UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

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

Permalink
https://escholarship.org/uc/item/1n6366h7

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors
Russin, Jacob
Kim, Joonhwa
Pavlick, Ellie
et al.

Publication Date
2025

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

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

Jacob Russin* **(jake_russin@brown.edu)**
Department of Computer Science
Department of Cognitive and Psychological Sciences
Brown University

Joonhwa Kim* **(joonhwa_kim@brown.edu)**
Department of Neuroscience
Brown University

Ellie Pavlick
Department of Computer Science
Brown University

Michael J. Frank
Carney Institute for Brain Science
Department of Cognitive and Psychological Sciences
Brown University

Abstract

Both humans and large language models (LLMs) perform better on some reasoning tasks when they are encouraged to think step by step. However, it is unclear whether these performance gains are based on similar principles. In this work, we investigate two hypotheses: (1) that these benefits arise due to the presence of local statistical structure in the training data, where intermediate steps of reasoning may be common but any specific reasoning trajectory is rare, and (2) that sequential processing improves reasoning by mitigating interference. Using LLMs and transformers trained on a synthetic dataset, we show how analogical distance effects previously observed in humans and LLMs may be explained by the presence of local statistical structure. Testing both humans and LLMs on a novel word analogy task, we find that interference caused by semantic similarity can hurt performance and drives humans to engage in a sequential reasoning process. Our findings show that both locality structure and interference may be key principles underlying the benefits of step-by-step thinking.

Keywords: analogy; large language models; chain-of-thought; sequential reasoning; interference

Introduction

Large language models have recently shown impressive reasoning performance in a number of domains, including analogy (Webb, Holyoak, & Lu, 2023; Musker, Duchnowski, Millière, & Pavlick, 2024), scientific reasoning (Luo et al., 2024), and logical or mathematical reasoning (Sprague et al., 2024). Much of this recent progress can be attributed to chain-of-thought (CoT) prompting (Wei et al., 2023; Kojima, Gu, Reid, Matsuo, & Iwasawa, 2023), where models are prompted to "think step by step" or given a few examples of how to do so. Humans also benefit from thinking step by step or reasoning sequentially in many domains (O'Reilly, Nair, Russin, & Herd, 2020), including analogy (Duncan, Chylinski, Mitchell, & Bhandari, 2017), and recent studies suggest that humans and LLMs may benefit from step-by-step thinking in some of the same domains (Liu et al., 2024). However, it is unclear whether the performance gains observed in both humans and LLMs are mediated by shared principles.

Recent work has investigated the principles underlying the effectiveness of CoT prompting in LLMs (Merrill & Sabharwal, 2023). One reason may be the presence of local statistical structure in the training distribution (Prystawski, Li, & Goodman, 2023): these models generalize poorly when they have to perform inference over variables that rarely co-occur, but can do so when prompted to generate intermediate variables before predicting the answer. This can help because one may be familiar with each individual reasoning step even though the specific trajectory from start to finish is unfamiliar.

This locality principle may explain why CoT prompting improves LLM performance in domains like math, where the space of all possible problems is combinatorially large (such that any individual problem will be rare in training), but the suite of reasoning steps required to solve them is relatively small (and common in training). This principle may also relate to the observation that both humans (Jones, Kmiecik, Irwin, & Morrison, 2022) and LLMs (Webb et al., 2023) exhibit analogical distance effects, performing better on "near" analogies (e.g., nose : smell :: tongue : taste) with words from the same semantic domain that are likely to co-occur in training, compared to "far" analogies (e.g., nose : smell :: antenna : signal) with words from different domains.

Research in computational neuroscience has emphasized that sequential processing in humans can help to resolve interference by ensuring that conflicting stimuli or 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.

These two principles are not necessarily mutually exclusive, and each may help to explain how sequential processing can facilitate reasoning. Here, we investigate both hypotheses in the domain of analogical reasoning in experiments with both humans and LLMs. First, we use LLMs and neural networks trained on a synthetic analogical reasoning dataset to study whether local statistical structure can explain the analogical distance effects that have been observed in both humans (Jones et al., 2022) and LLMs (Webb et al., 2023). Second, we study interference effects in a novel analogical reasoning task in humans and LLMs, and investigate whether CoT or sequential processing can mitigate this interference.

Analogical distance and locality

Neural networks trained on a synthetic task

To investigate whether the presence of local statistical structure can explain analogical distance effects, we trained transformer models from scratch on a synthetic analogical reasoning dataset (see Figure 1). In our synthetic task, each analogy

*Equal contribution

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

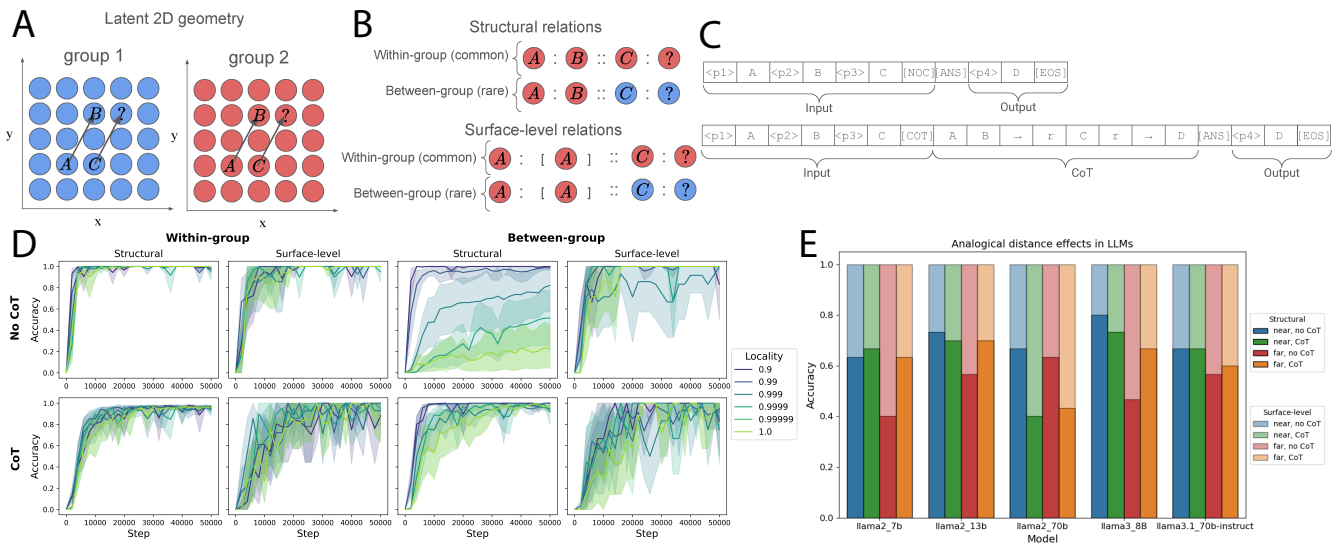

Figure 1: Analogical distance experiments. **A**. Latent 2D geometry in synthetic analogical reasoning task. Tokens were assigned to unique spatial positions in a latent 2D grid. Two example analogies of the form A : B :: C : D are shown. **B**. Tokens were separated into two groups. Analogies with tokens from the same group were much more frequent in the training data than those with tokens from different groups. In addition to the structural relations requiring knowledge of the latent 2D structure, surface-level relations simply required models to copy and add two special tokens (brackets). **C.** Example sequences from the synthetic dataset with and without chains of thought (CoT). **D**. Results of training a transformer neural network on the synthetic dataset. Models performed well on within-group analogies (left), regardless of whether the analogies tested structural relations or surface-level relations, and regardless of chain-of-thought (CoT) prompting. Models struggled on between-group structural analogies when these were rare in the training data ($\lambda > 0.999$), but performed well on between-group surface-level analogies (right-most column), and when they were prompted to produce chains of reasoning (bottom row). **E**. Analogical distance effects in large language models (LLMs). Models consistently exhibited analogical distance effects (blue vs red bars), performing better on near than far analogies. All LLMs performed well on surface-level relations (light colors). CoT prompting tended to improve performance on far analogies (orange vs red bars).

was given in a four-word proportional format (A : B :: C : D), where models were required to predict D.

Although the "words" that appeared in these synthetic problems were arbitrary tokens, we assumed a latent similarity structure governed by a fixed 2D geometry. Each token was assigned to one of 25 spatial locations in a 5x5 2D grid. Information about these locations was not made explicitly available to the model, which instead had to learn this latent geometry through training. The relation between a pair of tokens was defined to be the difference between their corresponding spatial positions ($\vec{r} = \vec{b} - \vec{a}$). An analogy was defined to be valid if the two pairs of tokens had the same spatial relation ($\vec{d} - \vec{c} = \vec{r} = \vec{b} - \vec{a}$). Thus, if the latent spatial position of each token was learned by the model, then analogies could be completed by extracting the structural relation between A and B and applying this relation to C ($\vec{r} = \vec{b} - \vec{a} \rightarrow \vec{d} = \vec{c} + \vec{r}$). We introduced local statistical structure into the data distribution by assigning each token to one of two groups. Tokens within the same group were observed in analogies much more frequently (λ) than tokens in different groups (1 - λ).

We also introduced a special "surface-level" relation that did not require knowledge of the latent structure of the grid. This special relation consisted of a simple string operation where additional tokens such as brackets were added before and after the target token (e.g., A : [A] :: C : [C]). This operation can be considered surface-level in the sense that the model has direct access to the brackets in the input, whereas the latent 2D relations would have to be learned through experience. The relation allowed us to test whether co-occurrence alone was sufficient for models to show a benefit of CoT, or whether they could generalize in some cases even with words that rarely co-occurred.

To allow testing with and without CoT, we introduced a CoT condition during training in which models saw the same problems followed by a special [COT] token, indicating that they had to produce a chain of the form: A B $\rightarrow$ r ; C r $\rightarrow$ D, where A, B, C, and D are the tokens from the original problem, and r is a token corresponding to the relation between A and B (see Figure 1C). A unique relation token was assigned to every possible linear relation (i.e., every possible difference vector) in the 5x5 grid. We also included separate trials where the models were only required to apply a given relation to a given C token, encouraging them to learn to utilize the r tokens generated during the CoT.

158,952 total problems were generated and 798 of these were held out for testing. 10% of the training data included

CoT reasoning. The locality parameter λ was varied from 0.9 to 1.0 to investigate the effects of a small number of between-group samples violating the locality constraint.

A standard transformer architecture was used for training (Touvron et al., 2023; Vaswani et al., 2017). The model had 12 layers, 8 attention heads, a hidden size of 64, a feedforward size of 128, and dropout = 0.1. 10 random initializations were trained using a cross-entropy loss and Adam optimizer with a learning rate of 0.001. Models were trained with a batch size of 256 for 50,000 steps.

Results The models performed well on held-out within-group analogies, showing they had learned the latent structure of the grid (see Figure 1D). However, when between-group analogies were rare during training ($\lambda > 0.999$), the models struggled to generalize this learned structure between groups. This is consistent with the finding that humans and LLMs perform worse on far analogies (Jones et al., 2022; Webb et al., 2023), where words come from different semantic domains and are therefore less likely to co-occur. However, the models performed well on between-group analogies involving surface-level relations, suggesting that analogical distance effects are not relevant when surface-level features can be used to extract relations.

Consistent with previous results investigating CoT benefits in LLMs more generally (Wei et al., 2023; Sprague et al., 2024), the models performed well on between-group analogies involving structural relations when they were prompted to produce a chain of reasoning. This shows that even on problems requiring latent structure to be extracted from tokens that rarely co-occurred during training, models were capable of completing analogies when reasoning step by step.

Analogical distance in large language models

We also tested whether similar phenomena emerge in LLMs trained on natural text by evaluating them on word analogy problems while manipulating analogical distance and CoT prompting. Word analogies were taken from a previous study that found an analogical distance effect in humans (Jones et al., 2022). Webb et al. (2023) showed that this effect was reproduced by large language models on the same problems, but did not systematically examine CoT prompting.

We tested five LLMs of different sizes from the Llama family (Touvron et al., 2023). These models were evaluated on 60 four-word proportional analogy problems given in the form A : B :: C : D. Half of the problems were far analogies (e.g., caffeine : drug :: rose : flower) and half were near (e.g., murder : crime :: gun : weapon). The original study did not manipulate whether the analogies were governed by surface features or structural relations, so we again augmented the dataset with character-level relations where the same A and C words from the original problems were surrounded by specific symbols such as brackets (e.g., accountant : [accountant] :: carrot : [carrot]). This again allowed us to manipulate the extent to which surface features could be used to complete the analogy, independent of analogical distance.

A short preamble was included at the beginning of each prompt: "Here are some word analogy problems of the form A : B :: C : D ('A is to B as C is to D')." We tested models in the few-shot learning setting, where 10 random problems and their answers were selected to be included in the prompt.

In the CoT condition, the models were additionally prompted to "Think step-by-step, showing your reasoning.", and each of the few-shot examples included an extra sentence describing the relation between A and B, before providing the answer (e.g., "Accountant is a kind of profession. What is a carrot a kind of? Answer: vegetable"). Models had to produce this reasoning themselves in the problems on which they were actually tested. Following Webb et al. (2023), the models were evaluated by measuring the average log probability assigned to the tokens in each answer choice.

Results The results for all models are shown in Figure 1E. All five models consistently performed better on near structural analogies compared to far structural analogies, reproducing the basic effect found by Webb et al. (2023).

While the results were mixed across the various models we tested, CoT prompting appeared to improve performance specifically on the far analogies, although the effect varied across the five models, and was reversed in Llama2-70b. Although it is difficult to draw any strong conclusions due to the variance across models, this trend was consistent with our hypothesis that producing chains of reasoning before generating a final answer would allow the models to generalize to pairs of words from different semantic domains, which may have been less likely to co-occur during training.

Finally, all five models performed nearly perfectly on all the surface-level relations we tested, regardless of analogical distance (see Figure 1E, lighter bars). This suggests that CoT is not required on analogy problems where surface-level features (in this case, token-level differences like additional brackets) form the basis of a relation, even when the words themselves are from different semantic domains.

Overall, our experiments with LLMs and neural networks trained on our synthetic task suggest that locality structure may explain analogical distance effects and the benefits of CoT prompting for certain kinds of analogical reasoning.

Interference in analogical reasoning

Interference effects in humans

Next, we investigated whether sequential reasoning or CoT-like processing is important for mitigating interference during analogical reasoning in humans and in LLMs. To do this, we designed a word analogy task modeled after Raven's Progressive Matrices (Raven & Raven, 2003), where we could manipulate 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 (see Figure 2A). 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

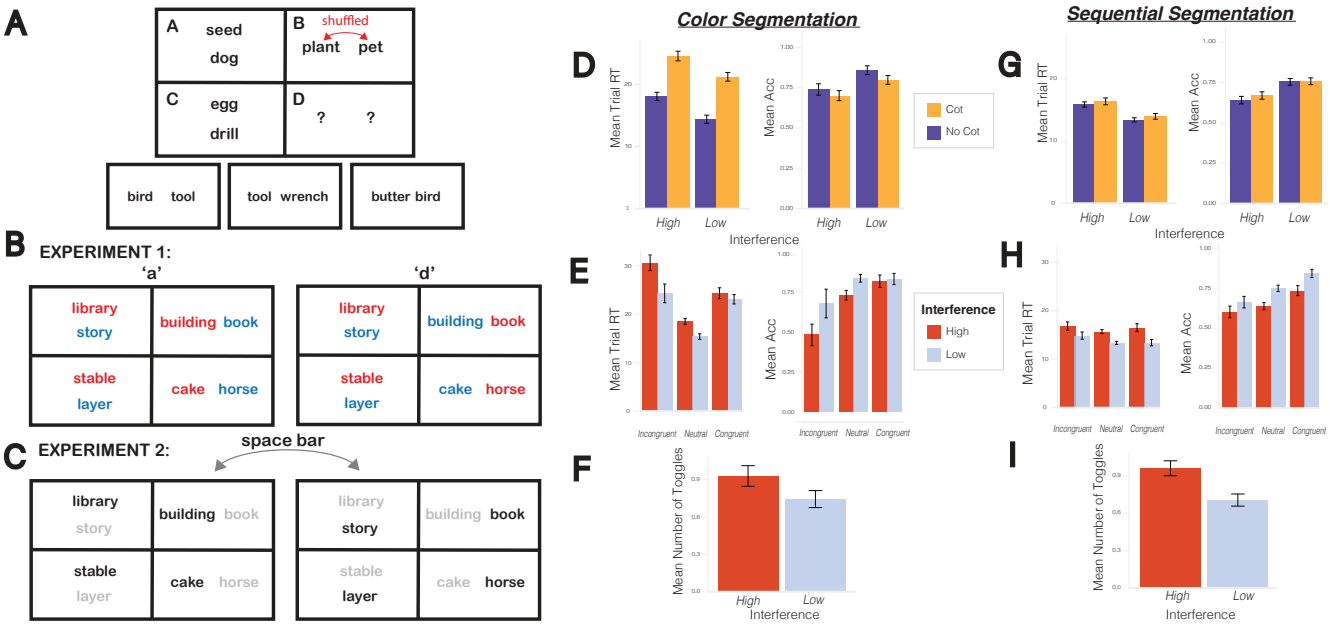

Figure 2: Human experiments. **(A)** 2x2 word analogy matrix problems were constructed by pairing two independent analogies. Both analogies had the form A : B :: C : D, where the A word from each of the two analogies appeared in the top left panel, the B words appeared in the top-right panel, etc. The task was to select the option for panel D that best completed both analogies simultaneously. The order of the two words in panel B was randomly shuffled, such that their respective pairings with the words from panel A must be inferred from the available answer choices. The example shown here is from the low interference condition. The answer is the first option: [bird tool]. **(B)** In Experiment 1, half of the participants were assigned to the Color Segmentation condition, in which they were trained to use buttons on the keyboard to manipulate the colors of words. Participants could toggle between two different color segmentations of the matrix using the 'a' and the 'd' keys — an incongruent segmentation (left side) where the colors were aligned with the incorrect reading, and a congruent segmentation (right side), where the colors were aligned with the correct reading. This example is from the high interference condition. Correct answers are shown in panel D for illustration purposes. **(C)** In Experiment 2, a Sequential Segmentation condition was included where participants could again toggle between two different interpretations of the matrix (only one shown), but could manipulate the presence of words on the screen rather than their colors. The space bar was used to alternate between the two analogies under a given segmentation. Here, words are shown in gray for illustration purposes, but were not visible at all during the experiment. **(D, G)** Accuracy and reaction time (RT) results for each condition in each experiment. **(E, H)** Accuracy and RTs split by whether more time was spent in the incongruent, neutral, or congruent segmentation. **(F, I)** Participants toggled between the two segmentations more frequently in the high interference condition.

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 the participants could complete a matrix, they had to infer which of these two readings was consistent with the available answers — that is, which pairs of words in the A and B panels could be related to get the correct answer. The answer choices were carefully constructed so that the problems were unambiguous.

We hypothesized that semantic similarity would play an important role in this inferential process, causing interference when the 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. In this condition, semantic similarity provided a veridical cue to the correct reading of the matrix. 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. In this case, semantic similarity was misleading, making the incorrect reading of the matrix more salient.

We hypothesized that this kind of interference would be mitigated when participants reasoned sequentially. We tested this hypothesis by manipulating the extent to which partici-

pants were encouraged to reason sequentially, in two different ways. In **Experiment 1**, we included a **Color Segmentation** condition, where the words in the matrix were color-coded according to one of the two possible readings of the matrix (see Figure 2B). By pressing special keys ('a', 's', 'd') on their keyboards, participants could toggle (using the 'a' and 'd' keys) between a *neutral* segmentation 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. The participants did not know which reading was correct 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.

In **Experiment 2**, we included a **Sequential Segmentation** condition, where individual analogies could be viewed one at a time according to one of the two readings of the matrix (see Figure 2C). As in the color segmentation condition, the 'a' and 'd' keys could be used to toggle between the two possible segmentations. However, because only one analogy was shown at a time, the space bar was additionally required to cycle between the two analogies within a given segmentation. Again, the 'd' key allowed a neutral view of the matrix, where all words were present on the screen simultaneously.

We hypothesized that each of these two conditions would encourage participants to reason sequentially through the two possible readings of the matrices, mitigating the interference caused by semantic similarity. In both experiments, we included a **Neutral** condition, where all the words in each matrix were always presented simultaneously in black and no toggling was required.

Word analogies were gathered from publicly available datasets[1], including college-level SAT analogy problems. In order to minimize the impact of vocabulary, we excluded analogies with especially low frequency words. We then randomly paired analogies to form matrices based on the cosine similarity ($\mathrm{s}(\cdot)$) of word vectors using spaCy[2]. For every possible pair of analogies, we computed an interference score:

$$\text{score} = \mathrm{s}(a_1, b_2) + \mathrm{s}(a_2, b_1) - (\mathrm{s}(a_1, b_1) + \mathrm{s}(a_2, b_2)) \quad (1)$$

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 and the datasets from which the analogies originated. We then removed any pairs we thought were ambiguous or otherwise inappropriate for the study and added distractor words used to create the answer choices. Our final dataset had 60 total matrix problems (30 for each condition).

26 participants (15 females, 11 males; age = 38.4±11.5 ys) for Experiment 1 and 54 participants (33 females, 21 males; age = 36.7±9.8 ys) for Experiment 2 were recruited via Prolific in online studies. In both experiments, we utilized a 2x2 design in which the CoT condition was varied between subjects and the Interference condition was varied within subjects. In each experiment, the task consisted of 60 trials where participants could toggle freely before selecting their answer from a set of 3 answer choices. In the Color Segmentation task, trials began in the neutral reading, but in the Sequential Segmentation task, trials were randomly initialized in one of the two segmentations to encourage participants to toggle. Participants were given 50 seconds to select an answer, and were provided with feedback after their selection was made.

Results The results from both experiments are shown in Figure 2D-I. In both experiments, participants showed slower reaction times (RTs) and lower accuracy in the high interference condition ($p < .001$; $p < .001$). This was consistent with our hypothesis that semantic similarity would cause interference and hurt performance.

Reaction times (RTs) were also slower in both the Color Segmentation and Sequential Segmentation conditions compared to the Neutral condition, suggesting that both of our manipulations encouraged participants to slow down and approach the problems in a step-by-step manner. However, this slowing did not result in improved accuracy in either the Color Segmentation or the Sequential Segmentation conditions compared to the Neutral condition, contrary to our original hypothesis. Furthermore, we found no significant interaction between either of these manipulations and the Interference condition in predicting accuracy. However, we did find that participants toggled between the available segmentations more often in the High Interference condition compared to the Low interference condition ($p = .047$; $p = .049$), suggesting that participants were sensitive to interference when choosing whether to toggle between readings (see Figure 2F and 2I).

Further analysis revealed that these two conditions interacted with interference effects in subtler ways. In both of the CoT conditions, participants who spent more time viewing the congruent segmentation, where the presentation of words was aligned with the correct reading of the matrix, performed better than those who spent more time in the incongruent segmentation and those in the neutral condition ($p = .011$; $p = .010$; see Figure 2E and 2H). In Experiment 1, we found a trending interaction between the time spent viewing the congruent vs incongruent segmentation and the interference condition, where the effect of interference on accuracy was reduced when participants spent more time viewing the congruent segmentation ($p = 0.067$). 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. In Experiment 2, the effect of interference was not exacerbated when participants spent more time in the incongruent segmentation, possibly due to the nature of the Sequential Segmentation condition, where words were only shown from one analogy at a time. Further investigation is needed to fully clarify the differences between these two

[1] https://github.com/asahi417/AnalogyTools

[2] https://spacy.io/

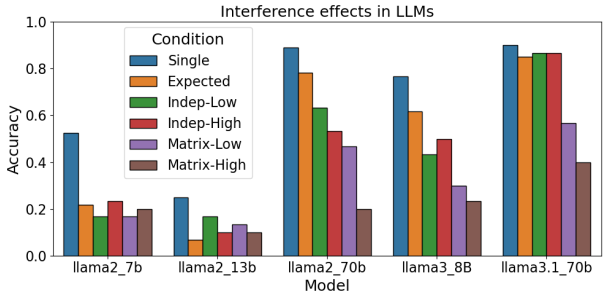

Figure 3: Interference effects in large language models.

ways of encouraging step-by-step thinking.

These preliminary findings suggest that semantic similarity can cause interference effects in analogical reasoning. When given the opportunity to manipulate the external presentation of stimuli to facilitate step-by-step thinking, participants chose to do so more often in the presence of greater semantic interference. However, the efficacy of this step-by-step reasoning process in mitigating interference depended on participants' viewing choices: in the Color Segmentation condition, interference was ameliorated when the right relationships were isolated, but was amplified when misleading semantic connections were highlighted.

Interference in large language models

To investigate whether LLMs exhibit the interference effects we observed in humans, we evaluated them on the exact same dataset of analogy matrix problems. Problems were given to the LLMs 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.

Results Preliminary results are shown in Figure 3. 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 Figure 3). 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. Furthermore, all high-performing models showed significant interference effects in this condition, performing better in the Low Interference condition than in the High Interference condition. This is 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 answering, they performed significantly 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.

Discussion

Although our findings are preliminary, they provide some evidence that the benefits of step-by-step reasoning in both humans and LLMs are related to two key principles: locality structure and interference. Our experiments using LLMs and neural networks trained on our synthetic dataset suggest that the analogical distance effects observed in humans (Jones et al., 2022) and in LLMs (Webb et al., 2023) may be explained by local statistical structure: CoT prompting improves performance specifically on far analogies where words come from different domains and are therefore unlikely to co-occur.

Our experiments testing humans and LLMs on a novel word matrix task show that while interference can also disrupt analogical reasoning, its interplay with sequential reasoning is more nuanced. In humans, the effectiveness with which sequential processing can mitigate this interference seems to depend on what is 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. Taken together, our findings highlight how local statistical structure and interference may be key factors contributing to the benefits of step-by-step reasoning in both humans and neural networks.

Acknowledgments We would like to thank all members of the Language Understanding and Representation Lab and the Laboratory of Neural Computation and Cognition at Brown University, as well as Taylor Webb and Declan Campbell for helpful discussions. MJF was supported by ONR grant N00014-23-1-2792. EP and JR were supported by NIH NIGMS COBRE grant #5P20GM103645-10.

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.

Duncan, J., Chylinski, D., Mitchell, D. J., & Bhandari, A. (2017, May). Complexity and compositionality in fluid intelligence. *Proceedings of the National Academy of Sciences*, *114*(20), 5295–5299. doi: 10.1073/pnas.1621147114

Jones, L. L., Kmiecik, M. J., Irwin, J. L., & Morrison, R. G. (2022, August). Differential effects of semantic distance, distractor salience, and relations in verbal analogy. *Psychonomic Bulletin & Review*, *29*(4), 1480–1491. doi: 10.3758/s13423-022-02062-8

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023, January). *Large Language Models are Zero-Shot Reasoners* (No. arXiv:2205.11916). arXiv. doi: 10.48550/arXiv.2205.11916

Liu, R., Geng, J., Wu, A. J., Sucholutsky, I., Lombrozo, T., & Griffiths, T. L. (2024, November). *Mind Your Step (by Step): Chain-of-Thought can Reduce Performance on Tasks where Thinking Makes Humans Worse* (No. arXiv:2410.21333). arXiv.

Luo, X., Rechardt, A., Sun, G., Nejad, K. K., Yáñez, F., Yilmaz, B., … Love, B. C. (2024, November). Large language models surpass human experts in predicting neuroscience results. *Nature Human Behaviour*. doi: 10.1038/s41562-024-02046-9

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

Musker, S., Duchnowski, A., Millière, R., & Pavlick, E. (2024, June). *Semantic Structure-Mapping in LLM and Human Analogical Reasoning* (No. arXiv:2406.13803). arXiv. doi: 10.48550/arXiv.2406.13803

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

O'Reilly, R. C., Nair, A., Russin, J. L., & Herd, S. A. (2020). How Sequential Interactive Processing Within Frontostriatal Loops Supports a Continuum of Habitual to Controlled Processing. *Frontiers in Psychology*, *11*. doi: 10.3389/fpsyg.2020.00380

Prystawski, B., Li, M. Y., & Goodman, N. D. (2023, November). *Why think step by step? Reasoning emerges from the locality of experience* (No. arXiv:2304.03843). arXiv. doi: 10.48550/arXiv.2304.03843

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_11

Sprague, Z., Yin, F., Rodriguez, J. D., Jiang, D., Wadhwa, M., Singhal, P., … Durrett, G. (2024, September). *To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning* (No. arXiv:2409.12183). arXiv. doi: 10.48550/arXiv.2409.12183

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., … Scialom, T. (2023, July). *Llama 2: Open Foundation and Fine-Tuned Chat Models* (No. arXiv:2307.09288). arXiv. doi: 10.48550/arXiv.2307.09288

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is All you Need. In I. Guyon et al. (Eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA* (pp. 5998–6008).

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

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., … Zhou, D. (2023, January). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* (No. arXiv:2201.11903). arXiv. doi: 10.48550/arXiv.2201.11903