Humans learn proactively in ways that language models don't

Simon Jerome Han (sjeromeh@stanford.edu)
Department of Psychology, Stanford University

James L. McClelland (jlmcc@stanford.edu)
Department of Psychology, Stanford University

Abstract

Do large language models (LLMs) learn like people do? We investigate this question with a simple task that compares human learning and LLM finetuning on the same set of novel inputs. We find that while humans learn and generalize robustly, finetuned LLMs largely fail to generalize from what they learned and are more influenced by prior expectations than humans are. We then analyze human solutions of our task and find that stronger performance is characterized by the proactive formation of efficient representations that aid learning and generalization. Although LLMs can use in-context learning to match the performance of humans who do not form these representations, and can use similar representations provided in-context to match the performance of those who, they do not form these representations on their own. Given these findings, we then consider how future theories of human learning might be built in the age of LLMs

Keywords: large language models; learning and generalization; finetuning; in-context learning; reversal curse

Introduction

Since the invention of the Transformer architecture, large language models (LLMs) have made remarkable progress towards capturing human-like, language-based intelligence. Across a wide variety of tasks ranging from syntax judgments (J. Hu, Mahowald, Lupyan, Ivanova, & Levy, 2024) to analogical reasoning (Webb, Holyoak, & Lu, 2023), LLMs now match or exceed average human performance. Such is the nature of their triumph that some have even begun to suggest that LLMs may soon become genuine examples of general intelligence (Bubeck et al., 2023), or unified theories of human cognition (Binz et al., 2024).

These developments raise the question of whether LLMs can serve as models of not only human inference, but also of human learning. Here, we are particularly interested in whether LLMs can capture the generalizability of human learning. By this we refer to the notion that humans are able to transform highly specific learning inputs into abstract representations of meaning that can be flexibly redeployed in new contexts much later in time – that after learning just one new sentence such as "X is bigger than Jupiter", we can expect humans to later not only be able to answer the question "is X bigger than Jupiter", but to also make reverse and transitive generalizations in response to other questions such as "is Jupiter smaller than X", or "is X larger than the Earth".

The question of whether LLMs can achieve this important quality of human learning remains an active area of research. In the case of pretraining, it cuts deeply into debates about whether LLMs are simply memorizing their training data (Bender, Gebru, McMillan-Major, & Shmitchell, 2021) or if they are indeed capable of truly generalizing beyond it (Lotfi et al., 2023; Lindsey et al., 2025). In the case of finetuning, existing evidence suggests that finetuned LLMs can go beyond pure memorization to a certain degree (Cohen, Biran, Yoran, Globerson, & Geva, 2024), but that they often fail to generalize in surprising and basic ways. For example, finetuned LLMs struggle to make even the simple reverse and transitive generalizations highlighted above (Berglund et al., 2023; Zhong, Wu, Manning, Potts, & Chen, 2023).

But how does LLM learning actually compare to human learning? Missing from our understanding is a concrete comparison where humans and LLMs each are each tasked with learning from the same inputs. Such a comparison is crucial if we are to make any conclusions about the degree of similarity between the two. For example, while we expect humans to be able to make reverse and transitive generalizations, human learning is also notoriously fragile, and even seemingly simple matters such as whether humans can generalize in a backwards direction during association learning (Wolford, 1971; Kahana, 2002) or digit recall (St Clair-Thompson & Allen, 2013) have led to extended investigations. It may well be the case that humans and LLMs display common failure patterns when asked to learn and generalize from the same inputs, even if these patterns are not as pronounced in humans.

Implementing such a comparison is impossible for pretraining but easy for finetuning; thus, here we introduce a simple language-based learning task that compares how human learning and LLM finetuning perform across a common set of learning items. We examine their learning trajectories and their abilities to generalize to novel test items that are semantically implied by the learned items, and we find that humans are more sample-efficient than LLMs during learning and more robust at generalizing during testing. Building on this, we consider the strategies that humans use in our task, and we find that our most successful human participants restructure their representations of the presented items in ways that support both learning and generalization, even though they were not instructed to do so. Finally, we find that LLMs can match human performance only if they are provided with similar representations in-context, and we conclude by considering the implications of these results for future theories of human learning.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

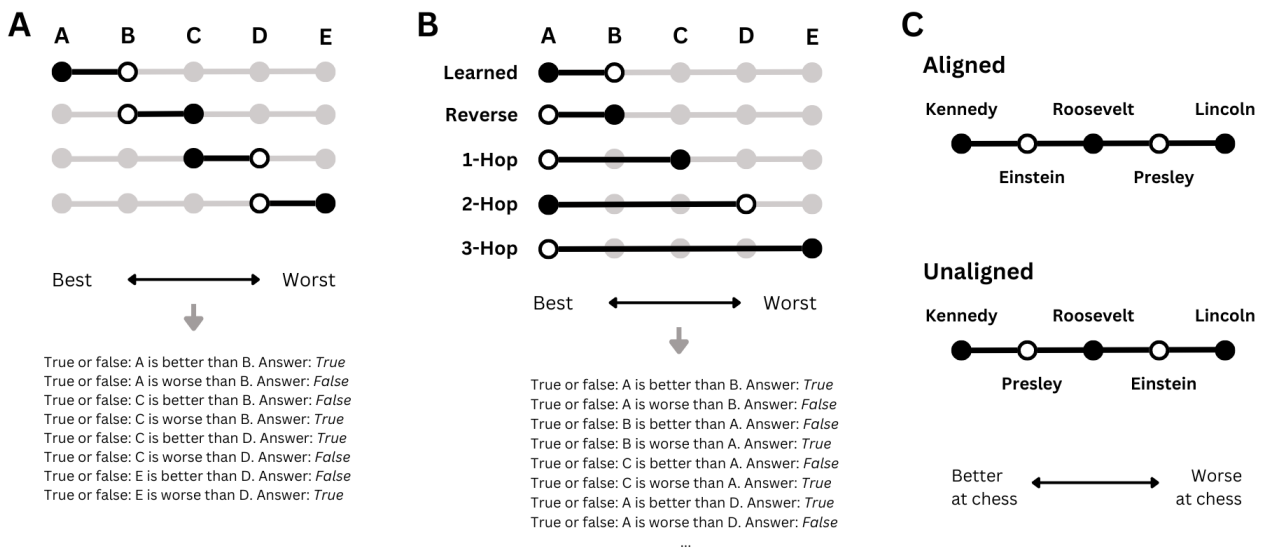

Figure 1: **The CRT task design.** A) In the learning stage, participants learn about neighboring entities from a ranked chain via a true/false cloze task. For each pair of entities, the order in which they appear in the cloze sentences is fixed. B) In the testing stage, participants make inferences about all possible pairs and orders of entities in the chain. These include the original learning stage items, as well as reverse and transitive 'n-hop' items. C) In the aligned condition, the second ranked entity is Einstein and the fourth ranked entity is Presley, while in the unaligned condition the second is Presley and the fourth is Einstein. All other entities are randomly shuffled across participants.

The Chain Rank Task

To compare human learning and LLM finetuning, we introduce the Chain Rank Task (CRT). The CRT is a simple task that is divided into a learning stage and a testing stage. During the learning stage, participants learn about how a ranked chain of entities relate to one another along a known dimension (Figure 1C). Participants learn about this chain from individual sentences describing the relations of neighboring pairs of entities, where the entities for each pair are always shown in the same order within each sentence (Figure 1A). These sentences resemble a cloze task where the participant's goal is to learn whether the correct completion for each sentence is 'true' or 'false' (e.g. "True or false: X is greater than Y. Answer:"); this allows the CRT to be presented to humans and LLMs with maximally similar inputs and outputs.

During the testing stage, participants are required to make judgments about every combination of entity pairs from the underlying chain (Figure 1B), again by responding to true/false cloze sentences. These sentences include both the original learning items and new items that require participants to make either reverse (`A > B` $\rightarrow$ `B < A`) or transitive (`A > B` $\wedge$ `B > C` $\rightarrow$ `A > C`) generalizations about what they learned. We refer to transitive generalizations as n-hop, where "n" represents the number of intermediate entities between two items in a chain. Our learning items are formulated such that all transitive generalizations include at least one reverse generalization.

To populate our chains, we use a set of five American historical figures, "Einstein", "Presley","Lincoln", "Roosevelt", and "Kennedy", and a single relation, "better/worse at chess." For each chain we randomly assign the three presidents to the first, third, and fifth positions, and the two non-presidents to the second and fourth positions. This allows for a maximally fair comparison between humans and LLMs: while both are familiar with the entities and relation, there is no existing ground truth that dictates how their chess-playing abilities should compare. Humans and LLMs also begin the task with prior expectations that differ from the rankings that they must learn in the study, ensuring that they are each faced with a genuine learning challenge.

This set of entities and relations also allows us to assess the effect that prior knowledge has on learning and generalization. Our chains have two configurations for non-presidents: one in which Einstein is ranked higher than Presley, and another in which Presley is ranked higher than Einstein. Participants who are presented with the latter configuration must therefore learn an underlying chain that diverges more from prior expectations than those who are presented with the former, even though the individual learning items diverge equally from prior expectations in either configuration. We refer to chains that abide by the former configuration as being in the "aligned" condition, and chains that abide by the latter configuration as being in the "unaligned" condition.

In sum, our task consists of five entities arranged in one of twelve combinations of hypothetical chess-playing ability. Each chain generates four pairs of neighboring entities used for the learning stage and ten pairs of entities used for the testing stage. In turn, these entity pairs become eight items for the learning stage ("True or False: X is better than Y" and "X is worse than Y" for each X/Y pair) and forty items for the testing stage ("X is better than Y", "X is worse than Y", "Y is better than X", "Y is worse than X" for each X/Y pair).

CRT for Humans

To assess human performance on the CRT, we recruited 72 US-based participants from Prolific. Six participants were randomly assigned to each of the twelve chains, ensuring balanced participant numbers between both alignment conditions and chain configurations. Participants were compensated with $0.20 for completing a screening task and a base rate of $9 for completing the main task, each described below.

Prescreening. Each participant was pre-screened a day before the main experiment to ensure that they were familiar with the entities in the experiment. The pre-screening survey required them to correctly identify the primary occupation of each entity from a multiple-choice format with eight options. Of the 200 participants that we screened, we excluded four who scored less than 5 out of 5 on the survey, leaving a pool of 196 participants from which we randomly sampled 72.

Learning stage. During the learning stage, participants sequentially viewed each learning item and responded by pressing 't' or 'f' to fill an input box with 'True' or 'False' before pressing the enter key to proceed to the next item. Correct answers were indicated by the input box turning green, while incorrect answers were indicated by the input box turning red and displaying the correct answer. Responses were self-paced during the response component, but the feedback component was limited to 1.5 seconds per item.

Participants completed twenty repetitions of the eight learning items. The items in each repetition were shuffled such that no two consecutive learning items contained overlapping entities. Given that there were only four underlying entity pairs, this was made possible by including one of two 'filler' items representing a fifth pair of entities ('Jefferson' versus 'Lincoln'), both unconnected to the learned chain, at the middle and end of each repetition. To encourage deeper encoding of the learning items, participants were also interrupted every four repetitions with a set of 10 simple arithmetic questions, also presented in true/false format. At the end of the learning stage, participants then answered 20 additional arithmetic questions to create a longer break between the learning and testing stages. Over the entire learning phase, participants therefore answered a total of 260 true/false questions. As an incentive, participants earned a bonus of 1 cent for each learning item that they answered correctly.

Testing stage. In the testing stage, items were presented in the same format as the learning stage, but no feedback was given. Participants viewed the 40 test items in a random order that was constrained by a balanced Latin square design that ensured that each of the four occurrences of any given entity pair occurred once across every block of ten items. Participants completed two cycles of testing, resulting in 80 test item responses in total. As an incentive, a bonus of 2 cents was awarded for each correctly answered test item.

Reflection stage. After completing the testing stage, participants filled out a 23 question survey. Several of these questions served as screening checks to exclude participants who reported using external aids or misunderstanding the task (we found none). The remaining questions focused on participants' reflections on the task. For these questions, participants reported on details such as which learning items were more or less challenging to learn, and what strategies they used to respond to test items.

Postscreening. As a final step in our human pipeline, we implemented an 80% accuracy threshold for the final four repetitions of the learning items. Participants who did not meet this threshold were excluded from further analysis. This ensured that our analysis of test stage generalization was based on only participants who successfully learned the learning items, and we report the exclusion rates below.

CRT for LLMs

To assess LLMs on the CRT, we evaluate the performance of finetuning across three open weight LLMs: DeepSeek 67B Chat (Bi et al., 2024), Llama 3.1 70B instruct and Llama 3 70B instruct (Dubey et al., 2024). These models allowed us to balance performance with feasibility, given the compute resources at our disposal.

Finetuning. Finetuning was performed using Low-Rank Adaptation (LoRA; E. J. Hu et al. (2021)), with hyperparameters r=128, dropout=0.1 and α=1 chosen from a light hyperparameter sweep that spanned a 2×4 grid of r and α values.

For each base LLM and each of the twelve possible chains we conducted five uniquely seeded finetuning runs. Each run had an initial learning rate of 1e−3 and lasted for 75 epochs. To remain consistent with the human experiments, each training batch included the two filler items alongside the eight learning items. After each run, finetuned model inferences for the test items were obtained using greedy decoding. All experiments utilized the Huggingface PEFT implementation of LoRA (Mangrulkar et al., 2022) to manage the finetuning and inference processes.

In-context learning. To complement our analysis, we also ran an in-context learning (ICL) experiment with the CRT and the same models. We conducted five uniquely seeded runs for each of the twelve chains, with each run consisting of five repetitions of the eight learning items and two filler items in random order. As the learning stage progressed, previous items and their correct answers were continuously added to the model's context window. During the testing stage, each item was then presented to the model with all 50 ($5\times(8+2)$) learning stage items included within the context window.

We ran two conditions of our ICL experiments. In the base condition, we prompted the model with a simple system prompt that asked it to use the provided context to respond to the current item by immediately producing the answer. In the chain-of-thought condition, we prompted the model with a modified system prompt that asked it to first reason about the current item and the provided context before producing an answer.

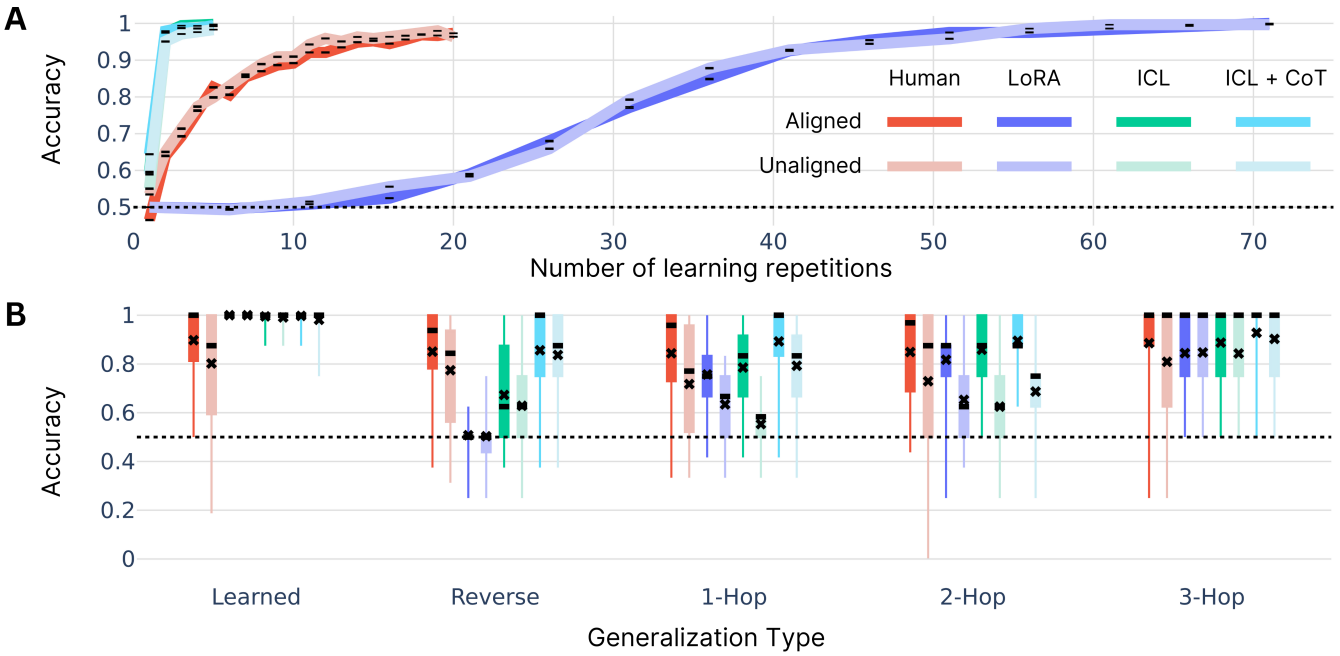

Figure 2: **Human and LLM performance on the CRT.** A) Learning curves. Black bars mark bootstrapped mean accuracy and SDE at each point. While humans learn steadily, in-context learning (ICL) learns nearly instantaneously and finetuning learns more gradually. B) Test stage accuracy. Black bars mark medians, crosses mark means. While all retain ceiling-level accuracy on learned items, only humans generalize robustly across generalization types. Humans and LLMs both demonstrate some degree of sensitivity towards alignment conditions, but this is statistically significant only in LLMs.

Results

Learning stage. During the learning stage, 81% of human participants in the aligned condition and 69% in the unaligned condition met our learning accuracy threshold. Given our sample size, this difference between conditions was not statistically reliable. Of the remaining 54 participants, final learning stage accuracy was high: mean accuracy in the final four learning repetitions was 0.9655 (SD = 0.0497) in the aligned condition and 0.9710 (SD = 0.0493) in the unaligned condition. There was no statistically reliable difference in final accuracy between conditions ($t(52) = -0.4084$, $p = .699$).

Accuracy for LLMs was at ceiling. In the final four learning repetitions, every finetuning run reached 100% accuracy on the learning items, and every ICL run surpassed our 80% accuracy threshold. These ICL runs achieved an overall mean accuracy of 0.9840 (SD = 0.0270) in the base condition and 0.9792 (SD = 0.0363) in the chain-of-thought condition, with negligible differences between alignment conditions. Although there were statistically significant differences in accuracy between the three specific LLM models when performing ICL ($F(2,717) = 53.54$, $p < .001$), these differences were small, as mean accuracy for all models was above 0.96.

Given our hyperparameters, humans learned faster than finetuned LLMs but slower than LLMs using ICL (Figure 2A). The latter were able to surpass 90% accuracy on the learning items within two repetitions, while humans on average took ten repetitions to exceed 90% accuracy and finetuning took 35. Overall differences in learning speed aside, there are two noteworthy aspects of these findings: first, ICL improved with more repetitions, suggesting that performance involved integrating information across multiple instances of the same context items. Second, the finetuning curves exhibited a long warm-up stage, a pattern often observed in the learning dynamics of multi-layer neural networks (Saxe, McClelland, & Ganguli, 2019) but not in humans on our task.

Testing stage. During the testing stage, human accuracy remained high (Figure 2B), with mean accuracy of 0.9039 (SD = 0.1362) in the aligned condition and 0.8580 (SD = 0.1585) in the unaligned condition. Although participants in the unaligned condition tended to have lower accuracy, this difference was not statistically significant ($t(52) = 1.11$, $p = .273$). Participant accuracy was high across all generalization types, and differences between alignment conditions within each type were also not statistically reliable.

By contrast, the overall accuracy of finetuned LLMs was at best moderate, with a mean accuracy of 0.7853 (SD = 0.0866) in the aligned condition and 0.7275 (SD = 0.0798) in the unaligned condition. This difference between alignment conditions was statistically significant ($t(357) = 6.581$, $p < .001$). While accuracy was at ceiling for learned items, it fell to chance levels for reverse items and to intermediate levels for 1-hop and 2-hop items across every LLM that we tested. Accuracy was notably higher for 3-hop items, and we speculate that this is because they involve endpoint entities that are easier to learn. A linear mixed effects model that predicted accuracy from the interaction between condition,

participant type (human or finetuned LLM) and generalization type found statistically significant interactions between participant type and all generalization types, further suggesting that human generalization performance was stronger than that of finetuned LLMs across the board.

Although LLMs performed better using ICL, their performance remained lower than that of humans. In particular, we found that LLMs using ICL with or without chain-of-thought were still very sensitive to alignment conditions for 1-hop and 2-hop transitive generalizations (Figure 2B). While humans may have also had some sensitivity to alignment conditions, this effect was numerically less pronounced and would require a larger sample to be statistically reliable.

How do humans complete the CRT?

Our results demonstrate fundamental differences between humans and LLMs on the CRT. Using finetuning, LLMs fell short of humans during both the learning and testing stages, with slower learning curves and less robust generalization. Even when they were allowed to perform ICL on all past learning items, the LLMs that we tested remained sensitive to alignment conditions in a way that humans were not.

If the mechanisms that underlie human learning on the CRT are different from finetuning or ICL, then what might instead explain human behavior? Our results suggest that humans rely on one of two strategies for representing and reasoning about the learning items. One strategy, which we will refer to as the 'ranker' strategy, involves actively recoding the learning items into a single representation of entity rankings that can be recalled for any test item. The other, which we will refer to as the 'non-ranker' strategy, involves memorizing each learning item independently, and recalling only relevant learning items for each test item.

We refer to two pieces of evidence for this. First, in the self-reflection stage of the human task, we asked participants to write down the underlying entity ranking from the experiment in order of learned chess ability. Then, we asked them if they had this explicit ranking in mind during the learning and testing stages. Participants responded to these questions using a five-point Likert scale that ranged from "Not at all" to "A great deal.", and we used their responses to categorize those who indicated that they had entity rankings in mind to a high degree ("quite a bit" or "a great deal") as rankers and those who indicated that they relied less on explicit entity rankings ("somewhat" or lower) as non-rankers.

Based on these self reports, 29 participants (14 in the aligned condition and 15 in the unaligned condition) could be classified as rankers and 25 (15 aligned, 10 unaligned) as non-rankers during the learning stage, while 37 participants (20 aligned, 17 unaligned) could be classified as rankers and 17 (8 aligned, 9 unaligned) as non-rankers during the testing stage. There was a slightly higher tendency for participants in the unaligned condition to adopt a ranking strategy during the learning stage (χ^2 = 207.36, $p < .001$), and no such association between condition and ranker status during the testing stage (χ^2 = 1.58, p = .209). Fisher's exact tests also suggest that participants who adopted a ranking strategy were more likely to write down the correct entity ranking than those who did not, regardless of whether they reported adopting the strategy during the learning stage (odds ratio = 17.17, $p < .001$) or the testing stage (odds ratio = 7.38, p = .003). These results indicate that participants did use at least two representation strategies for the entities during both learning and testing.

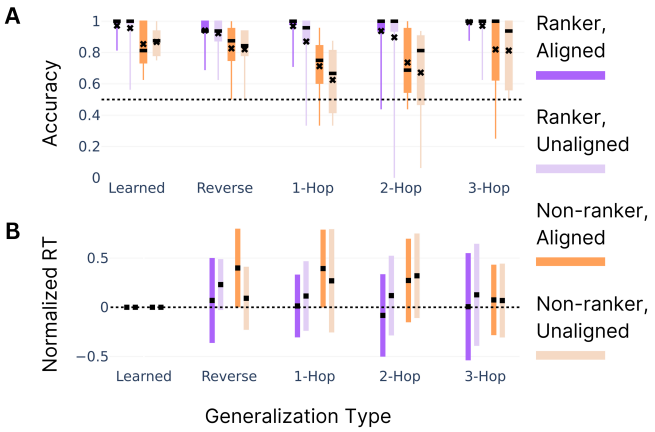

Figure 3: **Human CRT performance, by representation strategy.** A) Rankers perform at near ceiling level across all generalization types, while non-rankers are generally worse and particularly worse at transitive inferences. B) Despite high variance, rankers use roughly constant time to respond to all testing items, while non-rankers spend longer on reverse, 1-hop and 2-hop inferences. RTs are normalized on a participant basis with respect to test stage learned item RTs.

We also found that rankers and non-rankers performed differently on the CRT. During the testing stage, rankers tended not only to achieve higher accuracy (Figure 3A), but also to make faster judgments (Figure 3B). A linear mixed-effects model found a significant main effect of ranker status on accuracy (β = 0.1177, p = .035) as well as significant interaction effects between ranker status and generalization type for 1-hop (β = 0.1381, $p < .001$) and 2-hop (β = 0.0837, p = .004) items. A similar model for reaction times found a significant main effect of generalization type in non-rankers, with higher RTs for 1-hop (β = 1.190, p = .010) and reverse (β = 1.272, p = .006) items, as well as significant interaction effects between ranker status and generalization type, with rankers having lower RTs than non-rankers for 1-hop (β = -1.259, p = .023), 2-hop (β = -1.138, p = .040), and reverse (β = -1.095, p = .048) items.

In summary, rankers performed approximately 10 percentage points better than non-rankers on average, were particularly better than non-rankers at transitive items, and responded faster than non-rankers, especially for transitive items. These findings are consistent with the notion that forming more efficient representations leads to stronger performance – that by proactively thinking about how to best represent the underlying entity chain, rankers could general-

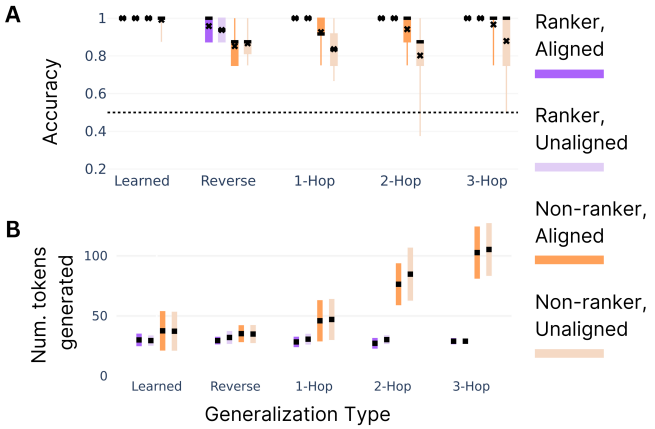

Figure 4: **ICL performance on the CRT given human-like item representations.** A) Given a ranked representation of the learning items, Llama 3.1 70B Instruct achieves perfect accuracy on test items. Given the right individual learning items, accuracy is lower but remains at human-comparable levels. B) Given a ranked representation, Llama 3.1 70B Instruct reasons with roughly constant time, as measured by number of reasoning tokens generated during chain-of-thought. Given the right individual items, reasoning time increases with the number of hops in a transitive inference.

ize more quickly and robustly than their non-ranker counterparts.

Capturing human-like behavior on the CRT

While we have shown that both finetuning and ICL fail to enable our LLMs to match human performance on the CRT, it is of course not true that this very simple task is beyond the capabilities of LLMs in general. Notably, we find that if we borrow the ideas about human representations from the previous section and use them to provide our LLMs with more information-dense item representations in-context, they become capable of performing at human levels.

To demonstrate this, we present a second ICL experiment with two conditions. In the 'ranker' condition, we provided Llama 3.1 70B Instruct with a ranked list of entities encoded as a single string, `Entity rankings: A > B > C > D > E`, rather than the individual learning items. In the 'non-ranker' condition, we provided the same model with only the learning items that were normatively required to confirm or reject a test item. For example, we provided all eight items in-context for 3-hop generalizations, but only the two items containing the same entity pair for reverse generalizations. In both conditions, the context provided to the model was therefore substantially shorter than that of our original experiments, but the prompt was otherwise the same as for the chain-of-thought ICL experiment.

With these representations in-context, Llama 3.1 70B was able to perform at much more human-like levels than in our original experiment (Figure 4A). In the ranker condition, it performed almost perfectly across all generalization types, and in the non-ranker condition it well exceeded its original ICL performance. Strikingly, these models used a roughly constant number of tokens to reason about test items in the ranker condition, but required more tokens for larger transitive generalizations in the non-ranker condition (Figure 4B). While these RT-like patterns did not strictly capture those observed in humans, they abide by the same basic principle: different representations give rise to different reasoning requirements during testing, and therefore different RT profiles.

Discussion

Here we introduced the Chain Rank Task, which we designed to enable fair comparison between human and LLM learning. Compared to the LLMs that we finetuned, humans learned more rapidly and generalized more robustly, and were also less influenced by prior expectations when doing so. Even when our LLMs had full access to all previous learning items and were able to generate chains of thought during in-context learning, their performance still fell short of that of humans.

Our analysis emphasizes the rich and principled approach to learning that is consciously executed by humans but difficult to detect in LLMs. Humans do not succeed at our task because they have better memory than finetuned LLMs – to the contrary, finetuned LLMs generalized poorly compared to humans despite having higher recall of the learning items. Rather, we suggest that human success can be attributed to their abilities to retrieve relevant memories, to use these memories to form new, unprompted representations of the learning items, and to then rely on these representations to efficiently make judgments during learning and testing. Although powerful, updating the autoregressive outputs of LLMs via finetuning does not do justice to the sophistication of this process.

As we pursue future theories of human learning that can capture human behavior on tasks such as the CRT, our in-context learning results highlight one potential direction – modeling human learning as a retrieval-augmented process. The notoriously limited nature of human working memory (Miller, 1956) means that ICL is not itself viable as a complete theory of human learning. However, the fact that our models could use ICL to generalize correctly when provided with the right representations suggests that it may be an important piece of the puzzle. Specifically, and with our analysis of human CRT solutions in mind, it seems natural to think of ICL as one important mechanism in a broader learning system that stores, retrieves, modifies and reasons over past memories in order to respond to present stimuli. A natural step is to think of this system as an analogue of a retrieval-augmented language model that includes a transformer-based 'reasoning module' and a separate memory store (Lewis et al., 2020; Borgeaud et al., 2022; Park et al., 2023). This idea has inherent connections to human brain literature on hippocampal involvement in transitive inference tasks (Shohamy & Wagner, 2008; Zalesak & Heckers, 2009), and presents a new extension of classic dual-systems theories such as Complementary Learning Systems Theory (McClelland, McNaughton, & O'Reilly, 1995) in the age of LLMs and in-context learning.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency* (pp. 610–623).

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., & Evans, O. (2023). The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*.

Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., ... others (2024). Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., ... others (2024). Centaur: a foundation model of human cognition. *arXiv preprint arXiv:2410.20268*.

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., ... others (2022). Improving language models by retrieving from trillions of tokens. In *International conference on machine learning* (pp. 2206–2240).

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... others (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Cohen, R., Biran, E., Yoran, O., Globerson, A., & Geva, M. (2024). Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, *12*, 283–298.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., ... others (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hu, J., Mahowald, K., Lupyan, G., Ivanova, A., & Levy, R. (2024). Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, *121*(36), e2400917121.

Kahana, M. J. (2002). Associative symmetry and memory theory. *Memory & cognition*, *30*(6), 823–840.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... others (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., ... Batson, J. (2025). On the biology of a large language model. *Transformer Circuits Thread*.

Lotfi, S., Finzi, M., Kuang, Y., Rudner, T. G., Goldblum, M., & Wilson, A. G. (2023). Non-vacuous generalization bounds for large language models. *arXiv preprint arXiv:2312.17173*.

Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S., & Bossan, B. (2022). *Peft: State-of-the-art parameter-efficient fine-tuning methods.* `https://github.com/huggingface/peft`.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, *102*(3), 419.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology* (pp. 1–22).

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537–11546.

Shohamy, D., & Wagner, A. D. (2008). Integrating memories in the human brain: hippocampal-midbrain encoding of overlapping events. *Neuron*, *60*(2), 378–389.

St Clair-Thompson, H. L., & Allen, R. J. (2013). Are forward and backward recall the same? a dual-task study of digit recall. *Memory & cognition*, *41*, 519–532.

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526–1541.

Wolford, G. (1971). Function of distinct associations for paired-associate performance. *Psychological Review*, *78*(4), 303.

Zalesak, M., & Heckers, S. (2009). The role of the hippocampus in transitive inference. *Psychiatry Research: Neuroimaging*, *172*(1), 24–30.

Zhong, Z., Wu, Z., Manning, C. D., Potts, C., & Chen, D. (2023). Mquake: Assessing knowledge editing in language models via multi-hop questions. *arXiv preprint arXiv:2305.14795*.