Exploring Causal and Compositional Reasoning in Large Language Models

Magnus Gjerde[1], Vanessa Cheung[1], & David Lagnado[1]
[1] Department of Experimental Psychology, University College London

Abstract

Large Language Models (LLMs) have shown surprising capabilities in reasoning tasks despite lacking direct physical experience with the world. We examine LLMs' ability to reason about object affordances through a tool innovation task where one must select unconventional objects to replace typical tools. In a study comparing GPT-3.5-turbo and GPT-4o with human participants (N=100), we found that while GPT-3.5 performed significantly worse than humans (38.7% vs. 85.8%), GPT-4o with chain-of-thought prompting achieved human-level performance (85.0%). Qualitative analysis revealed that both models could identify causally relevant object properties, but GPT-4o was superior in flexibly applying these properties in novel contexts. We argue that this success relies on compositional reasoning—the ability to decompose objects into abstract properties and recombine them for novel uses. Our findings suggest that LLMs' ability to reason about object affordances has progressed substantially, highlighting the need for further mechanistic research to characterise LLMs' underlying abilities.

Keywords: LLM; affordance; compositionality; causality; reasoning; generalisation

Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in various cognitive tasks, sparking debate about their understanding and reasoning abilities. One intriguing area in which to test LLMs is the domain of tool use and affordances, which concern the ability to reason about objects and their functional properties. This domain is particularly revealing because while LLMs have been trained on text about and images of object manipulations, they do not have the ability to interact with real-world objects or to learn directly from such interactions. This limitation is significant given that scholars like Lakoff and Johnson (1982) and Harnad (1990, 2024) have famously argued that the human conceptual system is built from the ground up, through experience and interactions with the real world. The "unreasonable effectiveness" of LLMs suggest that artificial cognitive systems can get quite far without real-world grounding, but can LLMs reason accurately about real-world tool and object uses?

Recent work has found that LLMs are increasingly sensitive to object affordances (Jones et al., 2022; Yiu et al., 2023). As LLMs' training data contains extensive reference to objects and tools being manipulated, it is plausible that LLMs can solve many affordance tasks through some form of memorisation. This gets to a central question about LLMs: is their often strong performance due to memorisation or some form of reasoning that enables them to deal with novelty and generalise (Krakauer & Mitchell, 2023; McCoy et al., 2023)?

To improve our understanding of how LLMs reason about object affordances and whether they can generalise, we test LLMs through a set of 20 original tool innovation tasks inspired by Yiu et al. (2023). In the task, human participants and LLMs are presented with a goal typically achieved using a specific tool (e.g., carrying water using a glass). They must then select from four options: two objects commonly associated with the typical tool but unsuitable for the task (e.g., a plate, a fork), one unrelated object that could accomplish the task (e.g., a swimming cap), and one irrelevant object. Using human performance as a baseline, we evaluate how well LLMs can judge object affordances in these unconventional scenarios.

While there is some evidence that LLMs can engage in creative problem-solving and use tools in unconventional ways in open responses (Tian et al., 2025), our task tests the models' ability to judge object affordances in a more controlled way. In our task, the model must answer multiple choice questions where the incorrect responses include objects that are closely associated with the typical tool but functionally inappropriate. This makes it challenging for models to give the correct response if they rely on simple associations. We also evaluate their chain-of-thought reasoning about these object properties.

We argue that success in this tool innovation task requires both causal and compositional reasoning abilities (Lake et al., 2017). To identify an unconventional object that can replace a typical tool, one must decompose both the tool and candidate objects into their abstract properties, then reason about which properties are causally necessary for the task. For example, to judge whether a swimming cap could replace a glass for carrying water requires decomposing both objects into properties (e.g., water-tightness, rigidity) and reasoning about which properties are causally essential. This makes the task particularly valuable for examining whether LLMs can go beyond mere association to reason about object properties and their causal roles.

Causal Reasoning and Generalisation

The causal reasoning abilities of LLMs have been extensively studied, with mixed evidence emerging from recent research (Lampinen et al., 2023; Thagard, 2024; Joshi et al., 2024; Bao et al., 2024; Binz & Schulz, 2023). Studies examining multiple varieties of causal reasoning, including causal discovery, effect inference, and counterfactual reasoning, have found that recent LLMs like GPT-4 perform remarkably well compared to earlier models, sometimes approaching human-level performance (Kıcıman et al.,

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

2023). For instance, on tasks spanning Pearl's causal hierarchy of association, intervention, and counterfactuals (Pearl & Mackenzie, 2018), GPT-4 achieved above-chance performance, doing particularly well in interventional reasoning tasks (Jin et al., 2023).

However, a key challenge in interpreting the results of LLM performance in cognitive tasks is determining whether their success stems from genuine reasoning or sophisticated pattern matching based on their training data (Kıcıman et al., 2023; McCoy et al., 2023). To test whether LLMs can generalise and deal with novelty (Chollet, 2019), researchers usually adopt one of three main approaches: creating novel examples of structurally familiar tasks (e.g., Zhang et al., 2024), creating structurally novel tasks (e.g., Wu et al., 2023), and creating tasks where there is a tension between associative retrieval and rule-bound (i.e., general) reasoning (e.g., Yiu et al., 2023). Associative retrieval relies on recognizing and retrieving patterns seen before, while rule-bound reasoning requires appreciating invariant principles that apply broadly.

Wu et al. (2024), for example, presented models with both standard and structurally modified versions of familiar reasoning tasks. One structurally modified task involved adding two numbers in a non-standard base-9 system. While all models showed decreased performance on the modified tasks, GPT-4's performance degraded less than other models, suggesting some capacity for generalization. This pattern has been found in other studies, with newer, larger models demonstrating superior performance on novel task variants compared to their smaller predecessors (for an example in maths, see Zhang et al., 2024). Research on analogical reasoning provides additional support, showing that LLMs can successfully tackle somewhat novel reasoning challenges (Webb et al., 2023, 2024). This suggests that LLMs possess some reasoning abilities that extend beyond mere retrieval of patterns encountered in their training data.

Affordances and Tool Innovation

Sensitivity to affordances—the action-possibilities available to an agent in an environment (Gibson, 2014)—is fundamentally connected to causal knowledge, as understanding what objects afford requires sensitivity to their causal properties. Early research comparing humans and statistical language models found that while humans could readily distinguish between afforded and non-afforded object uses (e.g., using a shirt versus glasses to dry feet), contemporary natural language processing models could not (Glenberg & Robertson, 2000). Two decades later, Jones et al. (2022) replicated this study using modern language models. They found that GPT-3, unlike smaller models like BERT and RoBERTa, showed moderate sensitivity to affordances despite never having physically interacted with objects. The authors suggest that this improvement may primarily be due to the model's larger scale.

A recent investigation of LLMs' understanding of object affordances was conducted by Yiu et al. (2023). They suggest that LLMs are capable of imitation, but not fully capable of innovation. Imitation, they argue, can involve a form of interpolative generalisation and novelty, which LLMs are capable of. For example, one can ask an LLM how heavy a three meter tall person would be, and it will be able to interpolate to give a plausible and novel answer. In contrast, they suggest LLMs struggle with innovation, which requires going beyond familiar patterns in the training data (Yiu et al., 2023, p. 3). They developed a tool task paradigm, which forms the inspiration of our study, where participants had to either select objects associated with typical tools (imitation) or identify unconventional objects that could replace typical tools (innovation). For instance, participants were asked to select an object to draw a circle without a compass, choosing between an associated but inappropriate object (a ruler), an unrelated but causally appropriate object (a round-bottomed teapot), and an irrelevant object.

Their results showed that while LLMs performed on par with humans in the imitation task (GPT-4: 83.3%, humans: 84.9%), they lagged behind in the innovation task (GPT-4: 75.9%, GPT-3.5: 58.9%, vs. human adults: 95.7%). The authors interpreted this as evidence that text-based learning may be insufficient for true tool innovation, which requires going beyond statistical co-occurrence patterns to understand abstract functional analogies and causal relationships between objects. However, given that GPT-4's performance is less than 10% behind that of human children (85.2%), these results might alternatively be viewed as demonstrating significant progress in LLMs' ability to reason about affordances.

The Present Study

In this work, we examine LLMs' ability to reason about object affordances through a modified version of Yiu et al.'s (2023) tool innovation task. Our design increases the task difficulty by including an additional associated option. We test the GPT models at different temperature settings (which determines how variable or deterministic the model's output is) to examine the robustness of their reasoning skills, and with a chain-of-thought (CoT) prompt to evaluate whether it improves their performance (Kojima et al., 2022; Wei et al., 2022).

We tested two preregistered hypotheses: (1) The mean accuracy of GPT-3.5-turbo and humans would differ significantly; (2) the mean accuracy of GPT-4o and humans would not differ significantly. The hypothesis that GPT-4o would perform on par with humans was based on GPT-4o being a later model and an overall improvement over the earlier GPT-4 (Islam & Moushi, 2024). We also explore the problem solving strategies of human participants and qualitatively review the reasoning output of LLMs given chain-of-thought prompts.

Method

The study was preregistered on AsPredicted.org [#184165]. We received ethical approval from the UCL Psychology Ethics Committee (EP/2018/005).

Participants and Materials

We recruited 100 UK participants (50 male, 50 female, age range 19-81, M = 42.3, SD = 12.7) through Prolific. Each participant completed the tool innovation task which included 20 questions and two attention checks. We developed these novel questions inspired by the task structure of Yiu et al. (2023), with the options being designed to be unambiguous and for there to be a clear answer. In each question, participants had to select an object to accomplish a goal typically achieved with a specific tool. The four options consisted of two objects associated with but inappropriate for the task, one unrelated but afforded object, and one irrelevant object. Associated objects were chosen based on contextual co-occurrence with the typical tool (e.g., a plate being associated with a glass). The full materials are available at https://github.com/magnus-gjerde/llm-tool-innovation/tree/main.

Your task is to keep your body warm. Normally, you would use a jacket to accomplish this task. However, a jacket is not available to you. At your disposal, you have the objects listed below. Which of these would you use to accomplish the task?

A bowl (irrelevant)	A boxer shorts (associated)
Curtains (afforded)	An umbrella (associated)

Figure 1: One example question from the test set.

Design and Procedure

The experiment employed a between-subjects design, comparing humans with two LLMs (GPT-3.5-turbo and GPT-4o). The LLMs were tested with both default and chain-of-thought (CoT) prompts, and at three temperature settings (0, 1, and 2). For each temperature setting, 25 responses were collected per model, with an additional 10 responses per model using CoT prompting at temperature 1. Humans saw the tasks in randomised order, while LLMs were presented with each task independently. The order of the object options were randomised for both the LLMs and humans. The human participants completed two attention checks and reported their problem-solving strategy afterward.

LLM responses were collected via OpenAI's API using Python. The default prompt requested that the model "only specify the chosen object", while the CoT prompt asked models to "evaluate the suitability of each option separately" and then specify their choice.

Data Analysis

Following Cheung et al. (in press), LLM scores were averaged by model run to enable statistical comparison with human responses. While this approach has limitations, as LLM responses are stochastic outputs rather than independent samples, it enables meaningful statistical analysis. It also enables us to visualise the variance in the LLMs' responses. For question-specific accuracy and the overall accuracy, all LLM responses were averaged regardless of model run.

For the main analysis, we compared mean accuracy of human and LLMs using a one-way ANOVA with model type as the independent variable. We report follow-up pairwise comparisons with Bonferroni-corrected *p*-values.

Results

Overall Accuracy

Overall, we found a significant effect of model type on mean accuracy, $F(2, 267) = 668.3$, $p < .001$, with significant differences between humans, GPT-3.5-turbo, and GPT-4o ($p < .001$, see Table 1). GPT-3.5-turbo's aggregate mean (38.7%) was substantially below both GPT-4o (74.5%) and humans (85.8%).

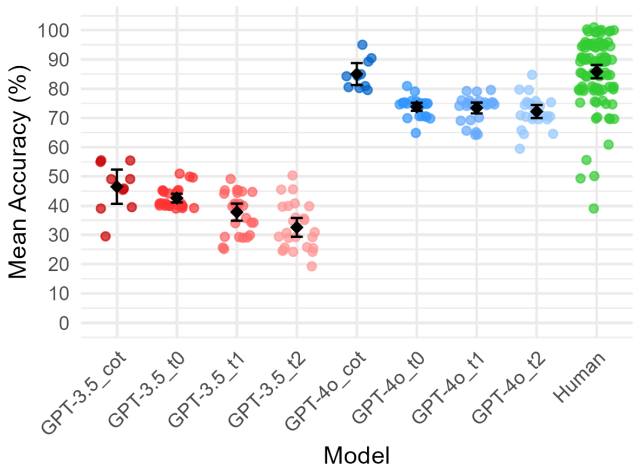

Figure 2: Mean accuracy of humans and the different configurations of the GPT models. On the x-axis, the t0, t1, and t2 labels represent the different temperature settings, while the "cot" label represents performance with chain-of-thought prompting. Mean accuracy and error bars (95% CI) are indicated in black. Jitters show individual data points.

Table 1: Pairwise Comparisons of Aggregated Mean Accuracies Between Humans, GPT-3.5, and GPT-4o

Contrast	Estimate	*SE*	Result
GPT-4o vs. Human	-0.11	0.01	$t(267) = -8.52$, $p < .001$
GPT-3.5 vs. Human	-0.47	0.01	$t(267) = -35.50$, $p < .001$
GPT-3.5 vs. GPT-4o	-0.36	0.01	$t(267) = -25.94$, $p < .001$

Note. Estimate, the difference between the means, is the ratio of correct responses. E.g., -0.11 means GPT-4o scored 11% lower.

Effect of Temperature and CoT Prompting

Comparing humans with different configurations of GPT-4o revealed significant differences between humans and GPT-4o at all temperature settings ($p < .001$, see Table 2). However, with chain-of-thought prompting, GPT-4o achieved comparable performance to humans (85.0% vs 85.8%, $p = 1.00$). CoT prompting substantially improved performance for both models, raising GPT-3.5's overall score by 8.8% and GPT-4o's by 11.9%. As CoT prompts were presented to models at temperature 1, we compared their CoT performance to that with a normal prompt at temperature 1. We found that CoT prompting significantly improved performance for both GPT-3.5 and GPT-4o compared to normal prompting ($t(33) = 3.13$, $p = .04$ and $t(33) = 6.57$, $p < .001$, respectively).

Temperature settings affected the models differently: while GPT-3.5's performance degraded steadily from temperature 0 to 2, GPT-4o maintained relatively stable performance across temperature settings, showing only minor degradation at temperature 2.

Table 2: Pairwise Comparisons of Mean Accuracy Between Humans and Configurations of GPT-4o

Contrast	Estimate	*SE*	*df*	*t*	*p*
GPT-4o_cot vs. Human	-0.01	0.03	180	-0.28	1.00
GPT-4o_t0 vs. Human	-0.12	0.02	180	-5.93	< .001
GPT-4o_t1 vs. Human	-0.12	0.02	180	-6.13	< .001
GPT-4o_t2 vs. Human	-0.14	0.02	180	-6.72	< .001

Note. The estimate is expressed as the ratio of correct responses.

Qualitative Analysis of Strategies

Problem-solving strategies. When asked about their approach to solving the tasks, 65% of human participants reported using visualization, 23% reported reasoning about object properties, and 12% relied on intuition. Two additional options—"I don't know or remember how I solved the tasks" and "Other" (with space for specification)—were available but not selected by any participants, suggesting the three main categories largely captured participants' perceived problem-solving strategies.

Qualitative analysis of GPT responses. While the accuracy scores reported above tell us something about the relative performance of the GPT models, it does not give one a clear impression of the ways in which models succeed or fail at the tasks. We therefore present a qualitative review of some of GPT-3.5 and GPT-4o's responses to the CoT prompts, selecting those responses where the two GPT models differ in illustrative ways.

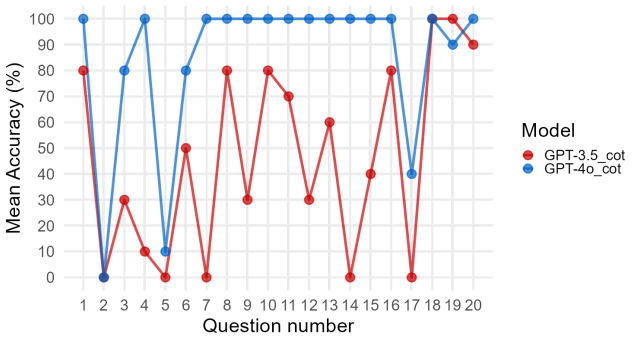

Figure 3: Comparison of mean accuracy per question between GPT-3.5 and GPT-4o models using CoT prompting. Note that questions were presented in randomised order in the experiment.

Question 4: Transporting water. This task involved substituting a glass used for transporting water. The afforded object was a swimming cap, while the two associated objects were a plate and a fork. GPT-3.5 achieved a score of 10% on this task, whereas GPT-4o scored 100%. Below are examples of how they evaluated the swimming cap and the plate:

GPT-3.5

"A **swimming cap** is not suitable for transporting water as it is designed to keep water out rather than hold it in."

"A **swimming cap** is made of waterproof material and can hold water. However, it is not the most practical choice for transporting water as it is typically worn on the head and may not be easy to handle."

Swimming cap: "… it is designed to be worn on the head and does not have the ability to hold water."

Plate: "While it may be able to hold some water, it is not an ideal vessel for transporting water as it is not designed for that purpose and may spill easily."

Plate: "While it may not be as efficient as a glass, a plate can still hold a decent amount of water and can be used to transport it from one place to another."

GPT-4o

"A **swimming cap** is made of a waterproof material … It is flexible and can form a small pouch-like shape when held properly. It can contain and transport water effectively despite not being a rigid container."

Swimming cap: "It can hold a reasonable amount of water if you invert it to form a makeshift container."

"A **plate** … is not very suitable for transporting water. While it can hold some amount of water if it has slightly raised edges, the water would spill easily during transportation as plates are shallow and not designed to contain liquids efficiently."

"If the **plate** has high sides, it might be slightly better, but overall, it is not very practical for transporting water over any distance."

In this instance, GPT-3.5 consistently pointed out that the plate is not an ideal option due to the risk of spilling water. However, it failed to make use of the fact that the property allowing swimming caps to keep water out can also, when held upside-down, enable them to hold water inside. While GPT-3.5 identified the relevant property, it was unable to

flexibly apply it to a novel use case. In contrast, GPT-4o consistently identified and correctly applied the relevant causal properties of the objects.

Question 7: Sweeping the floor. This task involved substituting a broom used for sweeping the floor. The afforded object was a wig, while the two associated objects were laundry detergent and a bucket. GPT-3.5 got a score of 0% on this task, whereas GPT-4o scored 100%. Below are examples of how they evaluated the wig and the bucket:

GPT-3.5	GPT-4o
"A **wig** would not be useful for sweeping the floor as it is a hair accessory and not a cleaning tool." **Wig:** "… it is a hairpiece and has no function in cleaning. It would not be effective in gathering dirt and would likely make a mess if used for this purpose." **Bucket:** "Suitable for sweeping the floor if used creatively. You could turn the bucket upside down and use it as a makeshift tool to push dirt and debris towards a corner or into a dustpan." "A **bucket** could potentially be used to help collect dust and debris, but it would not be very effective in actually sweeping the floor."	"A **wig** is made of hair or synthetic fibers. While unconventional, the strands of a wig could potentially be used to sweep light dust or dirt if there are no other options. It could act somewhat like the bristles of a broom if used carefully." "While not ideal, a **wig** with long hair could be used to gather dust and dirt more effectively than the other items available." "A **bucket** is rigid and designed to hold things, not sweep them. It cannot be used to move debris effectively."

Unlike GPT-4o, GPT-3.5 was unable to appreciate the functional analogy between hair and a broom's bristles. Instead, GPT-3.5 showed a lack of flexibility by emphasising that a wig is not a cleaning tool. GPT-4o solved the task well and successfully identified the causally relevant property of the wig.

Question 14: Fastening a piece of paper to a wall. This task involved substituting Blu Tack used to fasten a piece of paper to the wall. The afforded object was plasters and the two associated objects were rubber bands and a folder. GPT-3.5 scored 0% on this task and GPT-4o scored 100%. Below are examples of how they evaluated plasters and rubber bands:

GPT-3.5	GPT-4o
"**Plasters** could potentially be used similarly to Blu Tack by sticking them to the paper and then attaching it to the wall. However, plasters are not very sticky and may not hold the paper in place for very long." "**Plasters** (band-aids) are unlikely to be effective for fastening a piece of paper to a wall as they are designed for medical use on skin and not for sticking objects to surfaces." "**Rubber bands** could potentially be used to fasten the piece of paper to the wall by looping them around the corners of the paper and then stretching them across the wall to hold the paper in place."	"**Plasters** … have an adhesive strip that could very well stick a piece of paper to a wall. The adhesive on plasters might not be as strong as Blu Tack, but they can perform a similar function in this context." "**Plasters** have an adhesive side which is designed to stick to the skin, and they could serve as a temporary adhesive to affix the paper to the wall." "**Rubber bands** are flexible and can be stretched, but they are not inherently sticky. While they may be used to wrap around objects or bind things together, they cannot adhere paper to a flat wall without additional support or other materials."

As with the other cases discussed, GPT-3.5 demonstrated some ability to identify the causally relevant property—adhesiveness—but was overly focused on the typical medical use of plasters. It failed to recognize that plasters are sufficiently sticky to hold a paper to a wall. Instead, GPT-3.5 often favoured rubber bands, though its evaluations of them either depended on additional objects or were impractical. In contrast, GPT-4o provided accurate and task-appropriate evaluations of the objects.

Discussion

Our findings reveal several key insights about LLMs' ability to reason about object affordances and tool innovation. First, we found that while GPT-4o with standard prompting performed below human level, consistent with Yiu et al.'s (2023) results, it achieved human-comparable performance when we used chain-of-thought prompting. The substantial performance gap between GPT-3.5 and GPT-4o aligns with previous research showing larger models' superior capabilities (Zheng et al., 2023; Chang & Bergen, 2023), while GPT-4o's stability across temperature settings suggests more robust reasoning.

The qualitative analysis of model outputs revealed important differences in how the two LLMs approached the tasks. While both models could identify causally relevant properties, GPT-3.5 struggled to apply these properties in novel contexts, often fixating on objects' conventional uses. In other words, GPT-3.5 showed greater *functional fixedness*, where knowledge of an object's conventional function hinders one's ability to use it in a novel way (German & Defeyter, 2000). In contrast, GPT-4o demonstrated more flexible reasoning, consistently recognizing how object properties could be repurposed for novel applications. This aligns with previous findings that more advanced models show superior generalization capabilities (Zhang et al., 2024; Wu et al., 2023), while providing new qualitative insights into the nature of this improvement.

When asked how they solved the tasks, the majority of human participants reported using mental simulation, with the remainder reporting reasoning about properties or using intuition. This is consistent with research on the use of simulation in human causal reasoning (Johnson-Laird, 2010; Lagnado, 2021; Sloman & Lagnado, 2015). It also raises

interesting questions for future research. If humans report relying mostly on visualisation or simulation, how do LLMs solve the tasks? One benchmark, where solving tasks often require visualising scenes, finds that LLMs still perform well below humans (Philip & Hemang, 2024). Future behavioural research should aim to develop experiments that can help us dissociate different cognitive problem-solving strategies in humans and LLMs.

Causality and Compositionality

While the exact mechanisms by which LLMs encode and process causal relationships remain unclear, our results provide evidence that these models can apply causal knowledge. This aligns with recent findings showing LLMs' increasing performance on causal reasoning tasks (Kiciman et al., 2023), while our study highlights their ability to apply this knowledge flexibly in novel contexts.

We propose that compositional reasoning is key to the LLMs' success in our tool innovation task. The qualitative analysis revealed that both GPTs decomposed objects into abstract properties and assessed how these properties could be used. This ability to break down concepts into constituent parts and recombine them flexibly aligns with some research about LLMs' internal representations. Recent research has identified neural parameters corresponding to abstract concepts and relations within LLMs (Pavlick, 2023; Geva et al., 2021; Todd et al., 2024), and demonstrated that these representations can be systematically manipulated to influence model behavior (Geiger et al., 2021).

The stark difference between GPT-3.5 and GPT-4o's performance, particularly in their ability to recognize and apply object properties in novel contexts, suggests that more advanced LLMs possess enhanced compositional reasoning capabilities. While chain-of-thought prompting improved both models' performance (Kojima et al., 2022), GPT-4o's superior and more stable performance indicates more robust underlying representations. Understanding how these compositional abilities emerge with scale, and how they relate to the models' internal architecture and prompting strategies (Wang et al., 2023), remains important areas for future research.

Limitations and Future Directions

While our study provides insights into LLMs' reasoning capabilities, several limitations should be noted. First, as a behavioural study, it cannot address mechanistic questions about how increased scale leads to improved performance or which internal mechanisms enable successful task completion (Rai et al., 2024). Moreover, given LLMs' vast training data encompassing millions of texts and websites (Liu et al., 2024), it is challenging to ensure that tasks are truly novel rather than variants of patterns in the training data (Xu et al., 2024). Methods to identify task novelty relative to training data will be helpful for future research (Bordt et al., 2024).

A second limitation concerns the text-based nature of our evaluation. As Rutar et al. (2024) argue, true understanding of affordances involves the ability to interact with objects and learn from these interactions in real or simulated environments. While our results show that LLMs can make sophisticated judgments about object affordances, this differs from having an actionable mental model that enables direct object manipulation and learning. Future research might benefit from combining our approach with mechanistic studies examining how LLMs encode abstract properties like 'waterproof' or 'adhesive', including whether these encodings generalize across diverse contexts and enable genuine compositional reasoning with novel property combinations. Another promising path forward would be to give AI models tool affordance tasks that include images or videos, thereby testing their ability to integrate and reason across multiple modalities.

Conclusion

This study has aimed to advance our understanding of LLMs' capabilities through an investigation of their performance on tool innovation tasks. Our findings suggest that LLMs are increasingly able to make accurate judgements about object affordances, and that some of them achieve human-level performance when prompted appropriately. This suggests that LLMs can encode causal knowledge and apply it flexibly, which we argue is made possible through compositional reasoning. Future research combining behavioural and mechanistic approaches will be crucial to understand how LLMs encode and manipulate knowledge about affordances.

References

Bao, G., Zhang, H., Yang, L., Wang, C., & Zhang, Y. (2024). LLMs with chain-of-thought are non-causal reasoners. *arXiv preprint* arXiv:2402.16048.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., ... & Mindermann, S. (2024). Managing extreme AI risks amid rapid progress. *Science*, 384(6698), 842-845.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences, 120*(6), e2218523120.

Bordt, S., Nori, H., Rodrigues, V., Nushi, B., & Caruana, R. (2024). Elephants Never Forget: Memorization and Learning of Tabular Data in Large Language Models. arXiv preprint arXiv:2404.06209.

Chang, T. A., & Bergen, B. K. (2023). Language model behavior: A comprehensive survey. *Computational Linguistics, 50*(1), 293-350.

Cheung, V., Maier, M., & Lieder, F. (In press). Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences.*

Chollet, F. (2019). On the Measure of Intelligence. *arXiv preprint*. arXiv:1911.01547.

Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal abstractions of neural networks. *Advances in Neural Information Processing Systems, 34,* 9574-9586.

German, T. P., & Defeyter, M. A. (2000). Immunity to functional fixedness in young children. *Psychonomic Bulletin & Review*, 7, 707-712.

Geva, M., Schuster, R., Berant, J., & Levy, O. (2021). Transformer Feed-Forward Layers Are Key-Value Memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 5484-5495).

Gibson, J. J. (2014). The Ecological Approach to Visual Perception. Psychology Press.

Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language, 43*(3), 379-401.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena, 42*(1-3), 335-346.

Harnad, S. (2024). Language Writ Large: LLMs, ChatGPT, Grounding, Meaning and Understanding. a*rXiv preprint*. arXiv:2402.02243.

Islam, R., & Moushi, O. M. (2024). GPT-4o: The Cutting-Edge Advancement in Multimodal LLM. *Authorea Preprints.*

Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Lyu, Z., Büth, K., Gonzalez, F., Kleiman-Weiner, M., Suchard, C., & Schölkopf, B. (2023). CLADDER: A Benchmark to Assess Causal Reasoning Capabilities of Language Models. *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*.

Johnson-Laird, P. N. (2010). Mental models and human reasoning. *Proceedings of the National Academy of Sciences, 107*(43), 18243-18250.

Jones, C. R., Chang, T. A., Coulson, S., Michaelov, J. A., Trott, S., & Bergen, B. (2022). Distrubutional Semantics Still Can't Account for Affordances. In *Proceedings of the Annual Meeting of the Cognitive Science Society, 44*(44).

Joshi, N., Saparov, A., Wang, Y., & He, H. (2024). LLMs Are Prone to Fallacies in Causal Inference. *arXiv preprint*. arXiv:2406.12158.

Kıcıman, E., Ness, R., Sharma, A., & Tan, C. (2023). Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint.* arXiv:2305.00050.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems, 35*, 22199-22213.

Lagnado, D. A. (2021). Explaining the Evidence: How the mind investigates the world. Cambridge University Press.

Lakoff, G., & Johnson, M. (1982). Metaphors We Live By. University of Chicago Press.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences, 40,* e253.

Lampinen, A., Chan, S., Dasgupta, I., Nam, A., & Wang, J. (2024). Passive learning of active causal strategies in agents and language models. *Advances in Neural Information Processing Systems, 36.*

Liu, Y., He, H., Han, T., Zhang, X., Liu, M., Tian, J., ... & Ge, B. (2024). Understanding LLMs: A comprehensive overview from training to inference. *arXiv preprint.* arXiv:2401.02038.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023). Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint* arXiv:2309.13638.

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences, 120*(13), e2215907120.

Pavlick, E. (2023). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A, 381*(2251), 20220041.

Pearl, J., & Mackenzie, D. (2018). The Book of Why: The New Science of Cause and Effect. Basic Books.

Philip and Hemang. 2024. SimpleBench: The Text Benchmark in which Unspecialized Human Performance Exceeds that of Current Frontier Models. https://simple-bench.com/

Rai, D., Zhou, Y., Feng, S., Saparov, A., & Yao, Z. (2024). A Practical Review of Mechanistic Interpretability for Transformer-Based Language Models. *arXiv preprint*. arXiv:2407.02646.

Rutar, D., Markelius, A., Schellaert, W., Hernández-Orallo, J., & Cheke, L. (2025). GIBSONA: General Interaction Battery: Simple Object Navigation and Affordances [Preprint]. University of Cambridge.

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology, 66*(1), 223-247.

Thagard, P. (2024). Can ChatGPT Make Explanatory Inferences? Benchmarks for Abductive Reasoning. arXiv preprint arXiv:2404.18982.

Tian, Y., Ravichander, A., Qin, L., Bras, R. L., Marjieh, R., Peng, N., ... & Brahman, F. (2025). MacGyver: Are Large Language Models Creative Problem Solvers?. *arXiv preprint arXiv:2311.09682*.

Todd, E., Li, M., Sharma, A., Mueller, A., Wallace, B. C., & Bau, D. (2024, May). Function Vectors in Large Language Models. *International Conference on Learning Representations (ICLR)*.

Wang, B., Min, S., Deng, X., Shen, J., Wu, Y., Zettlemoyer, L., & Sun, H. (2023, July). Towards Understanding Chain-of-Thought Prompting: An Empirical Study

of What Matters. *The 61st Annual Meeting Of The Association For Computational Linguistics.*

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

Webb, T., Holyoak, K. J., & Lu, H. (2024). Evidence from counterfactual tasks supports emergent analogical reasoning in large language models. *arXiv preprint* arXiv:2404.13070.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35,* 24824-24837.

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., ... & Kim, Y. (2023). Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint* arXiv:2307.02477.

Xu, C., Guan, S., Greene, D., & Kechadi, M. (2024). Benchmark Data Contamination of Large Language Models: A Survey. *arXiv preprint* arXiv:2406.04244.

Yiu, E., Kosoy, E., & Gopnik, A. (2023). Transmission versus truth, imitation versus innovation: What children can do that large language and language-and-vision models cannot (yet). *Perspectives on Psychological Science,* 17456916231201401.

Zhang, H., Da, J., Lee, D., Robinson, V., Wu, C., Song, W., ... & Yue, S. (2024). A careful examination of large language model performance on grade school arithmetic. *arXiv preprint* arXiv:2405.00332.

Zheng, S., Zhang, Y., Zhu, Y., Xi, C., Gao, P., Zhou, X., & Chang, K. C. C. (2023). GPT-Fathom: Benchmarking Large Language Models to Decipher the Evolutionary Path towards GPT-4 and Beyond. *arXiv preprint* arXiv:2309.16583.