Principles of Animal Cognition for LLM Evaluations: A Case Study on Transitive Inference

Sunayana Rane*
Department of Computer Science
Princeton University
srane@princeton.edu

Cyrus F. Kirkman*
Department of Psychology
University of California Los Angeles
cyruskirkman@g.ucla.edu

Graham Todd*
Department of Computer Science and Engineering
New York University Tandon
gdrtodd@nyu.edu

Amanda Royka*
Department of Psychology
Yale University
amanda.royka@yale.edu

Ryan M.C. Law*
MRC Cognition and Brain Sciences Unit
University of Cambridge
ryan.law@mrc-cbu.cam.ac.uk

Erica A Cartmill†
Departments of Anthropology and Psychology
Indiana University Bloomington
ericac@iu.edu

Jacob G. Foster†
Department of Informatics
Luddy School of Informatics, Computing, and Engineering
Cognitive Science Program
Indiana University Bloomington
jacobf@iu.edu

Abstract

It has become increasingly challenging to understand and evaluate LLM capabilities as these models exhibit a broader range of behaviors. We respond to this challenge by taking inspiration from another field which has developed ideas, practices, and paradigms for probing the behavior of complex intelligent systems: animal cognition. We present five core principles from animal cognition, and explain how they provide invaluable guidance for understanding LLM abilities and behavior. We ground these principles in an empirical case study, and show how they can provide a richer contextual picture of one particular reasoning capability: transitive inference.

1 Motivation

Evaluating the capacities of large language models (LLMs) presents a unique challenge. LLMs can converse in a sustained and often coherent manner and demonstrate some capabilities in domains such as Theory of Mind [12, 19], and analogical reasoning [22]. Nonetheless, LLMs fail on seemingly simple tasks, often in unpredictable ways. As such, it can be difficult to determine whether LLMs successes reflect true emergent cognitive capacities or whether their performance stems from shallower

*Equal contribution as first author.
†Equal advising/senior authors. Listed alphabetically.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

solutions, such as recapitulation of patterns in training data. Furthermore, existing LLM evaluation methods often report only a single numerical performance metric, limiting the inferences we can draw about the full extent and limits of a model's abilities.

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 alternate 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, 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:

1. Design control conditions with an adversarial attitude (P1) [10, 2, 9, 16]
2. Establish robustness to variations in stimuli (P2) [3, 7, 8]
3. Delve into failure types, moving beyond a success/failure dichotomy (P3) [13, 11, 18, 21]
4. Clarify differences between mechanism and behavior (P4) [17, 14, 1]
5. Meet the organism (or, more broadly, intelligent system) where it is, while noting systemic limitations (P5) [10, 24, 4]

These principles provide broad guidance that can help our machine learning research community effectively study the behavioral properties of LLMs and other foundation models. In addition to identifying these principles, we demonstrate their value in an empirical case study. We select a cognitive capability probed widely and thoroughly across animal literature: Transitive Inference (TI), or the ability to extrapolate ordinal relationships between items that have not been directly compared. Not only is TI well-studied across the animal kingdom from wasps [20] to rats [15] to humans [5]; it is the kind of domain-general reasoning capacity that is fundamental to other capacities that would be desirable for LLMs to exhibit, such as numerosity and causality. [23]

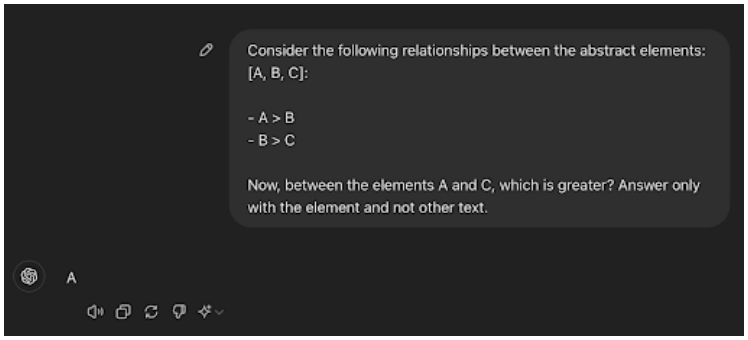

Figure 1: One task structure for probing TI in our series of TI tasks adapted to LLM's modalities.

3 Methods

What does probing Transitive Inference in an LLM look like? Of course, it's possible to simply query a model directly to define or demonstrate TI (and they will generally do so quite persuasively). 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 1. We note which of our core principles (P1-P5) are used in each set of experiments.

3.1 "greater" and "bigger" operators

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 "greater" or "bigger." Figure 1 shows this type of task structure. Varying the "greater" and "bigger" operators serves as a simple adversarial control (**P1**); if transitive inference were being used, performance should be insensitive to this variation.

3.2 Symbolic Distance Effect

When presented with TI tasks, humans, pigeons, and macaque monkeys are more accurate when making comparisons between abstract elements at a greater distance (i.e., those having more intervening terms between them). This is known as the Symbolic Distance Effect [6]. We report experimental results of measuring for an analogous symbolic distance effect in the LLM's responses 3. Having demonstrated the expected behavior with the operator "greater", we probe it for the operator "bigger", testing for robustness to variations in stimuli (**P2**). We then analyze the specific pattern of failures (**P3**) as a function of variation in stimulus.

3.3 Trial Structure

Our experiments on the symbolic-distance effect demonstrate the characteristic brittleness of LLMs to minor prompt variations; this was revealed through analyzing the pattern of failures (**P3**). While these findings indicate a lack of robustness in TI ability, we turn to our fifth principle (**P1**) and attempt to "meet the organism where it is" by going beyond a simple claim about brittleness to minor prompt variation. Instead, we try to find a setting where a system with transitive inference capabilities should exhibit the expected behavior (without noise from complex prompts or minor prompt variation). We turn to a simpler trial structure that is frequently used in animal cognition studies. This structure minimizes word use and instead focuses on a deep look at the underlying behavior.

This new control condition closely resembles the trial-based learning version of the task presented to animal subjects. This condition required the language model to integrate ordinal information over a sequence of interactions. With this condition, we can start to separate mechanism from behavior (**P4**); if the same mechanism (transitive inference) were used across tasks, we would expect similar performance on this task variant. Instead, performance dropped to chance (see Figure 2, control conditions increasing in complexity from left to right).

4 Results

While the model showed excellent performance on initial versions of this task, performance dropped significantly after the introduction of control conditions to address alternative explanations (Figure 2). In addition, 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 3). Taken together, this probing demonstrates the lack of a generalizable TI ability in GPT-4o.

5 Discussion

While these results should cast serious doubt on the deductive reasoning capacities of LLMs, they also show how careful study design inspired by animal cognition can give a deeper insight into whether and where such important abilities are present. For example, when designing an experiment to answer a simple question like "Do LLMs understand the concept of transitivity?," minor prompt variations can lead us to sweeping "yes" or "no" conclusions, while the answer often lies somewhere in between. Figures 2 and 3 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 behavior pattern. 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

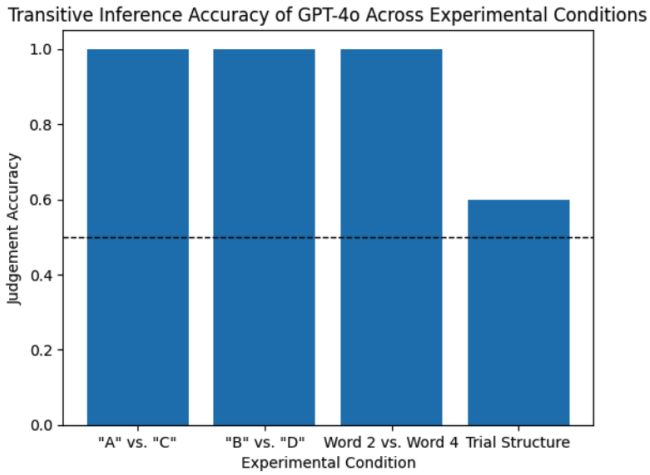

Figure 2: Accuracy of the GPT-4o language model in a transitive inference task across a variety of task structures. Performance was perfect across the first three conditions, but fell dramatically in the last, indicating that the model does not robustly demonstrate TI abilty.

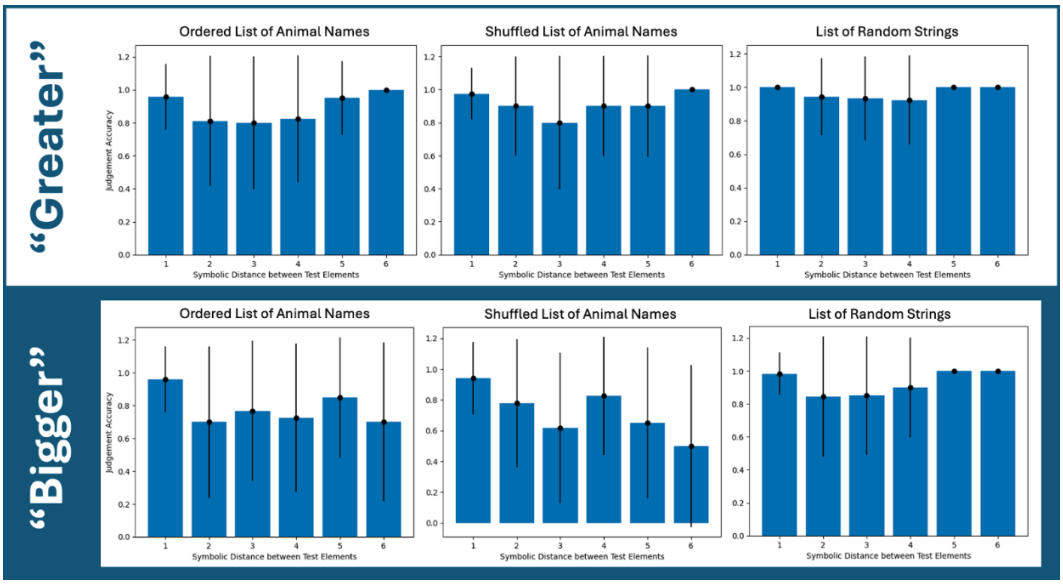

Figure 3: Performance of GPT-4o in a transitive inference task utilizing a single-trial and sentence-structured inquiry. A single transitive descriptor was replaced ["greater" (top) vs. "bigger" (bottom)] in the following prompt: Consider the following relationships between the abstract elements: A is greater than B, B is greater than C. . . , and F is greater than G. Now, between the elements elem1 and elem2, which is greater? Relative accuracy is provided with 95% CI. Animal name lists were either ordered by relative size of animal or shuffled; random strings were thirty-six random characters. In similar tasks, humans, pigeons, and macaque monkeys are more accurate when making comparisons between abstract elements between more intervening terms, also known as the Symbolic Distance Effect. However, GPT-4o's errors exhibit the opposite pattern in two versions of the task: the LLM is more accurate for elements with smaller symbolic distances, which is consistent with associative, non-inferential models of TI performance.

demonstrate advanced capabilities like TI under certain conditions, their performance can easily be affected by noise in the environment, and each animal species operates in only their own modalities. Having to probe for intelligent behavior under these constraints has led to a rich literature of ideas,

principles, experimental paradigms and theoretical frameworks that, when applied to LLMs and other foundation models, have the potential to greatly 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 provided 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, credible 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.

References

[1] Sylvain Alem et al. "Associative mechanisms allow for social learning and cultural transmission of string pulling in an insect". In: *PLoS biology* 14.10 (2016), e1002564.

[2] Christophe Boesch. "Identifying animal complex cognition requires natural complexity". In: *Iscience* 24.3 (2021).

[3] Elizabeth M Brannon and Herbert S Terrace. "Representation of the numerosities 1–9 by rhesus macaques (Macaca mulatta)." In: *Journal of Experimental Psychology: Animal Behavior Processes* 26.1 (2000), p. 31.

[4] Sergey Budaev et al. "Decision-making from the animal perspective: bridging ecology and subjective cognition". In: *Frontiers in Ecology and Evolution* 7 (2019), p. 164.

[5] Cyril Burt et al. "Experimental tests of higher mental processes and their relation to general intelligence". In: (1911).

[6] MR D'amato and M Colombo. "The symbolic distance effect in monkeys (Cebus apella)". In: *Animal Learning & Behavior* 18.2 (1990), pp. 133–140.

[7] Donald N Farrer. "Picture memory in the chimpanzee". In: *Perceptual and Motor Skills* 25.1 (1967), pp. 305–315.

[8] Sharon L Greene. "Feature memorization in pigeon concept formation". In: *Discrimination processes* (1983).

[9] Marta Halina. "Methods in comparative cognition". In: (2023).

[10] Scarlett R Howard and Andrew B Barron. "Understanding the limits to animal cognition". In: *Current Biology* 34.7 (2024), R294–R300.

[11] Karline RL Janmaat. "What animals do not do or fail to find: A novel observational approach for studying cognition in the wild". In: *Evolutionary Anthropology: Issues, News, and Reviews* 28.6 (2019), pp. 303–320.

[12] Michal Kosinski. "Evaluating large language models in theory of mind tasks". In: *arXiv e-prints* (2023), arXiv–2302.

[13] Alia Martin and Laurie R Santos. "What cognitive representations support primate theory of mind?" In: *Trends in cognitive sciences* 20.5 (2016), pp. 375–382.

[14] Behzad Nematipour, Marko Bračić, and Ulrich Krohs. "Cognitive bias in animal behavior science: A philosophical perspective". In: *Animal Cognition* 25.4 (2022), pp. 975–990.

[15] William A Roberts and Maria T Phelps. "Transitive inference in rats: A test of the spatial coding hypothesis". In: *Psychological Science* 5.6 (1994), pp. 368–374.

[16] Laasya Samhita and Hans J Gross. "The "clever Hans phenomenon" revisited". In: *Communicative & integrative biology* 6.6 (2013), e27122.

[17] Sara J Shettleworth. "Animal cognition and animal behaviour". In: *Animal behaviour* 61.2 (2001), pp. 277–286.

[18] Sara J Shettleworth. "Do animals have insight, and what is insight anyway?" In: *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 66.4 (2012), p. 217.

[19] James WA Strachan et al. "Testing theory of mind in large language models and humans". In: *Nature Human Behaviour* (2024), pp. 1–11.

[20] Elizabeth A Tibbetts et al. “Transitive inference in Polistes paper wasps”. In: *Biology letters* 15.5 (2019), p. 20190015.

[21] DA Washburn, RKR Thompson, and DL Oden. “Monkeys trained with same/different symbols do not match relations”. In: *38th Annual Meeting of the Psychonomic Society, Philadelphia, PA*. 1997.

[22] Taylor Webb, Keith J Holyoak, and Hongjing Lu. “Emergent analogical reasoning in large language models”. In: *Nature Human Behaviour* 7.9 (2023), pp. 1526–1541.

[23] Thomas R Zentall et al. “Concept learning in animals”. In: *Comparative Cognition & Behavior Reviews* 3.1 (2008), pp. 13–45.

[24] Shaowu Zhang et al. “Visual working memory in decision making by honey bees”. In: *Proceedings of the National Academy of Sciences* 102.14 (2005), pp. 5250–5255.