Position: Principles of Animal Cognition to Improve LLM Evaluations

Sunayana Rane [*1] **Cyrus F. Kirkman** [*2] **Graham Todd** [*3] **Amanda Royka** [*4] **Ryan M. C. Law** [*5]
Erica A. Cartmill [+6] **Jacob G. Foster** [+7 8]

Abstract

It has become increasingly challenging to understand and evaluate LLM capabilities as these models exhibit a broader range of behaviors. In this position paper, we argue that LLM researchers should draw on the lessons from another field which has developed a rich set of experimental paradigms and design practices for probing the behavior of complex intelligent systems: animal cognition. We present five core principles of evaluation drawn from animal cognition research, and explain how they provide invaluable guidance for understanding LLM capabilities and behavior. We ground these principles in an empirical case study, and show how they can already provide a richer picture of one particular reasoning capability: transitive inference.

1. Introduction

In the early 20th century, a horse named Clever Hans gained international fame for his ability to solve arithmetic calculations—including addition, division, fractions, and telling time—and even "talk" by tapping his hoof on a grid of numbers and letters. Hans could complete a wide variety of tasks with a high degree of accuracy, and toured throughout Germany performing as "the first talking animal." However, an empirical investigation by comparative psychologists revealed that Hans was not performing calculations at all. Instead, he was unconsciously responding to subtle, involuntary cues from his owner—such as micro-movements and facial expressions—that guided him to the correct answer. Interestingly, Hans' owner was unaware of these unintentional signals, genuinely believing that the horse was acting independently (for a review, see Samhita & Gross, 2013). This phenomenon, now known as the "Clever Hans Effect," remains a cautionary tale in the study of intelligence and communication, and emphasizes the importance of careful and rigorous experimental design when investigating behavior for hallmarks of intelligence.

This historic case study raises critical questions in the modern age of artificial intelligence: How and when can we be sure that large language models (LLMs) exhibit the cognitive capacities of humans and other evolved organisms? As LLMs have become larger and more sophisticated, researchers have largely focused on the creation of novel tasks designed such that high performance can be taken as evidence of an underlying cognitive capacity. This approach has been used to probe language models for theory of mind (Kosinski, 2023; Strachan et al., 2024), abstract and analogical reasoning (Chollet, 2019; Webb et al., 2023), planning (Momennejad et al., 2024), and even "general intelligence" (Bubeck et al., 2023). While undeniably useful for benchmarking model performance and improvements over time, these tasks also raise the possibility that LLMs might "cheat" and achieve high performance merely through the sheer scale of their parameter space and training data. Indeed, there is some indication that LLM performance on many such tasks is "brittle" and vulnerable to small changes in problem formulation (McCoy et al., 2023; Lewis & Mitchell, 2024), which can lead to unexpected, unreasonable downstream behavior (Rane, 2024). Further, many existing evaluation tasks produce only a single numerical performance metric, limiting the inferences that can be drawn about the full extent and limits of a model's abilities. In this paper we argue not for the use of a new *task* but rather for the adoption of a set of *principles* that can guide the creation of new evaluation methods. **We argue that the core principles introduced in this paper, drawn from methods in animal cognition research, can help us develop more robust evaluations for LLMs.**

[*]Equal contribution,[+]Equal advising/senior authors listed alphabetically. [1]Department of Computer Science, Princeton University [2]Department of Psychology, University of California Los Angeles [3]Department of Computer Science and Engineering, New York University Tandon [4]Department of Psychology, Yale University [5]MRC Cognition and Brain Sciences Unit, University of Cambridge [6]Department of Anthropology, Cognitive Science Program, and Program in Animal Behavior, Indiana University Bloomington [7]Department of Informatics and Cognitive Science Program, Indiana University Bloomington [8]Santa Fe Institute. Correspondence to: Sunayana Rane <srane@princeton.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

2. Principles

How, then, can we more robustly and thoroughly evaluate the strengths and weaknesses of LLMs to better inform science and policy about model safety and utility? Here, we turn to another field which has substantial experience probing other "black box" intelligences – animal cognition. Animal cognition researchers do not have the luxury of simply asking their subjects whether they understand a particular concept. Instead, they develop rigorous experimental paradigms that tease out cognitive capacities while eliminating as many alternative explanations as possible. This allows for a rich, mechanistic, and functional understanding of cognitive capacities. Animal cognition researchers also have to adapt experimental paradigms to test understanding of the same abstract concept in various different animals. Over decades of careful study, they have developed adaptable and robust experimental paradigms which can be adapted to probe LLMs' abilities from various angles, characterizing where they succeed and precisely how they fail. Taking inspiration from a wide range of such studies and model organisms, we present five "core principles" that offer useful lessons to improve LLM evaluation. We first list the principles, along with several representative papers exemplifying each one. Then, we provide an in-depth exposition of the principles and their application to LLMs.

1. Design control conditions with an adversarial attitude (**P1**) (Howard & Barron, 2024; Boesch, 2021; Halina, 2023; Samhita & Gross, 2013)
2. Establish robustness to variations in stimuli (**P2**) (Brannon & Terrace, 2000; Farrer, 1967; Greene, 1983)
3. Analyze failure types, moving beyond a success and failure dichotomy (**P3**) (Martin & Santos, 2016; Janmaat, 2019; Shettleworth, 2012; Washburn et al., 1997)
4. Clarify differences between mechanism and behavior (**P4**) (Shettleworth, 2001; Nematipour et al., 2022; Alem et al., 2016)
5. Meet the organism (or, more broadly, intelligent system) where it is, while noting systemic limitations (**P5**) (Howard & Barron, 2024; Zhang et al., 2005; Budaev et al., 2019)

2.1. Design control conditions with an adversarial attitude (P1)

When attempting to determine whether a system has a certain capacity, it is essential to design tasks that enable reasonable inferences. However, more often than not, successful task performance has multiple potential explanations, requiring careful consideration of alternative processes. Researchers in animal cognition must therefore adopt an adversarial attitude to account for these possibilities and mitigate biases in design or interpretation, such as anthropomorphism and anthropocentrism—that is, the tendency to project human traits onto animals, and the assumption that human ways of thinking or behaving are the default or most important standard.

Take, for example, an early study of nonhuman primate memory in which experimenters aimed to test chimpanzees' ability to select a stimulus that matched a previously-shown sample (Farrer, 1967). Chimps were provided with a "match-to-sample" task consisting of repeated trials in which they were first shown a sample stimulus, then a brief delay, before being presented with several choices (one of which was correct). Chimps were reinforced with food for selecting the correct match. While they quickly mastered the task with high proficiency ($> 90\%$), additional control conditions revealed they didn't rely on the same simple matching strategy as humans would. When the sample was removed in a follow-up control condition, chimps somehow maintained near-perfect performance. They had learned a chain of 17 unique "if, then" conditional discriminations to select the correct choice and solve the task! Despite the many similarities between humans and nonhuman animals, problem-solving can occur through means that are functionally different, though often equally valid.

Within animal cognition research, it is common for a single study to be accompanied by multiple follow-ups, each controlling for an alternative explanation; often, experiments are directly or conceptually replicated across many different labs after publication to test robustness and probe plausible alternative explanations (Boesch, 2021; Halina, 2023). This adversarial approach—in which authors take it upon themselves to propose alternative explanations to their theories and then design and run studies to specifically test them—enables stronger inferences about the mechanisms underpinning a given behavior. By adopting a similar adversarial attitude in their research, LLM researchers can produce rigorous experiments and objective interpretations of results that are less susceptible to alternative hypotheses.

2.2. Establish robustness to variations in stimuli (P2)

One of the primary ways to design rigorous control conditions is through careful selection of experimental stimuli. Most cognitive capacities of interest (e.g., analogical reasoning, planning, etc.) are domain-general: they are useful precisely because they are applicable in a wide range of situations. It is essential, then, to carefully design experiments such that performance hinges on the domain-general cognitive process and not a contingent or local property of the particular experimental stimuli. For instance, if a system is capable of analogical reasoning, then it should be able to apply such reasoning over a variety of colors, shapes, objects, etc. To claim that a system exhibits analogical

reasoning (in lieu of simpler explanations like associative learning), researchers must show that a system's successes are not restricted to a single stimulus or class of stimuli.

Imagine you are tasked with differentiating each face of a six-sided die. An obvious option to humans might be to simply count the number of spot elements on each face, but, adopting an adversarial attitude, are there other ways the problem can be solved? For example, could you differentiate faces by simply comparing the total amount of black vs. white pixels present? Brannon & Terrace, 2000 faced this problem while investigating how macaque monkeys represent numerosity, or the ability to perceive and estimate the quantity of items in a set. They carefully controlled for this pixel differentiation strategy by altering the size of numerical elements but fixing relative pixel area. However, this change opened up another way to differentiate stimuli without counting: faces of the die could now be differentiated by the relative size of each element. This required yet another stimulus condition in which element size was quasi-randomly varied, such that the previous contingency between size and numerosity was broken. Brannon & Terrace, 2000 used seven different stimulus sets in total, each addressing a possible non-numeric approach to solving the problem and, thus, ensuring that the domain-general capacity of numerosity was being measured.

In addition to variations in the features of experimental stimuli, other factors, like the size of stimulus banks, have also been shown to yield substantial and meaningful changes in behavior. Bodily et al., 2008 utilized the aforementioned match-to-sample task to investigate identity-matching, but, as their primary experimental manipulation, systematically varied the number of utilized stimuli across conditions. Here, the experimental protocol and types of visual stimuli were exactly the same, but the number of stimulus options differed. Interestingly, they found that pigeons will adopt an bottom-up strategy of sole associative memorization when faced with a smaller stimulus bank (e.g., 24 stimuli), but will systemically switch to a top-down rule-based strategy when faced with larger stimulus banks (up to 768 stimuli). In other words, subjects adopted two very different problem-solving strategies when faced with different stimulus bank sizes, even though the stimuli and experiment were functionally identical across conditions.

As we attempt to characterize the full range of LLM behaviors (both beneficial and deleterious), it is vital to probe thoroughly for robustness to variations in stimuli. If an LLM has drastically different behavioral responses to variations in stimuli, it could be an indication of over-reliance on priors (e.g. only performing well on stimuli similar to those it has seen before). Conversely, if an LLM demonstrates true domain-generality across many stimulus types, it becomes more likely that the LLM is not solely relying on correlations and instead has an effective world model for the task.

2.3. Analyze failure types, moving beyond a success and failure dichotomy (P3)

Oftentimes, a surprising failure can be more informative than an expected success. Because the financial and moral cost of running experiments on live animals is high, animal cognition researchers must design efficient experiments that maximize the information that can be gained, even in the event that subjects do not behave as expected. For an empirical example of how failure types can be informative in better understanding LLM behavior, see subsection 7.2. Furthermore, understanding the ways in which a subject fails can allow you to predict future failures and gain deeper understanding of its behavior. Analogously, we argue that LLM evaluations can benefit from a richer account of the particular kinds of situations in which models succeed and fail. For instance, the question *"Can LLM model X perform logical reasoning?"* might be replaced with *"In which scenarios does LLM model X exhibit logical reasoning, and what are the features of the scenarios which give rise to that behavior? How might these features be functionally extrapolated to other scenarios?"* We note, however, that researchers must be cautious when generating explanations of differences in performance – it is easy to anthropomorphize or assume that an animal or model is failing due to a cognitive bias found in humans. The field of animal cognition also has several principles that can aid in parsimonious interpretation and modeling of behavioral processes while simultaneously combating anthropomorphism. One of these is Morgan's Canon (Waters, 1939; Epstein, 1984; Zentall, 2018), which states:

> "In no case may we interpret an action as the outcome of the exercise of a higher psychical faculty, if it can be interpreted as the outcome of one which stands lower in the psychological scale."

In other words, interpretations of behavior should rely upon the simplest possible explanation—such as instinct or learned associations—*before* attributing more complex cognitive abilities like reasoning or insight. Adoption of similar perspectives has the potential to increase validity and boost theoretical alignment within LLM research.

2.4. Clarify differences between mechanism and behavior (P4)

Both AI researchers and comparative psychologists aim to make claims about the unobservable capacities of their respective subjects of study. Moreover, both groups examine similar capacities such as inference, theory of mind, and analogical reasoning, among others. However, these capacities

cannot be directly seen. Instead, researchers must rely upon observable behaviors (e.g., actions recorded in the wild or in an experiment or output generated in response to a given prompt) in order to make inferences about whether their subjects possess a certain capacity. Due to the close relationship between the behaviors we observe and the inferences they enable us to make, it can be easy to conflate the two. In order to be clear about one's claims and facilitate discussion about the soundness of behavior-mechanism inferences, it is of paramount importance to clarify when claims are about observable behaviors as opposed to referencing the mechanisms producing those behaviors.

Animal cognition researchers and theoreticians have developed a useful shorthand for maintaining this distinction. When discussing a behavior that seems to align with all of the non-cognitive, non-mechanistic criteria for a capacity, then they attach the label "functional" (Macedonia & Evans, 1993; Hall & Brosnan, 2017; Byrne & Whiten, 1991; Kline, 2015). Consider this example of functional teaching: adult meerkats will provision pups with pre-injured prey when they are too young to hunt on their own, a behavior that decreases in frequency as the pup ages. While this behavior meets important criteria for teaching, most definitions of teaching require that teachers represent something about their pupil's knowledge (or lack thereof) (Thornton & McAuliffe, 2006). Even in the absence of positive or negative evidence that meerkats satisfy this criteria, researchers can still productively discuss this as an instance of meerkats functionally teaching pups how to handle prey without implicitly making claims about the mechanisms driving this behavior.

LLM researchers could embrace a similar distinction. For example, when LLMs pass developmentally-inspired mental state reasoning tasks, this might seem consistent with possessing a Theory of Mind (Kosinski, 2024). However, without evidence of an emergent causal model underpinning their responses (Gopnik & Wellman, 1992) and in the absence of sufficient adversarial controls, a term such as "functional Theory of Mind" could serve as a useful marker that a behavioral benchmark has been met, but that we still lack sufficient evidence for understanding mechanism.

2.5. Meet the organism (*or, more broadly, intelligent system*) where it is, while noting systemic limitations (P5)

When testing the capacities of any intelligent system, there is always a risk that their performance on a given task may not reflect their underlying competence. Nonhuman animals, for example, could fail even the most elegant experiment for a variety of reasons such as a lack of motivation, neophobia, perceptual limitations, etc. As such, researchers studying animal cognition try to account for these considerations when designing tasks. In this way, researchers try to "meet the organism where it is." Studies should be adversarial from a theoretical perspective via engagement of alternative explanations, but they should never adversarially put the target species at a disadvantage (Howard & Barron, 2024). For example, fish should not be judged on their ability to ride a bike any more than humans should not be judged for their ability to breathe underwater. If subjects cannot fully engage with the task at hand, then the results are of little use for determining the subjects' true capabilities.

Much like nonhuman animals, it is reasonable to expect that LLMs might require some accommodations when designing studies in order to give them the best chance at succeeding. Some of this might take the form of careful prompt design. In other cases, this could take the form of awareness of innate shortcomings in certain domains, such as physical reasoning. Given that LLMs are un-embodied systems without the benefit of direct experience with the physical world, testing them on problems related to physical reasoning obscures capacities that LLMs could apply without issue in other domains (Webb et al., 2023).

Another way that animal researchers meet animal models "where they are" is to carefully account for subject-specific differences when constructing experiments. For instance, take the famous example of Pavlov's dogs: after repeated training with tone + food pairings, dogs were shown to salivate to the tone alone. This salivation *could* be evidence of associative learning, but there are several alternative explanations that need to be ruled out: perhaps, for example, dogs are simply more likely to salivate after being fed repeated times (this is called "sensitization"). A control condition for sensitization is to simply present a different dog with food only, then measure differences in salivation. But what if the two dogs innately have different salivation rates? Or what if one learns a bit slower than another? A better option, then, is a within-subject experiment in which each dog is presented with two tones: one paired with food and one not. In animal behavior research, within-subject experiments enable strong experimental effects with efficient data collection–sometimes with as few as two subjects (Kirkman et al., 2022).

3. Alternative Views

Contrasting with our view that principles of animal cognition research can help us better understand and improve LLM abilities, there is the perspective that probing LLMs using techniques developed for non-linguistic subjects is not worthwhile. Given how much data and compute LLMs consume, and how dependent they are on language, some take the view that they are different enough from all forms of biological cognition to not warrant further comparison with biological intelligent systems. Animal cognition is,

however, inherently a study of a variety of extremely diverse forms of intelligent behavior. Considering that animal cognition researchers have nevertheless been able to identify design practices and experimental paradigms that are useful in studying the staggering range of these intelligent behaviors, we believe that some of these principles and practices will similarly prove invaluable for understanding and improving LLM behavior. However, while we take the view that probing LLM behavior for underlying understanding is important, and that studies of animal behavior have developed transferrable methods for how to do this in a variety of intelligent systems, it is also possible that these methods will not yield results as interesting as we would anticipate.

4. Case Study: Transitive Inference

These five principles provide broad guidance that can help our machine learning research community effectively study and holistically evaluate the behavioral properties of LLMs and other foundation models. Here, we demonstrate their value in an empirical case study.

Much of animal cognition research prioritizes the study of domain-general reasoning capacities to reveal the fundamental mechanisms that drive behavior, rather than focusing solely on specific, task-dependent abilities. These capacities, which enable animals to solve problems and infer relationships across a variety of contexts, are considered the foundational building blocks upon which more specialized cognitive skills can be built. By identifying these underlying principles and comparing them across species, comparative cognition researchers gain insight into how animals confront novel challenges, adjust to changing environments, and display adaptive flexibility in their behavior. One domain-specific cognitive capability that has been studied widely and thoroughly across the animal kingdom is Transitive Inference (TI), or the ability to extrapolate ordinal relationships between items that have not been directly compared.

TI is a fundamental form of reasoning in which, given some premises, a relational conclusion between stimuli can be inferred. For example, if A is greater than B and B is greater than C, then A must, transitively, be greater than C. TI has historically been used to evaluate the reasoning ability of young pre-verbal children (Burt et al., 1911), but has also been well studied throughout the animal kingdom, from rats (Roberts & Phelps, 1994) to wasps (Tibbetts et al., 2019). There are several experimental approaches to studying TI in animals, the most common of which is the n-term series task (see subsection 6.2).

TI has many specific adaptive functions, such as rank estimation in social animals (Cheney & Seyfarth, 1986; Shettleworth, 2004). However, given its deep functional homology across diverse taxa, TI appears to also serve as a foundational building block for a wide range of logical reasoning phenomena. Both causality—the capacity to infer when one event (the cause) influences or directly results in another event (the effect)—and numerosity—the ability to perceive and estimate the quantity of items in a set—rely on TI as a foundational cognitive process (Halpern, 2016; Zalesak & Heckers, 2009; McGonigle & Chalmers, 1986). TI is the kind of domain-general reasoning fundamental to logical capacities that are desirable for LLMs to exhibit.

Another advantage of evaluating the TI ability of LLMs lies in the wealth of behavioral models that have been developed to illustrate and emulate the phenomenon (see Vasconcelos, 2008 for an in-depth review). Models span both bottom-up approaches, based upon associative and reinforcement principles (Wagner & Rescorla, 1972; Wynne, 1998; Siemann & Delius, 1998), and top-down approaches, relying upon cognitive rule-based principles (Harris & McGonigle, 1994; Bryson & Leong, 2006). Some models of TI have even been directly applied to basic neural networks (Frank et al., 2003; De Lillo et al., 2001; Wu & Levy, 2001). This diverse set of models provides valuable functional frameworks from which TI can be compared and evaluated in LLMs. Furthermore, the neural architecture facilitating TI has been the subject of substantial investigation and is well understood across multiple animal species (Dusek & Eichenbaum, 1997; Zalesak & Heckers, 2009; Ramawat et al., 2023).

This rich foundation for evaluating, quantifying, and modeling TI allows for parallels to be drawn between established animal cognition paradigms and exploratory machine learning systems (see subsection 7.2). By leveraging these insights, domain-general reasoning capacities such as TI (which are fundamental to logical reasoning and adaptive problem-solving) can be more holistically investigated and understood in LLMs.

5. Why probe non-linguistic task behavior in a *language* model?

Language is an essential part of human intelligence (Rumelhart, 1993; McClelland et al., 2020; Rabovsky et al., 2018). However, disentangling linguistic acuity (*sounding* intelligent) from underlying world understanding (*actually being* intelligent) can help us develop more nuanced, useful evaluations for LLM behavior. Evaluations that are agnostic to language use, but are dependent on an underlying understanding of a fundamental principle or concept such as TI, can help us tease apart when the performance of an LLM is driven merely by patterns of word co-occurrences, and when it is instead the result of a deeper, more robust and human-compatible understanding of the world. Past work has shown that foundation models that achieve superhuman performance in some tasks lag far behind humans and many animals in other, seemingly simpler tasks (such as numeros-

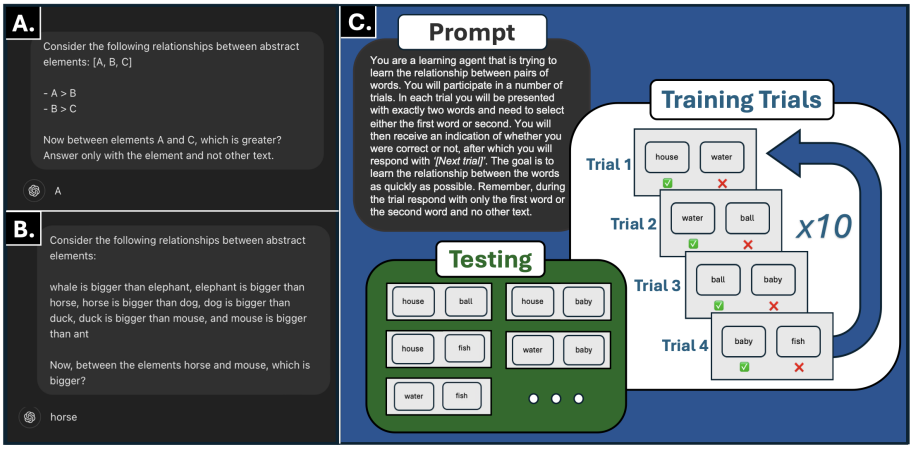

Figure 1. Three task structures, adapted to align with LLM's modalities, used to probe transitive inference (TI). ***Panel A***: A simple 3-term query for TI between elements [A, B, C]. ***Panel B***: An example of the size-ranked animal element list used in Experiment 1. Element list, ordinal operators, and queried pairs were varied across iterations. ***Panel C***: The trial-based *n-term* task used in Experiment 2. The model was first presented with an introductory prompt outlining contingencies and ideal behavior. Following the prompt were training trials in which two adjacent pairs of neutral stimuli were presented and the model was prompted to choose one option; choices received differential "in/correct" feedback in the following message (visualized with green check mark and red 'x', respectively). The message following feedback was another trial with a new adjacent pair. Seven total stimuli (only four illustrated in Panel C) meant six adjacent pair trials, which were repeated 10x each in randomized order until they were adequately learned. Last, the model was tested with trials consisting of two familiar elements in a novel pairing.

ity, see Rane et al. 2024). Therefore, LLM evaluations must account for these large variations in performance across simple and complex tasks, and must move away from simplistic statements towards a richer, more nuanced understanding of the strengths and weaknesses of LLMs. Research on animal (and indeed human; e.g., Duncan et al. 2000) cognition has been probing nonverbal intelligence for decades, and has formalized experimental principles that allow for evaluation and quantification of fundamental traits we identify with "intelligence," probing for a deeper, robust understanding of the world.

6. Methods

What does probing Transitive Inference in an LLM look like? It is possible to simply query a model directly to define or demonstrate TI (and they will generally do so quite persuasively; see Figure 1). However, such a glib explanation may mask a deeper misunderstanding of the concept. To investigate this question, we adopt a series of classic TI paradigms from the animal cognition literature that were tailored specifically for GPT-4o (Figure 1). We note which of our core principles (**P1**-**P5**) are used in each set of experiments.

6.1. Experiment 1: Transitive Operator & Element Manipulation

One way to probe TI ability in an intelligent system that converses in natural language is to simply use language to ask the LLM which element is "bigger." Figure 1 shows this type of task structure. Varying the ">" and "bigger than" operators serves as a simple adversarial control (**P1**, **P2**); if general transitive inference were being used, performance should be insensitive to this variation.

We then analyze the specific pattern of failures (**P3**) as a function of variation in stimulus (**P2**). Three stimuli sets were used: ranked words (transitively-linked animal names ranked from biggest to smallest size), reverse rank (incorrectly ranked animal names in reverse order of biggest to smallest), and random strings (no transitive link between words).

6.2. Experiment 2: Trial Structure (*n-term* task)

Experiments manipulating operator type demonstrate a characteristic brittleness of LLMs caused by minor prompt variations; this was revealed through analyzing the pattern of failures (**P3**). While these findings indicate a lack of robustness in TI generalizability, we turn to our fifth principle (**P5**) and go beyond minor prompt variation, instead developing an experiment with minimal noise from complex prompts and without an explicit operator type or transitively-linked elements. We turn to a robust trial-structured task frequently

used in animal cognition studies of TI called the "*n-term* task." This trial-based structure is inherently less linguistic as it is operator-agnostic. Our *n-term* task is designed to "note systemic limitations" **(P5)** that may arise from abstracting the task away from the linguistic domain. That being said, we began by giving the LLM *some* useful linguistic information (more than an animal might receive) by first prompting the model with information regarding ideal performance parameters (see Figure 1 for full prompt).

The language model was then presented with a series of consecutive choice trials, each consisting of two words systematically chosen for transitive neutrality. Seven word stimuli were chosen, and were randomly paired across 10 iterations of this task. Within one iteration, pairs remained consistent and were bound in an ascending order (*AB, BC...FG*, such that A was always correct and B was always incorrect). After the language model guessed one of two options, it was differentially "reinforced" with a response of "in/correct." Sequential trials were presented in a quasirandom order, in which there could be no more than three consecutive repeats of one trial type. Correct word order within trials was alternated randomly. Piloting showed that the model was able to learn these pairwise discriminations within 3-5 trials, so we presented 10 of each pair for a total of 60 training trials per iteration.

After training was complete, we tested for TI by presenting novel non-adjacent pairs. The pairs consisted of two words that were used in training, but that had never been presented together before. Tested words were not near terminal ends of the chain (e.g., *A* or *G*), and, therefore, had previously been both a correct and incorrect choice on various trials (e.g., *BF*). This test condition required the language model to integrate ordinal information over a sequence of abstract interactions with no explicit ordinal operator. With this condition, we can start to separate mechanism from behavior (**P4**); if the same mechanism (transitive inference) were used across all possible contexts, we would expect similar performance on this task variant. Instead, performance dropped to chance levels (see Figure 2). This experimental structure minimizes word use and instead focuses on a deeper look at the underlying behavior, and this seems to show TI behavior to be brittle in application.

7. Empirical Results

While the model showed excellent performance on initial versions of this task, performance dropped significantly after the introduction of control conditions that abstracted the task (Table 1). Careful probing revealed that the language model's performance was not robust, but instead highly contingent on the specific words used in the prompt and test set (see Figure 2). Taken together, this probing demonstrates the lack of a generalizable TI ability in GPT-4o.

7.1. Experiment 1 Results

As to be expected, performance in all conditions was highest for element pairs that were explicitly provided in the prompt. Although errors here were few, further investigation into error types (**P3**) of adjacent pairs revealed that the model was more likely to answer correctly for element pairs near the beginning and at the very end of the list compared to those near the middle. This behavior is analogous to the serial position effect, which has been well-documented in humans and animals (Bryant & Trabasso, 1971; Woocher et al., 1978; DiMattia & Kesner, 1984). Results across operator and element types are reported in Table 1 and Figure 2.

7.2. Symbolic Distance Effect

When presented with TI tasks in long chains of five or more paired elements, humans, pigeons, and macaque monkeys are more accurate when making comparisons between novel pairs of abstract elements at a greater distance (i.e., those having more intervening terms between them). In other words, $B > F$ is easier than $B > D$. This phenomenon is known as the Symbolic Distance Effect (SDE; D'amato & Colombo, 1990). When investigating error types (**P3**) across symbolic distances in Experiment 1, we found evidence of behavior analogous to the SDE (see Fig. 2). Normalized accuracy scores were found to be positively correlated with symbolic distance in all six conditions.

7.3. Experiment 2 Results

Experiment 2 removed the ordinal operator from the task and instead depended upon an *n-term* trial-based task in which correct choices were differentially reinforced by the experimenter. Pairwise element distinctions were learned relatively quickly, with the model learning the correct binary choice within 3-5 trials.

Despite fast learning and maintained performance across training trials, the model performed at chance when tested with novel non-adjacent pairs of elements with symbolic distances between 2-6. No differences between symbolic distances were observed (likely due to floor effects). In other words, the model did not show evidence of TI ability when the ordinal operator was removed in the *n-term* task.

8. Discussion

These results demonstrate how careful study design inspired by animal cognition can provide a deeper insight into whether and where important behavioral capacities are present in LLMs. For example, when designing an experiment to answer a simple question like "Do LLMs understand the concept of transitivity?" minor prompt variations can

Table 1. Accuracies across TI tasks. As transitive elements and ordinal operators were systematically removed across experiments, evaluation performance dropped substantially from 100% to 40% (roughly chance in the trial-based *n-term* task).

TASK	TRANSITIVE ELEMENTS	ORDINAL OPERATOR	ACCURACY	S.D.
A VS. C	X	X	1.00	0.00
B VS. D	X	X	1.00	0.00
RANKED ANIMAL	X	X	0.97	0.04
RANDOM STRING		X	0.92	0.09
REVERSE RANKED ANIMAL		X	0.78	0.20
N-TERM TRIAL-BASED			0.40	0.50

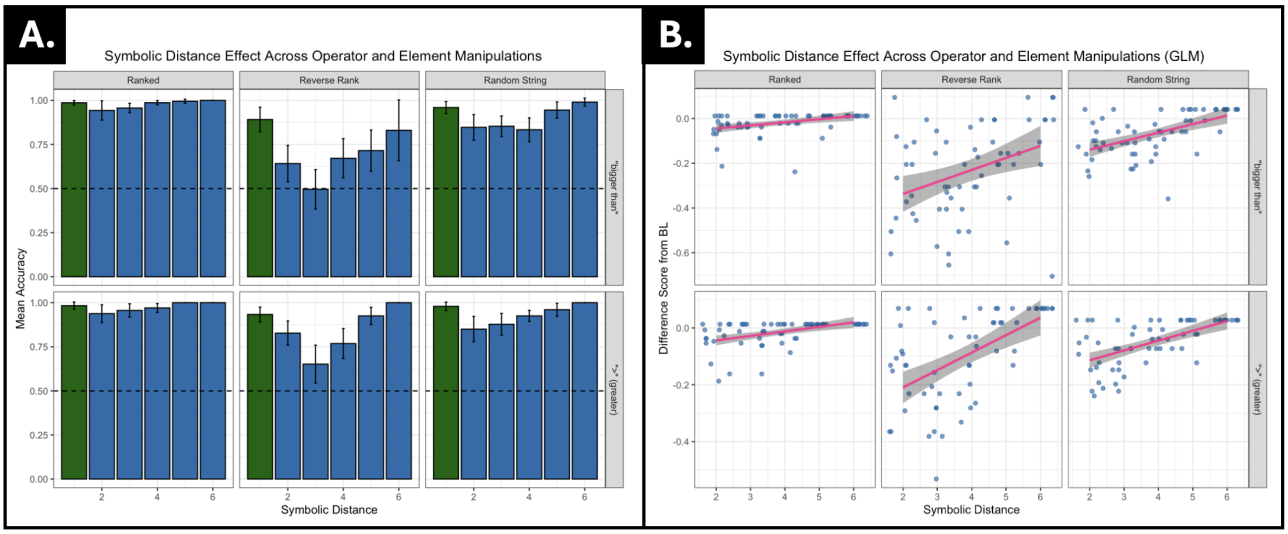

Figure 2. Performance of GPT-4o in a transitive inference task utilizing a single-trial inquiry and the Symbolic Distance Effect. Transitive descriptors were replaced [*"bigger" (top rows) vs. ">"(bottom rows)*] and sample stimuli were changed [*rank (left column): animal names ranked from largest to smallest size vs. reverse rank (middle column): animal names in reverse and incorrect size order vs. random string (right column): eight random characters*] in the base prompt (see Figure 1). Seven sample stimuli elements (pulled from a bank of 10 possible elements) and six pairs were simultaneously presented within one trial. All possible element combinations were queried for each prompt; the distance between elements is the "Symbolic Distance." Elements that were presented as pairs in the prompt had a symbolic distance of 1, elements that were one step removed had a symbolic distance of 2 and so on, with a maximum symbolic distance of 6. ***Panel A***: mean accuracy of each element query compared by symbolic distance. Green bars represent "familiar" queries that were featured in the initial prompt, while blue bars represent queries where transitive logic could be used; plots include 95% CI error bars. Dashed line shows chance performance (50%). ***Panel B***: evidence that, for this task presentation, GPT-4o shows the "Symbolic Distance Effect" also present in humans and animals, where accuracy is positively correlated with symbolic distance between tested element pairs. Jittered points show the relative difference between the accuracy of each query type and the baseline (BL) accuracy (i.e., accuracy on the distance 1 queries; green bars in Panel A) for each condition. Generalized linear model shown in pink with 95% CI shading.

lead us to sweeping "yes" or "no" conclusions, while the answer often lies somewhere in between (**P3**). Figure 2 and Table 1 illustrate just how much of a distribution exists between those two overly simplistic answers. If we ignore this nuance, we risk missing critical details about an LLM's behavioral patterning. The five core principles we have laid out in this paper are a distillation of many decades of animal cognition researchers grappling with similar behavioral confounds – animals often only demonstrate advanced capabilities like TI under certain conditions, their performance can easily be affected by noise in the environment, and each animal species operates with unique sensory modalities and in a particular ecological niche. Having to probe for intelligent behavior under these constraints has led to a rich literature of concepts, principles, experimental paradigms and theoretical frameworks that, when applied to LLMs and other foundation models, have the potential to illuminate the full nature of the model's behavior.

It is particularly difficult to understand and characterize LLM behavior because their surface-level responses may not reflect a deeper conceptual understanding. Grounding work on LLM evaluation in the existing scaffolding pro-

vided by animal cognition experiments and paradigms gives us a foundation on which to develop in-depth behavioral studies that can truly probe concept-level understanding in LLMs. By putting our five principles into practice, we go beyond simplistic binary statements of whether LLMs "do" or "don't" exhibit advanced concept learning and reasoning capabilities. Instead, these five principles can help us conduct in-depth, nuanced analyses of the range of behavior these models exhibit—and surface the capacities they may (or may not) possess. This work introduces these principles and takes a first step towards demonstrating how they can concretely be used in empirical studies of advanced concept learning and reasoning abilities in LLMs.

Acknowledgements

The authors would like to thank the Diverse Intelligences Summer Institute (DISI) and funding from the John Templeton Foundation (Grant 63138) for making this collaboration possible.

Impact Statement

This paper's contribution is designed to improve LLM evaluation protocols, with a particular focus on improving models' robustness, trustworthiness, and safety. As such, there are substantial societal implications of our work, which are discussed in the main text.

References

Alem, S., Perry, C. J., Zhu, X., Loukola, O. J., Ingraham, T., Søvik, E., and Chittka, L. Associative mechanisms allow for social learning and cultural transmission of string pulling in an insect. *PLoS biology*, 14(10):e1002564, 2016.

Bodily, K. D., Katz, J. S., and Wright, A. A. Matching-to-sample abstract-concept learning by pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 34(1):178–184, 2008. doi: 10.1037/0097-7403.34.1.178.

Boesch, C. Identifying animal complex cognition requires natural complexity. *Iscience*, 24(3), 2021.

Brannon, E. M. and Terrace, H. S. Representation of the numerosities 1–9 by rhesus macaques (Macaca mulatta). *Journal of Experimental Psychology: Animal Behavior Processes*, 26(1):31–49, 2000.

Bryant, P. E. and Trabasso, T. Transitive inferences and memory in young children. *Nature*, 232:456–458, 1971. doi: 10.1038/232456a0.

Bryson, J. J. and Leong, J. C. S. Primate errors in transitive 'inference': A two-tier learning model. *Animal Cognition*, 10:1–15, 2006.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Budaev, S., Jørgensen, C., Mangel, M., Eliassen, S., and Giske, J. Decision-making from the animal perspective: bridging ecology and subjective cognition. *Frontiers in Ecology and Evolution*, 7:164, 2019.

Burt, C. et al. Experimental tests of higher mental processes and their relation to general intelligence. 1911.

Byrne, R. W. and Whiten, A. Computation and mindreading in primate tactical deception. 1991.

Cheney, D. L. and Seyfarth, R. M. The recognition of social alliances by vervet monkeys. *Animal Behaviour*, 34(6): 1722–1731, 1986.

Chollet, F. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

De Lillo, C., Floreano, D., and Antinucci, F. Transitive choices by a simple, fully connected, backpropagation neural network: Implications for the comparative study of transitive inference. *Animal Cognition*, 4:61–68, 2001.

DiMattia, B. V. and Kesner, R. P. Serial position curves in rats: Automatic versus effortful information processing. *Journal of Experimental Psychology: Animal Behavior Processes*, 10(4):557–563, 1984. doi: 10.1037/0097-7403.10.4.557.

Duncan, J., Seitz, R. J., Kolodny, J., Bor, D., Herzog, H., Ahmed, A., Newell, F. N., and Emslie, H. A neural basis for general intelligence. *Science*, 289(5478):457–460, 2000.

Dusek, J. A. and Eichenbaum, H. The hippocampus and memory for orderly stimulus relations. *Proceedings of the National Academy of Sciences*, 94:7109–7114, 1997.

D'amato, M. and Colombo, M. The symbolic distance effect in monkeys (cebus apella). *Animal Learning & Behavior*, 18(2):133–140, 1990.

Epstein, R. The principle of parsimony and some applications in psychology. *The Journal of Mind and Behavior*, 5(2):119–130, 1984.

Farrer, D. N. Picture memory in the chimpanzee. *Perceptual and Motor Skills*, 25(1):305–315, 1967.

Frank, M. J., Rudy, J. W., and O'Reilly, R. C. Transitivity, flexibility, conjunctive representations, and the hippocampus. ii. a computational analysis. *Hippocampus*, 13:341–354, 2003.

Gopnik, A. and Wellman, H. M. Why the child's theory of mind really is a theory. 1992.

Greene, S. L. Feature memorization in pigeon concept formation. *Discrimination processes*, 1983.

Halina, M. Methods in comparative cognition. 2023.

Hall, K. and Brosnan, S. F. Cooperation and deception in primates. *Infant Behavior and Development*, 48:38–44, 2017.

Halpern, J. Y. Sufficient conditions for causality to be transitive. *Philosophy of Science*, 83(2):213–226, 2016.

Harris, M. and McGonigle, B. A model of transitive choice. *Quarterly Journal of Experimental Psychology*, 47B:319–348, 1994.

Howard, S. R. and Barron, A. B. Understanding the limits to animal cognition. *Current Biology*, 34(7):R294–R300, 2024.

Janmaat, K. R. What animals do not do or fail to find: A novel observational approach for studying cognition in the wild. *Evolutionary Anthropology: Issues, News, and Reviews*, 28(6):303–320, 2019.

Kirkman, C., Wan, H., and Hackenberg, T. D. A behavioral-economic analysis of demand and preference for social and food reinforcement in rats. *Learning and Motivation*, 77:101780, 2022. ISSN 0023-9690.

Kline, M. A. How to learn about teaching: An evolutionary framework for the study of teaching behavior in humans and other animals. *Behavioral and Brain sciences*, 38: e31, 2015.

Kosinski, M. Evaluating large language models in theory of mind tasks. *arXiv e-prints*, pp. arXiv–2302, 2023.

Kosinski, M. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121, 2024.

Lewis, M. and Mitchell, M. Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. *arXiv preprint arXiv:2402.08955*, 2024.

Macedonia, J. M. and Evans, C. S. Essay on contemporary issues in ethology: Variation among mammalian alarm call systems and the problem of meaning in animal signals. *Ethology*, 93(3):177–197, 1993.

Martin, A. and Santos, L. R. What cognitive representations support primate theory of mind? *Trends in cognitive sciences*, 20(5):375–382, 2016.

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., and Schütze, H. Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*, 117(42):25966–25974, 2020. doi: 10.1073/pnas.1910416117.

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

McGonigle, B. and Chalmers, M. Representations and strategies during inference. In McGonigle, B. (ed.), *Reasoning and Discourse Processes*, pp. 141–164. Academic Press, London, 1986.

Momennejad, I., Hasanbeig, H., Vieira Frujeri, F., Sharma, H., Jojic, N., Palangi, H., Ness, R., and Larson, J. Evaluating cognitive maps and planning in large language models with cogeval. *Advances in Neural Information Processing Systems*, 36, 2024.

Nematipour, B., Bračić, M., and Krohs, U. Cognitive bias in animal behavior science: A philosophical perspective. *Animal Cognition*, 25(4):975–990, 2022.

Rabovsky, M., Hansen, S. S., and McClelland, J. L. Modelling the n400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, 2 (9):693–705, 2018.

Ramawat, S., Marc, I. B., Ceccarelli, F., Ferrucci, L., Bardella, G., Ferraina, S., Pani, P., and Brunamonti, E. The transitive inference task to study the neuronal correlates of memory-driven decision making: A monkey neurophysiology perspective. *Neuroscience and Biobehavioral Reviews*, 152:105258, 2023.

Rane, S. The Reasonable Person Standard for AI. In *Forty-first International Conference on Machine Learning*, 2024.

Rane, S., Ku, A., Baldridge, J., Tenney, I., Griffiths, T., and Kim, B. Can generative multimodal models count to ten? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

Roberts, W. A. and Phelps, M. T. Transitive inference in rats: A test of the spatial coding hypothesis. *Psychological Science*, 5(6):368–374, 1994.

Rumelhart, D. E. *Some problems with the notion of literal meanings*, pp. 71–82. Cambridge University Press, 1993.

Samhita, L. and Gross, H. J. The "clever hans phenomenon" revisited. *Communicative & integrative biology*, 6(6): e27122, 2013.

Shettleworth, S. J. Animal cognition and animal behaviour. *Animal behaviour*, 61(2):277–286, 2001.

Shettleworth, S. J. Cognitive science: rank inferred by reason. *Nature*, 430(7001):732–733, 2004. doi: 10.1038/430732b.

Shettleworth, S. J. Do animals have insight, and what is insight anyway? *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 66(4):217, 2012.

Siemann, M. and Delius, J. D. Algebraic learning and neural network models for transitive and non-transitive responding. *European Journal of Cognitive Psychology*, 10:307–334, 1998.

Strachan, J. W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., et al. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, pp. 1–11, 2024.

Thornton, A. and McAuliffe, K. Teaching in wild meerkats. *Science*, 313(5784):227–229, 2006.

Tibbetts, E. A., Agudelo, J., Pandit, S., and Riojas, J. Transitive inference in polistes paper wasps. *Biology letters*, 15(5):20190015, 2019.

Vasconcelos, M. Transitive inference in non-human animals: An empirical and theoretical analysis. *Behavioural Processes*, 78(3):313–334, 2008.

Wagner, A. R. and Rescorla, R. A. Inhibition in pavlovian conditioning: Application of a theory. In Boakes, R. A. and Haliday, M. S. (eds.), *Inhibition and Learning*, pp. 301–336. Academic Press, New York, 1972.

Washburn, D., Thompson, R., and Oden, D. Monkeys trained with same/different symbols do not match relations. In *38th Annual Meeting of the Psychonomic Society, Philadelphia, PA*, 1997.

Waters, R. H. Morgan's canon and anthropomorphism. *Psychological Review*, 46(6):534–540, 1939. doi: 10.1037/h0055191.

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

Woocher, F. D., Glass, A. L., and Holyoak, K. J. Position discriminability in linear orderings. *Memory & Cognition*, 6:165–173, 1978. doi: 10.3758/BF03197437.

Wu, X. and Levy, W. B. Simulating symbolic distance effects in the transitive inference problem. *Neurocomputing*, 40:1603–1610, 2001.

Wynne, C. D. L. A minimal model of transitive inference. In Wynne, C. D. L. and Staddon, J. E. R. (eds.), *Models of Action: Mechanisms for Adaptive Behavior*, pp. 269–307. Erlbaum, Mahwah, NJ, 1998.

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

Zentall, T. R. Morgan's canon: Is it still a useful rule of thumb? *Ethology*, 124(7):449–457, 2018. doi: 10.1111/eth.12750.

Zhang, S., Bock, F., Si, A., Tautz, J., and Srinivasan, M. V. Visual working memory in decision making by honey bees. *Proceedings of the National Academy of Sciences*, 102(14):5250–5255, 2005.