, making the incorrect reading of the matrix more salient (see analogies in Fig. 1B).

We hypothesized that this kind of interference would be mitigated when participants reasoned sequentially, so we manipulated the extent to which participants were encouraged to reason sequentially. We included a CoT condition where the words could be color-coded according to one of the two possible readings of the matrix (Fig. 1B). By pressing the keys ('a', 's', 'd'), participants could toggle between a *neutral* seg-

mentation where all words were shown in black, a *congruent* segmentation where the colors of words aligned with the correct reading of the matrix, and an *incongruent* segmentation where the colors aligned with the incorrect reading. It was unknown which segmentation was congruent in advance — the 'a' and 'd' keys determined colors based only on the positions of the words in panel B, which were randomized on each trial.

26 online participants (15 females, age=38.4±11.5ys) were recruited via Prolific, using a 2x2 design in which the CoT condition was varied between subjects and the Interference condition was varied within subjects. Over 60 trials, participants were given 50 seconds to toggle freely and select their answer from a set of 3 answer choices, followed by feedback. We randomly paired analogies to form matrices based on the cosine similarity of word vectors using spaCy and used these to compute an interference score. Analogy pairs with the most semantic interference (i.e., the highest positive scores) were placed in the High Interference condition. Analogy pairs in the Low Interference condition all had negative scores and were matched on frequency.

To investigate whether LLMs exhibit the interference effects observed in humans, we evaluated them on the same dataset of analogy matrix problems, given in three different conditions:

- **Single**: To establish a baseline, we evaluated the LLMs on the same analogies from the matrices, given one at a time.
- **Independent**: To isolate the effect of performing two analogies at once, we tested the LLMs in a condition where both analogies were given simultaneously but could be completed independently. In this case, the two analogies were presented sequentially rather than in a matrix format, and the B words were not shuffled, so there was no ambiguity about which B words belonged to which analogy.
- **Matrix**: To simulate the conditions experienced by the human participants, we also gave the LLMs text-based versions of the full matrix problems. In this case, the two analogies were presented in four "panels" ("Panel A: library, store; Panel B: building, book; ...") and the words in the B panel were randomly shuffled.

We used few-shot prompting with 10 examples given in context. A short preamble containing instructions about the nature of the task was included in the beginning of each prompt. We evaluated the LLMs by measuring the average log probability (Webb, Holyoak, & Lu, 2023).

Results

Participants showed slower reaction times and lower accuracy in the high interference condition ($p < .001$; Fig. 1), supporting our hypothesis that semantic similarity would cause interference. Participants also toggled between the available segmentations more often in the high interference condition compared to the low interference condition ($p < .05$), supporting our hypothesis that participants were sensitive to interference when choosing whether to toggle (Fig. 1C).

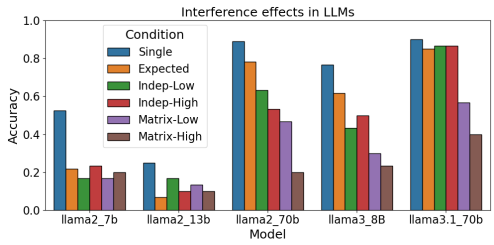

Figure 2: Interference effects in large language models.

Further analysis revealed subtler interactions with interference effects. Participants who spent more time viewing the congruent segmentation performed better than those who spent more time in the incongruent segmentation and those in the neutral condition ($p = .01$, Fig. 1E). Moreover, the effect of interference on accuracy was reduced when participants spent more time viewing the congruent segmentation (trending interaction between congruent vs incongruent segmentation time spent and interference condition; $p = 0.067$). This suggests that when participants could utilize color to consider each reading of the matrix sequentially, interference was reduced when they spent more time considering the correct reading, but interference was exacerbated when they spent more time considering the incorrect reading. While the causality of this relationship between toggling and accuracy is yet to be explored, this result validates that humans use the toggling to aid sequential processing of the interpretation they believe to be correct, and will facilitate further work in humans and LLMs that investigates the choice to sequentially process.

Preliminary LLM results are shown in Figure 2. Three of the models (Llama2-70b, Llama3-8b, Llama3.1-70b) performed particularly well on single analogies, achieving accuracies comparable to those observed in humans on the matrix problems. However, when these models were given two problems simultaneously in the Independent condition, they performed significantly worse. In two out of the three best-performing models (Llama2-70b, Llama3-8b) accuracy was worse than expected from the Single analogy accuracy, under the assumption that each of the two problems would be performed independently (compare orange and green/red bars in Fig. 2). This shows that the presence of another analogy in the problem interfered with the models' ability to do each problem in isolation, even when the two analogies were completely independent of one another. Performance was further degraded in the Matrix condition, where the format was more challenging and the B words were shuffled.

All high-performing models showed significant interference effects in this condition, performing better in the Low Interference condition than in the High Interference condition, consistent with the interference effects we found in humans on the same problems. Some preliminary experiments were performed using CoT prompting, but the results were inconclusive (not shown). For example, when models were prompted to generate each possible reading of a given matrix before an-

swering, they performed worse than without such prompting. Further experimentation is required to understand whether other kinds of CoT prompts would mitigate the interference effects observed in the Matrix condition, how these specific types of chaining parallel that of goal-directed attention in humans, and the mechanisms underlying the benefit of chaining.

Conclusion

Our experiments testing humans and LLMs on a novel word matrix task show that interference can disrupt analogical reasoning and can also drive engagement in sequential processing in humans. The effectiveness of sequential processing in mitigating interference in humans was related to what was attended during the step-by-step reasoning process: interference was reduced when the right relationships were isolated but exacerbated when misleading connections were prioritized. This dynamic may also explain why we could not immediately improve LLM performance on the same problems with explicit CoT prompting. Although our findings are preliminary, they provide some evidence that the benefits of step-by-step reasoning in humans interact with the key principle of interference, and will facilitate further investigation of the mechanisms that underlie when and how chaining benefits LLMs.

References

Campbell, D., Rane, S., Giallanza, T., Sabbata, N. D., Ghods, K., Joshi, A., ... Webb, T. W. (2024, October). *Understanding the Limits of Vision Language Models Through the Lens of the Binding Problem* (No. arXiv:2411.00238). arXiv.

Merrill, W., & Sabharwal, A. (2023, October). *The Expressive Power of Transformers with Chain of Thought* (No. arXiv:2310.07923). arXiv. doi: 10.48550/arXiv.2310.07923

Musslick, S., & Cohen, J. D. (2020, November). *Rationalizing Constraints on the Capacity for Cognitive Control.* doi: 10.31234/osf.io/vtknh

Raven, J., & Raven, J. (2003). Raven Progressive Matrices. In *Handbook of nonverbal assessment* (pp. 223–237). New York, NY, US: Kluwer Academic/Plenum Publishers. doi: 10.1007/978-1-4615-0153-4$_1$1

Webb, T., Holyoak, K. J., & Lu, H. (2023, July). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526–1541. doi: 10.1038/s41562-023-01659-w