Contextual Effects in LLM and Human Causal Reasoning

Zach Studdiford [1 2] **Gary Lupyan** [1]

Abstract

What type of knowledge is required to infer the outcomes of everyday actions such as a glass being more likely to break when it falls onto tile than onto a carpet? One possibility is that such inferences requires highly robust and general world models. Another possibility is that making such inferences is a much more contextual process, the success of which depends on the particulars of the scenario being probed. We evaluate causal inferences in people and LLMs and show that although human accuracy far exceeds that of LLMs, there is a surprising degree of alignment in human and LLM performance. Both show a high degree of specificity. Seemingly superficial differences in probing causal knowledge matter for both people and LLMs. We then show that prompts that elicit more integrated patterns of attention predict both higher model accuracy and closer alignment to human performance.

1. Introduction and related work

What information does a transformer get "for free" when it learns to predict the next word in a sentence? Accurately predicting the continuation of "*Ali bumps the vase off the table and the vase BLANK*" requires knowing that vases shatter, that things move when bumped, and that these events unfold in a coherent causal sequence. Moreover, the ability to generalize this knowledge to unfamiliar sentences– such as predicting the outcome of *"Ali knocks the **lamp** off the table and the vase BLANK"* indicates a capacity for reasoning beyond memorized associations, suggesting some internalized understanding of the relations this text describes.

Large language models (LLMs), trained to reduce next-token prediction error demonstrate capacities in reasoning[3] about physical causality (Kıcıman et al., 2023) and temporal sequences (Xiong et al., 2024) despite only being exposed to these relations in text. This gap between the words an LLM makes predictions over and the real-world meanings these words refer to, raises the question of whether for the model, these linguistic symbols are grounded in any meaningful way (Mollo & Millière, 2023). Although these models cannot perceive or intervene on the world, they are trained on language produced by humans who do (Pearl, 2009; Gopnik, 2010). The language humans generate, far from reflecting mere associations, encodes the statistical regularities of the temporal sequences (Talmy, 1995), agentic relationships, and (Levison & Lessard, 1990), and causal structure (Gleitman & Gleitman, 1997) of events in the world. Language is used not only to describe these events, but also functions to draw attention to the relevant details, foregrounding some elements and backgrounding others. This is often accomplished by subtle differences in syntax: "The horse was tied to a tree" and "The tree was tied to a horse" have different causal implications (Gleitman et al., 1996). Syntactic constructions in some languages (Spanish, Japanese) de-emphasize the role of an agent in causing something to happen ("The vase broke itself"), while constructions in others (English) do the opposite ("Ali broke the vase") (Fausey & Boroditsky, 2008). In short, causal structure is amply represented in language.

If LLMs have indeed learned the causal knowledge encoded in language, what form does that knowledge take? One possibility is that causal structure is encoded in abstract world models: structured, symbolic representations supporting reasoning and generalization over temporal sequences of events (Tenenbaum & Niyogi, 2003). These models have been proposed as central to human causal cognition in developmental (Goddu & Gopnik, 2024) and computational (Tenenbaum & Griffiths, 2002) accounts of how humans causal learning. By making probabilistic hypotheses about the world over latent variables encoded in the model, learners can infer unseen relations from sparse input data and make robust generalizations to new situations under uncertainty (Tenenbaum & Niyogi, 2003). These models have been used to explain how humans acquire knowledge of intuitive physics (Xu et al., 2021), infer goal-directed behavior, and update their world knowledge during play in childhood (Gopnik,

[1]Department of Psychology, University of Wisconsin-Madison [2]Department of Computer-Science, University of Wisconsin-Madison. Correspondence to: Zach Studdiford <studdiford@wisc.edu>.

ICML 2025 Workshop on Assessing World Models.

[3]We use the term "reasoning" broadly, applying it both to logical reasoning and common-sense reasoning that is the target of the present investigation.

2010; Goddu & Gopnik, 2024).

Another possibility is that reasoning is supported by a form of context-sensitive pattern-matching—where inputs are used to construct context-sensitive schemas learned through experience with specific events—without necessarily invoking an abstract/generative model. Margolis (1987) characterize this as reasoning by "re-cognition"—reusing familiar patterns when interpreting new situations. Chater & Oaksford (1999) argue that much of human cognition can be explained as probabilistic tuning to the statistical structure of the environment, rather than reasoning logically over a symbolic world model. Indeed, classic reasoning experiments such as the Wason selection task show that while people systematically fail to reason over a set of logical relations abstractly, they can do so effectively when the same relations are framed in familiar schemas (Wason, 1968). Similarly, people often infer spatial relationships by drawing on contextually grounded representations as opposed to symbolic models, erroneously exchanging relative, egocentric terms like "in front of" for "North of", and assuming alignment in these directions (Tversky, 1992). This preference for surface-level familiarity over symbolic models has been observed in domains such as category and concept learning (Verheyen et al., 2008) probabilistic judgment (Kahneman & Tversky, 1972) and analogical reasoning (Gentner & Maravilla), suggesting that deliberative and symbolic *system two reasoning* is the exception rather than the rule in human cognition Kahneman (2011)

In efforts to align LLM reasoning with that of humans, recent work has argued for the augmentation of neural networks trained for next word prediction with abstract world models (LeCun, 2022). However, we argue that evidence from cognitive science suggests human reasoning often relies on heuristics/schemas learned from familiar contexts, inferring the outcome of novel scenarios by drawing on these context-specific patterns. To the extent that these patterns are embedded in the distribution of word co-occurrences in language, we expect that a model that has learned this distribution will converge to similar human-like errors in reasoning. To test this claim, we evaluate everyday causal inference scenarios in both humans and LLMs. Specifically, we ask whether variation in the *attention mechanism* of LLMs predicts human accuracy for scenarios that require similar world models, with arbitrary differences in content. Such a relationship would indicate that causal reasoning in both humans and LLMs is often dependent on statistical patterns embedded in language, as opposed to abstract symbolic models.

2. Methods

2.1. Causal Reasoning Assessment

To evaluate world models in LLMs and humans, we design a set of stimuli evaluating causal reasoning in everyday situations, concerning grounded interactions and relations between agents and objects. Drawing inspiration from Ivanova et al. (2024), each stimulus is composed of two parts (C, R): a context state C and an outcome R. For example: C: *The painting is in front of Ali. Ali turns around.* R: *The painting is behind Ali*. We constructed 429 prompts in 11 categories spanning egocentric and geocentric spatial relations, references between people, states of objects, actions, and two *nonfollows* categories, where the action included in C does not produce any change to the described situation. Full descriptions and examples of all categories can be seen in Figure 1.

For each prompt, participants and LLMs were presented with a BLANK which must be filled with a binary choice option. For example, *"The painting is in front of Ali. Ali turns around. The painting is BLANK Ali."*, with options being (*in front*, *behind*). To test the robustness of causal inference, we included a *result-completion* condition where the BLANK appears in the end (result segment) of the prompt (*"Which result follows from the premises?"*), and a *context-completion* condition where the BLANK appears in the beginning (context segment) (*"Which context must be true given the result outcome?"*).

2.2. Evaluating human causal inference

An initial group of 142 participants recruited from Amazon Mechanical Turk were shown 63 prompts each. On each trial, participants saw the full prompt (including the "BLANK") and pressed the spacebar to reveal the two options after which they could choose which option best completed the prompt. This design allowed us to measure two reaction times (RTs), a prompt processing time (*read-RT*) and a response time *choice-RT*, defined as the time between the spacebar press and the choice of one of the two options. Accuracy was defined as the proportion of correct choices.

2.2.1. Testing consistency of human responses

To evaluate whether human errors were robust or reflected random variation, we collected additional data from 80 participants on the 259 prompts that had accuracies below 80%, tested in bins of 37 prompts. After going through all the prompts, these participants were again shown prompts for which they made an error, as well as a random sample of 40% of prompts for which they responded correctly, and given the opportunity to respond again as well as to justify their answer.

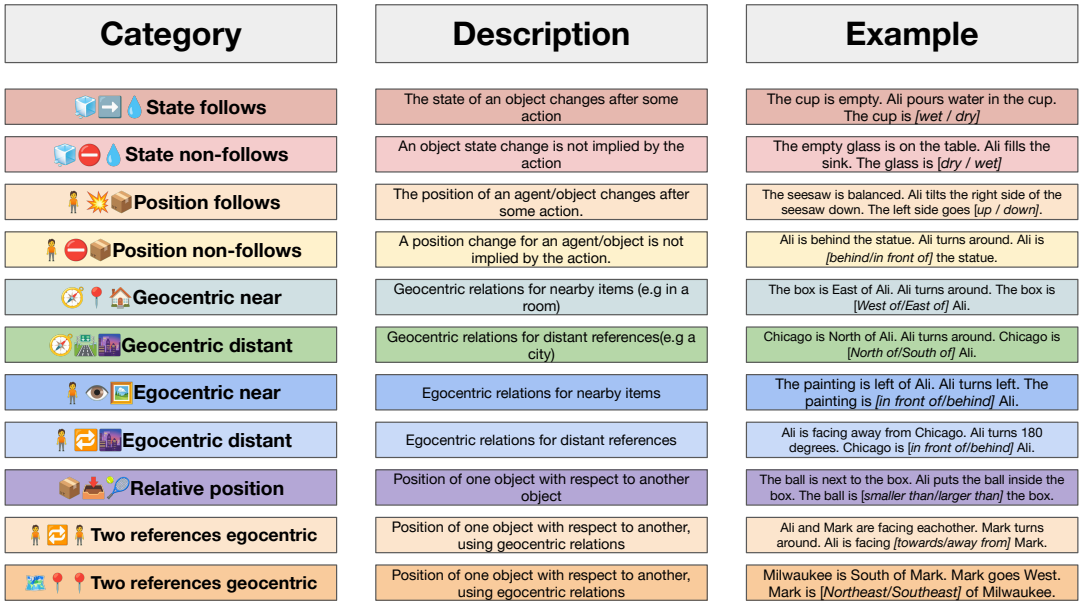

Figure 1. We evaluate 11 categories of causal reasoning in humans and LLMs, including cases where the action changes the outcome vs. not, and a variety of spatial transformations including uses of egocentric and allocentric reference frames. . The two options for each BLANK are in brackets for all examples, with the correct choice being the first listed. Appendix C contains the prompts with with the highest and lowest accuracies for for humans and LLMs for each category.

2.3. Evaluating LLM causal inference

We evaluated several open source LLMs: `gemma-2-2b`, `gemma-2-9b`, `gemma-2-27b` as well as models with fewer parameters: `gpt2-small` and `gpt2-xl`. We also queried three frontier models: `gpt-4`, `gpt-4.5`, and `gpt-o3`. For the open source gemma models, we compared the log probabilities of the two response option tokens as a measure of accuracy, computed as $\log p(res_{correct}) - \log p(res_{incorrect})$. For closed-source models we measured accuracy as a binary correct/incorrect response.

2.3.1. Measure of LLM Attention

To quantify the extent to which LLMs attend to the prompt (C, R), we computed an attention-weighted lookback vector $\mathbf{v}_{\text{attn}}^{(\ell)}$ at each layer ℓ, where each component corresponds to the average attention mass allocated to tokens at a specific normalized distance bin (1/12 of the prompt length). Specifically, given the attention mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V,$$

We extracted non-normalized attention scores $QK^\top$ at the final token x_t for each layer ℓ, yielding a distribution over prior tokens. Each score was weighted by the relative lookback distance from x_t (scaled to $[0, 1]$), then aggregated into 12 equal-length bins. The resulting vector $\mathbf{v}_{\text{attn}}^{(\ell)} \in \mathbb{R}^{12}$ has uniform dimensionality for all prompts, and captures the average distribution of attention at each token range. To quantify the spread of this distribution, we computed a scalar value of entropy for each layer vector for all prompts. Intuitively, a higher entropy score $\mathcal{H}(\mathbf{v}_{\text{attn}}^{(\ell)})$ indicates the distribution of model attention is more diffuse while processing the prompt.

3. Results

In both humans and LLMs[4], we observe strikingly low absolute accuracies ($\mu_{\text{human}} = 0.71$, $SD_{\text{human}} = 0.16$; $\mu_{\text{LLM}} = 0.54$, $SD_{\text{LLM}} = 0.22$). Additionally, human accuracies are consistent when retested for all but three categories (*geocentric-near*: $t = 10.05$; *relative-position*: $t = 5.16$; *geocentric-two-references*: $t = 4.54$, $p < .01$), suggesting that fail states are consistent and not the result of random error. Despite relatively low accuracy in humans and seemingly near-chance accuracy in LLMs, people and models are surprisingly well aligned at the item level ($r = 0.51$) and even more well aligned at the category level ($r = 0.88$). That is, the prompts on which the models score

[4]All LLM results reported in this section refer to `gemma-2-27b-it` unless otherwise noted (the open source model showing the highest correlation with human accuracy). Results for all models can be seen in Appendix C.

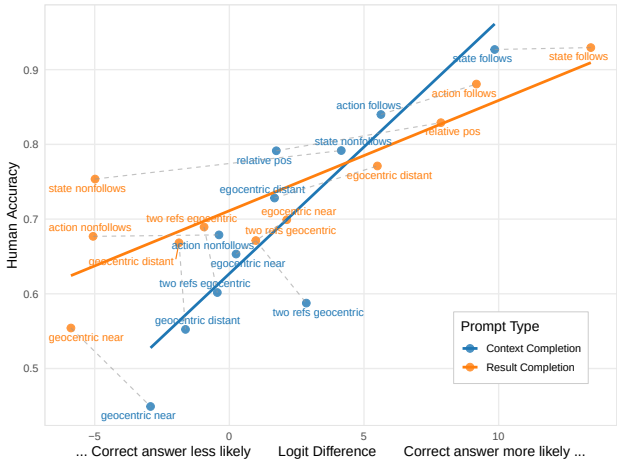

Figure 2. LLM and Human category accuracy for context and target prompt types. The x axis indicates the average LLM logit difference between the correct and incorrect response for a category and prompt format. The y axis indicates mean human response. Categories in the upper left above the regression lines have higher mean accuracy for people relative to LLMs, while categories in the lower right below the regression lines have higher mean accuracy for LLMs relative to people.

systematically below chance are (largely) the same prompts that give people the most trouble. One of the latest frontier "reasoning models" (`gpt-o3`) scores 0.83, which exceeds human performance. However, alignment with human accuracy is only marginally improved from `gemma-2-27b-it` (r_{o3} = 0.54, $r_{gemma-27b}$ = 0.51).

3.1. Types of misalignment

Despite high human-LLM correlations, inspection of Figure 1 reveals several clear cases of misalignment.

3.1.1. Differences in spatial reasoning

One source of misalignment is for categories related to spatial reasoning. Categories involving egocentric relations (*egocentric near*, *two references egocentric*, *egocentric distant*) are relatively easier for people and more difficult for LLMs. Conversely, categories testing allocentric representations of space (*geocentric near*, *geocentric distant*) show poor accuracy in both humans and LLMs. *two references geocentric* shows greater LLM performance relative to human accuracy, particularly for the context-completion condition.

3.1.2. Sensitivity to prompt format

Humans and LLMs also diverge in the effect of prompt format on accuracy. Both *state nonfollows* and *position nonfollows* show strong interactions with prompt type such that LLM performance improves for context prompt types of those categories by 0.45 and 0.21, respectively. Humans are also subject to differences in the prompt format: for *two references geocentric*, *geocentric near*, and *geocentric distant*), humans show significant improvement in the result-completion condition relative to context-completion. Interestingly, these same categories show no significant interaction effect for prompt-format in model accuracies. These marked improvements in accuracy are surprising given that there are minimal-differences in prompt-format result-completion and context-completion conditions, placing the BLANK in either the context or result portion of the stimulus.

3.2. LLM attention is uniquely predictive of human accuracy

The entropy of the attention distribution vector $\mathcal{H}(\mathbf{v}_{\text{attn}}^{(\ell)})$ emerges as a significant predictor of human accuracy, even when controlling for prompt category, prompt type (context or target condition), model accuracy, prompt length, and the average trigram frequency of the prompt ($\beta_{\max} = 0.1917$, $p < .001$). Surprisingly, this signal is *strongest* in smaller parameter models (`gemma-2-2b`). Although these models have chance accuracy and show minimal alignment in output logits with human accuracy ($r_{(gemma-2-2b,human)} = 0.09$, $\mu^{\text{accuracy}}_{\text{gemma-2-2b}} = 0.45$), their QK attention activations capture meaningful structure: the variance in human accuracy explained by the attention entropy of `gemma-2-2b` exceeds that predicted by accuracy in even frontier models such as `gpt-o3` ($R^2_{acc_{o3}} = 0.29$, $R^{2\,\mathcal{H}_{\text{attn}}}_{\text{gemma-2-2b}} = 0.31$). The high predictive power of QK^T activations in `gemma-2-2b`, despite low alignment in accuracy, suggests that the attention mechanism is predictive of human responses entirely independent of encoded semantic knowledge.

4. Discussion

Theories of abstract world models posit the existence of a de-contextualized symbol representation used in causal reasoning (Tenenbaum & Griffiths, 2002). However, our evaluation of LLM and human common-sense reasoning reveals consistent fail states and patterns of uncertainty in both LLMs and humans, for problems that on their face seem trivially easy. Our evaluations assess mundane, familiar actions and relations–dropping objects, turning around, walking forwards, orienting towards a cardinal direction–and yet we observe broad variance even for *logically equivalent questions* as a function of small changes to the prompt format in both populations. While there are some differences in the categories that are most difficult for LLMs and humans–LLMs struggling with egocentric relations while humans struggle with geocentric relations–the overall high

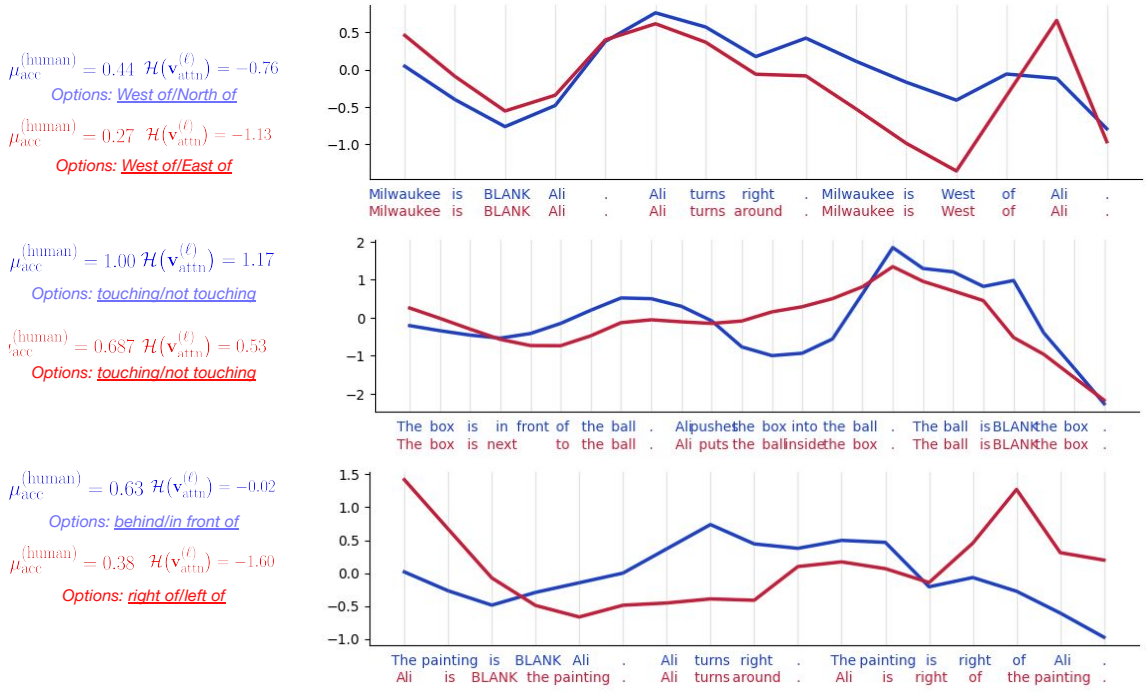

Figure 3. Distributions of the lookback measure $\mathbf{v}_{\text{attn}}^{(\ell)}$ for logically similar prompts at layer 12 (`gemma-2-2b`, the layer with greatest entropy prediction coefficient after controlling for confounds). Blue and red correspond to versions of the prompt with higher and lower accuracy and entropy, respectively. The completion options for the blank are indicated for all prompts, where the first option listed is the correct completion. Recall that the order of the completion options are randomized on each trial. Entropy units are normalized to zero mean and unit variance.

alignment in accuracies at both the item and category level suggests common factors explaining both systems' behavior.

Despite the differences in "training data" for LLMs and humans, the entropy of the attention distribution at intermediate layers of `gemma-2b` emerges as predictive of responses in both humans and LLMs above prompt type, prompt format, and model accuracy. This measure corresponds to how distributed the QK attention signal is across the prompt: high-entropy distributions correspond to diffuse attention, while low entropy distributions indicate concentration on a smaller subset of the prompt tokens. We observe a strong predictive relationship between higher entropy and higher accuracy in humans, despite the low human alignment of logit outputs in `gemma-2-2b`. Figure 3 shows exampes of logically similar prompts (within the same evaluation categories), where the entropy of the attention activations in `gemma-2-2b` predict large discrepancies in human accuracy, despite a small number of single word differences between prompts.

The present work is not the first to demonstrate converging fail states between humans and LLMs: Dasgupta et al. (2024) show evidence of content effects in LLMs similar to those seen in humans on the Wason task (Wason, 1968), and propose two possibilities: (1) that these biases are learned as a result of "parroting" the underlying human data generation process, or (2) that these content effects emerge naturally because they are generally useful for predicting semantic regularities reflected in text. We argue that the ability of activations in the QK^T matrix of small models to predict human accuracy beyond the output logits of much larger models (`gpt-o3`) suggests the latter: latent features in natural language reflect aspects of causality in the world, and these features are beneficial for predicting the outcomes of novel causal relations reflected in text. Moreover, large differences in *human* accuracy for logically similar prompts suggest that, in some instances, these heuristics might be exploited in lieu of consulting an abstract world model. In either case, the effectiveness of the attention mechanism in predicting human reasoning errors leaves open the possibility for future work exploring the extent to which latent features in the language distribution play a causal role in reasoning in both LLMs and humans.

References

Chater, N. and Oaksford, M. Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, 3(2): 57–65, 1999. doi: 10.1016/S1364-6613(98)01273-X.

Dasgupta, I., Lampinen, A. K., Chan, S. C. Y., Sheahan, H. R., Creswell, A., Kumaran, D., McClelland, J. L., and Hill, F. Language models show human-like content effects on reasoning tasks, 2024. URL https://arxiv.org/abs/2207.07051.

Fausey, C. M. and Boroditsky, L. English and spanish speakers remember causal agents differently. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30, pp. 1109–1114. Cognitive Science Society, 2008. URL https://escholarship.org/uc/item/4425600t.

Gentner, D. and Maravilla, F. Analogical reasoning. In Little, T. R. and Thompson, R. F. (eds.), *International Handbook of Thinking and Reasoning*.

Gleitman, L. and Gleitman, H. What is a language made out of? *Lingua*, 100(1–4):29–55, 1997. doi: 10.1016/S0024-3841(93)00036-8.

Gleitman, L. R., Gleitman, H., Miller, C., and Ostrin, R. Similar, and similar concepts. *Cognition*, 58(3):321–376, 1996. doi: 10.1016/0010-0277(95)00686-9. URL https://doi.org/10.1016/0010-0277(95)00686-9.

Goddu, M. K. and Gopnik, A. The development of human causal learning and reasoning. *Nature Reviews Psychology*, 3(5):319–339, 2024. doi: 10.1038/s44159-024-00300-5. URL https://doi.org/10.1038/s44159-024-00300-5.

Gopnik, A. How babies think. *Scientific American*, 303 (1):76–81, 2010. URL https://www.jstor.org/stable/26002102.

Ivanova, A. A., Sathe, A., Lipkin, B., Kumar, U., Radkani, S., Clark, T. H., Kauf, C., Hu, J., Pramod, R. T., Grand, G., Paulun, V., Ryskina, M., Akyürek, E., Wilcox, E., Rashid, N., Choshen, L., Levy, R., Fedorenko, E., Tenenbaum, J., and Andreas, J. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models, 2024. URL https://arxiv.org/abs/2405.09605.

Kahneman, D. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011. ISBN 9780374275631.

Kahneman, D. and Tversky, A. Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3 (3):430–454, 1972. doi: 10.1016/0010-0285(72)90016-3.

Kıcıman, E., Ness, R., Sharma, A., and Tan, C. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023. doi: 10.48550/arXiv.2305.00050. URL https://arxiv.org/abs/2305.00050. To appear in TMLR.

LeCun, Y. A path towards autonomous machine intelligence, June 2022. URL https://openreview.net/pdf?id=BZ5a1r-kVsf. Position paper.

Levison, M. and Lessard, G. Application of attribute grammars to natural language sentence generation. In Deransart, P. and Jourdan, M. (eds.), *Attribute Grammars and their Applications*, volume 461 of *Lecture Notes in Computer Science*, pp. 298–312. Springer, Berlin, Heidelberg, 1990. doi: 10.1007/3-540-53101-7_21. URL https://doi.org/10.1007/3-540-53101-7_21.

Margolis, H. *Patterns, Thinking, and Cognition: A Theory of Judgment*. University of Chicago Press, Chicago, IL, 1987.

Mollo, D. C. and Millière, R. The vector grounding problem. *arXiv preprint arXiv:2304.01481*, 2023. URL https://arxiv.org/abs/2304.01481. Available at https://doi.org/10.48550/arXiv.2304.01481.

Pearl, J. *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, 2nd edition, 2009.

Talmy, L. Narrative structure in a cognitive framework. In Duchan, J. F., Bruder, G. A., and Hewitt, L. E. (eds.), *Deixis in Narrative*, pp. 39–90. Psychology Press, London, 1995. ISBN 9780203052907. Chapter accessed via Taylor & Francis eBooks.

Tenenbaum, J. and Griffiths, T. Theory-based causal inference. In Becker, S., Thrun, S., and Obermayer, K. (eds.), *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002. URL https://proceedings.neurips.cc/paper_files/paper/2002/file/e77dbaf6759253c7c6d0efc5690369c7-Paper.pdf.

Tenenbaum, J. B. and Niyogi, S. Learning causal laws. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, pp. 1127–1132. Psychology Press, 2003. ISBN 9781315799360. doi: 10.4324/9781315799360-175.

Tversky, B. Distortions in cognitive maps. *Geoforum*, 23(2): 131–138, 1992. doi: 10.1016/0016-7185(92)90011-R.

Verheyen, S., Ameel, E., Rogers, T. T., and Storms, G. Learning a hierarchical organization of categories. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pp. 697–702. Cognitive Science Society, 2008.

Wason, P. C. Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20(3):273–281, 1968. doi: 10.1080/14640746808400161.

Xiong, S., Payani, A., Kompella, R., and Fekri, F. Large language models can learn temporal reasoning. *arXiv preprint arXiv:2401.06853*, 2024. URL https://doi.org/10.48550/arXiv.2401.06853. ACL 2024 (main).

Xu, K., Srivastava, A., Gutfreund, D., Sosa, F., Ullman, T., Tenenbaum, J., and Sutton, C. A bayesian-symbolic approach to reasoning and learning in intuitive physics. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 2478–2490. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/147540e129e096fa91700e9db6588354-Paper.pdf.

Appendix A: Category accuracy and human correlation for frontier open source and proprietary models

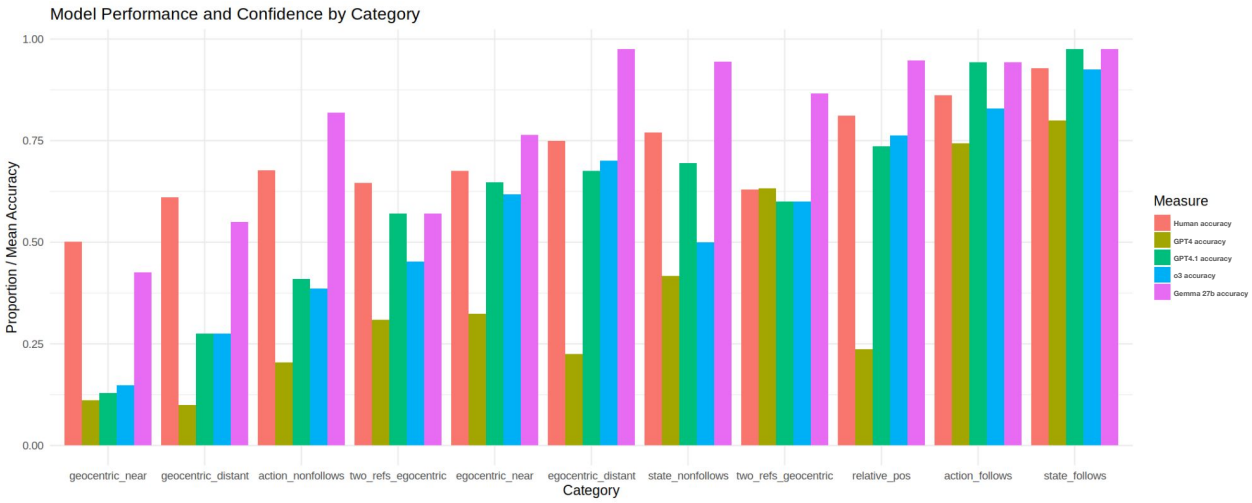

Figure 4. Mean category accuracy for humans, GPT4, GPT4.1, GPT-o3, and gemma-27b-it.

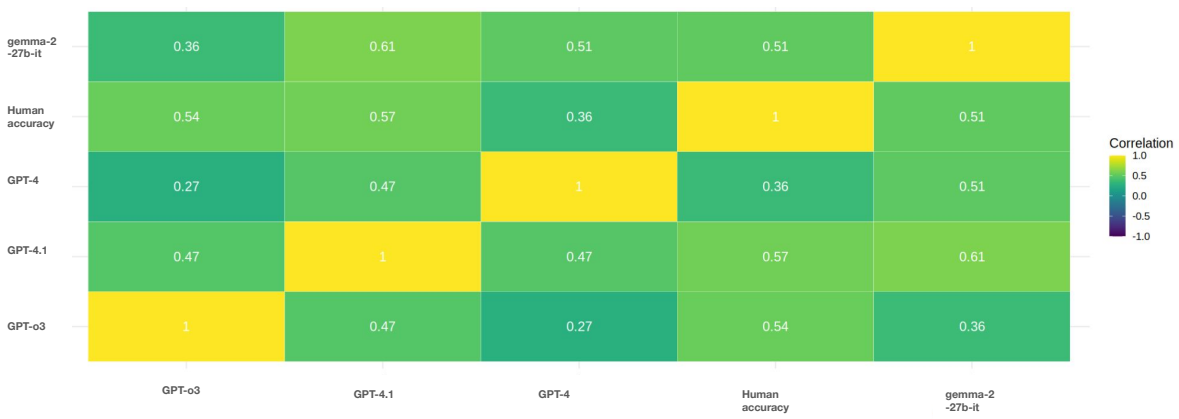

Figure 5. Mean correlations for humans and frontier models GPT4, GPT4.1, GPT-o3, and gemma-27b-it.

Appendix B: Measures of QK^T attention activations

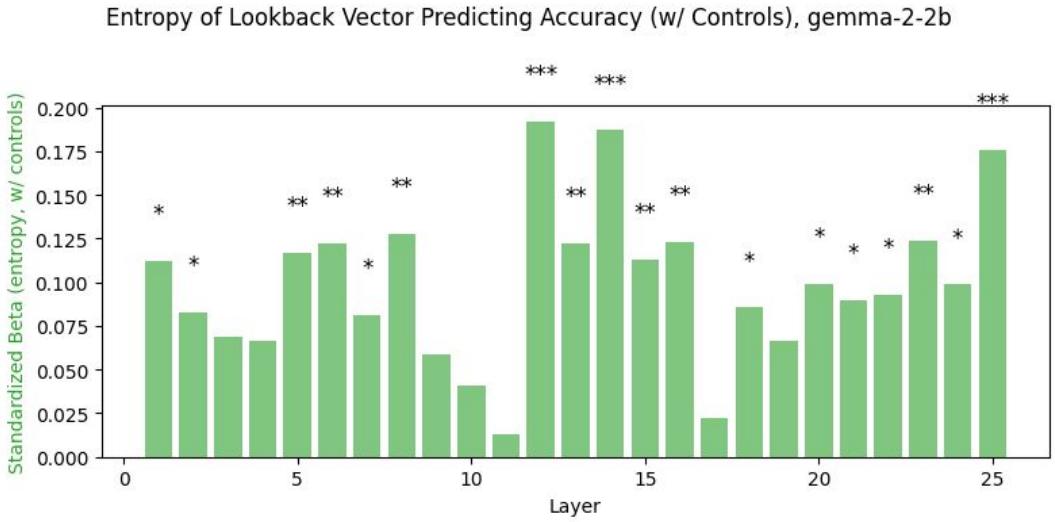

Figure 6. Layer coefficients for `gemma-2-2b` attention entropy, accounting for confounds of category, prompt type, prompt length, trigram frequency, and gemma-2-27b-it accuracy.

Appendix C: Prompt-Level Accuracy and Reaction Times

two refs egocentric

Top 5 Model Accuracy

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Ali is facing away from Mark. Ali turns around. Ali is facing BLANK Mark.	0.8500000	15.2500	1382.1000	4980.890
Ali and Mark are facing away from eachother. Mark turns around. Mark is facing BLANK Ali.	0.9000000	13.6250	1429.1714	4882.571
Mark is behind Ali. Mark walks fowards. Mark is getting BLANK Ali.	0.7500000	12.6875	737.6250	3962.600
Ali is facing towards Mark. Ali turns around. Ali is facing BLANK Mark.	0.7000000	10.6250	959.7857	6435.400
Mark is behind Ali. Ali turns around. Ali is facing BLANK Mark.	0.9090909	9.0625	3756.2455	3566.070

Bottom 5 Model Accuracy

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Mark is in front of Ali. Mark walks forwards. Mark is getting BLANK Ali.	0.5454545	-11.6875	1390.287	6064.650
Ali is to the right of Mark. Mark looks to his left. Mark BLANK Ali.	0.7500000	-11.2500	1802.050	5940.971
Mark is behind Ali. Mark walks backwards. Mark is getting BLANK Ali.	0.6190476	-11.1250	2189.044	6069.988
Mark is in front of Ali. Ali turns around. Ali is facing BLANK Mark.	0.8125000	-10.8750	1590.314	7391.629
Ali is to the right of Mark. Mark looks to his left and Ali looks to his right. Ali is facing BLANK Mark.	0.7619048	-10.5625	1423.250	11026.000

Top 5 Human Accuracy

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Mark is BLANK Ali. Ali turns around. Ali is facing towards Mark.	0.9333333	-3.6250	3966.308	4643.583
Mark is behind Ali. Ali turns around. Ali is facing BLANK Mark.	0.9090909	9.0625	3756.245	3566.070
Ali and Mark are facing away from eachother. Mark turns around. Mark is facing BLANK Ali.	0.9000000	13.6250	1429.171	4882.571
Ali is to the BLANK Mark. Mark looks to his left. Mark can't see Ali.	0.8750000	-2.7500	1765.787	4500.600
Ali and Mark are facing BLANK eachother. Mark turns around. Mark is facing towards Ali.	0.8666667	-5.3125	2095.308	4721.069

BOTTOM 5 HUMAN ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Mark and Ali are facing BLANK eachother. Mark turns around. Ali is facing away from Mark.	0.1904762	-9.7500	2336.800	684.000
Ali is to Mark's BLANK. Ali turns around. Ali is to Mark's right.	0.3333333	-4.2500	2557.933	3369.750
Mark is facing BLANK Ali. Ali turns around. Mark is facing away Ali.	0.3636364	-5.0625	5501.050	6989.775
Mark is to the BLANK Ali. Mark looks to his left and Ali looks to his left. Ali is facing towards Mark.	0.3636364	2.8750	12707.433	4695.800
Mark and Ali are both facing BLANK. Ali turns to his left. Ali is facing away from Mark.	0.4285714	2.2500	3148.533	3587.067

geocentric distant

TOP 5 MODEL ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Ali is South of Milwaukee. Ali walks North. Ali is getting BLANK Milwaukee.	0.5625000	13.3125	1428.7167	7480.012
Ali is East of Milwaukee. Ali walks West. Ali is getting BLANK Milwaukee.	0.6190476	12.0625	1518.2750	10317.089
Ali is South of Milwaukee. Ali walks South. Ali is getting BLANK Milwaukee.	0.8181818	7.1250	836.8333	4574.163
Ali is directly North of Milwaukee. Ali walks East. Ali is BLANK Milwaukee.	0.7000000	6.4375	2737.3000	8906.057
Ali is directly BLANK of Milwaukee. Ali walks West. Ali is Southwest of Milwaukee.	0.7142857	6.3750	8512.5500	4734.450

BOTTOM 5 MODEL ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Milwaukee is North of Ali. Ali turns around. Milwaukee is BLANK Ali.	0.5000000	-9.6250	1579.086	4271.788
Milwaukee is West of Ali. Ali turns around. Milwaukee is BLANK Ali.	0.6000000	-9.1875	1200.143	6628.400
Milwaukee is South of Ali. Ali turns around. Milwaukee is BLANK Ali.	0.6363636	-8.9375	2979.300	3654.656
Milwaukee is North of Ali. Ali turns right. Milwaukee is BLANK Ali.	0.6250000	-8.6875	5267.150	7893.200
Milwaukee is North of Ali. Ali turns left. Milwaukee is BLANK Ali.	0.6363636	-8.3125	1220.533	7596.955

TOP 5 HUMAN ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Ali is directly South of Milwaukee. Ali walks West. Ali is BLANK Milwaukee.	0.9047619	4.000	4118.5667	7185.130
Milwaukee is BLANK of Ali. Ali walks East. Ali is getting further from Milwaukee.	0.9047619	2.750	1170.0300	4526.408
Milwaukee is West of Ali. Ali walks East. Ali is getting BLANK Milwaukee.	0.8750000	-1.125	1931.2750	4235.273
Ali is directly BLANK Milwaukee. Ali walks East. Ali is Northeast of Milwaukee.	0.8666667	-0.625	3392.8250	4902.918
Ali is South of Milwaukee. Ali walks South. Ali is getting BLANK Milwaukee.	0.8181818	7.125	836.8333	4574.163

BOTTOM 5 HUMAN ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Milwaukee is BLANK Ali. Ali turns around. Milwaukee is East of Ali.	0.1818182	-5.3750	1705.50	9651.000
Milwaukee is BLANK Ali. Ali turns around. Milwaukee is West of Ali.	0.2666667	-5.4375	4730.10	3286.167
Milwaukee is BLANK Ali. Ali turns around. Milwaukee is South of Ali.	0.3125000	-4.2500	2266.05	3013.125
Milwaukee is East of Ali. Ali turns around. Milwaukee is BLANK Ali.	0.3333333	-8.1875	1835.20	7535.933
Milwaukee is BLANK Ali. Ali turns left. Milwaukee is West of Ali.	0.3750000	-0.8750	2540.64	1897.880

egocentric near

TOP 5 MODEL ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
The painting is in front of Ali. Ali turns around. The painting is BLANK Ali.	0.9090909	12.6250	929.250	3426.709
Ali is in front of the painting. Ali turns around. The painting is BLANK Ali.	0.8000000	11.8750	1103.300	5311.056
The painting is left of Ali. Ali turns around. The painting is BLANK Ali.	0.5000000	9.7500	4717.967	7656.078
The painting is behind Ali. Ali turns around. The painting is BLANK Ali.	0.9523810	9.3125	1296.723	4801.236
The painting is left of Ali. Ali turns right. The painting is BLANK Ali.	0.8500000	8.3750	2155.415	4621.415

BOTTOM 5 MODEL ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Ali is behind the painting. Ali turns around. Ali is BLANK the painting.	0.2500000	-11.6250	1946.400	3389.400
The painting is to Ali's right. The painting is moved right. The painting is to Ali's BLANK.	0.7500000	-7.3125	1859.010	6901.017
The painting is in front of Ali. Ali turns right. The painting is BLANK Ali.	0.7619048	-7.1250	1971.264	6817.136
Ali is right of the painting. Ali turns around. Ali is BLANK the painting.	0.5714286	-6.6875	4473.171	14483.057
Ali is BLANK the painting. Ali turns around. Ali is right of the painting.	0.3809524	-5.1250	1271.875	1723.100

TOP 5 HUMAN ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
The painting is behind Ali. Ali turns around. The painting is BLANK Ali.	0.9523810	9.3125	1296.723	4801.236
The painting is in front of Ali. Ali turns around. The painting is BLANK Ali.	0.9090909	12.6250	929.250	3426.709
The painting is behind Ali. Ali turns right. The painting is BLANK Ali.	0.9000000	4.6875	1992.911	4921.360
The painting is in front of Ali. Ali turns left. The painting is BLANK Ali.	0.9000000	-4.0000	2304.170	6582.055
The painting is left of Ali. Ali turns left. The painting is BLANK Ali.	0.8750000	2.5000	2052.170	5217.164

BOTTOM 5 HUMAN ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
The painting is to Ali's BLANK. The painting is moved left. The painting is to Ali's right.	0.2000000	1.250	1737.900	8295.75
Ali is behind the painting. Ali turns around. Ali is BLANK the painting.	0.2500000	-11.625	1946.400	3389.40
The painting is to Ali's left. The painting is moved left. The painting is to Ali's BLANK.	0.3000000	0.375	1631.340	3009.14
Ali is BLANK painting. Ali turns around. Ali is behind the painting.	0.3333333	2.375	2510.500	3559.46
Ali is BLANK the painting. Ali turns around. Ali is right of the painting.	0.3809524	-5.125	1271.875	1723.10

action nonfollows

TOP 5 MODEL ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
The box is floating in the water. Ali puts a BLANK in the box. The box floats.	0.9545455	11.5000	3143.940	4244.280
The box is in the bin. Ali moves the BLANK. The bin doesn't move.	0.8750000	9.6250	4296.764	3886.518
The ball is BLANK the box. The ball moves right. The ball is higher than the box.	0.7142857	9.0000	6100.150	4487.725
The box is above the tunnel. Ali goes under the tunnel. The box is BLANK the tunnel.	0.8500000	8.5625	1384.122	8710.222
The book is under the table. Ali moves the book. The table BLANK.	0.6250000	8.4375	3203.012	6137.833

BOTTOM 5 MODEL ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
The wine glass is on the table. Ali bumps the lamp. The wine BLANK.	0.6500000	-17.000	1675.2000	6115.489
The glass is on the table and the cup is on the floor. Ali picks up the glass. The cup BLANK.	0.3000000	-16.125	529.3333	541.450
Ali is in the wagon. Ali pushes on the wagon. The wagon BLANK.	0.2500000	-15.625	3354.3500	3330.100
Ali is in the wagon. Ali pulls on the wagon. The wagon BLANK.	0.6000000	-14.375	3307.6429	7324.229
The wine glass is on the table. Ali bumps the BLANK. The wine doesn't spill.	0.5238095	-12.500	1401.2800	3263.533

TOP 5 HUMAN ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
The ball is above the box. The ball moves right. The ball is BLANK the box.	1.0000000	3.0625	3477.091	5971.400
The box is floating in the water. Ali puts a BLANK in the box. The box floats.	0.9545455	11.5000	3143.940	4244.280
The ball is inside the box. Ali pushes the box forward. The ball is BLANK the box.	0.9375000	-7.1250	1958.880	6148.483
The ball is under the box. The ball moves right. The ball is BLANK the box.	0.9375000	-0.5625	8267.540	5690.542
The box is BLANK the tunnel. Ali goes under the tunnel. The box is above the tunnel.	0.9375000	4.3750	6048.960	4148.520

BOTTOM 5 HUMAN ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
The bowling ball is close to Ali. Ali throws the bowling ball. The bowling ball is BLANK Ali.	0.0952381	-6.4375	2543.0000	6984.550
Ali is in the wagon. Ali pushes on the wagon. The wagon BLANK.	0.2500000	-15.6250	3354.3500	3330.100
The ball is close to Ali. Ali throws the BLANK. The ball is close to Ali.	0.2666667	-12.0625	3501.8571	4984.033
The glass is on the table and the cup is on the floor. Ali picks up the glass. The cup BLANK.	0.3000000	-16.1250	529.3333	541.450
Ali is behind the statue. Ali turns around. Ali is BLANK.	0.3125000	-10.9375	2992.5667	5471.950

action follows

TOP 5 MODEL ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
The ball is on the table. Ali tilts the table. The ball BLANK.	0.9523810	18.875	1732.017	3549.046
The box is floating in the water. Ali puts a bowling ball in the box. The box BLANK.	1.0000000	18.750	1661.558	4645.192
The wagon is on top of the hill. Ali pushes the wagon. The wagon BLANK.	0.9500000	16.750	1346.964	3668.933
The book is on the shelf. Ali bumps the shelf. The book BLANK the shelf.	0.8571429	16.500	1502.200	4328.255
Ali is behind the wagon. Ali pushes on the wagon. The wagon BLANK.	1.0000000	16.250	1377.625	4051.331

BOTTOM 5 MODEL ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
The wine glass is on the BLANK. Ali bumps the wine glass. The wine spills on the floor.	0.2666667	-11.4375	4214.900	5121.275
The ball is in the box. Ali moves the box. The ball BLANK.	0.9523810	-6.5000	2054.250	2472.464
The wine glass is on the table. The wine glass falls over. The wine spills on the BLANK.	0.4545455	-6.5000	2954.956	3674.290
Ali is BLANK the wagon. Ali pulls on the wagon. The wagon moves.	0.7142857	-3.2500	2633.475	3625.511
The ping pong ball is on the floor. Ali rolls a tennis ball BLANK the ping pong ball. The ping pong ball moves.	0.8750000	-1.4375	4531.910	4972.089

TOP 5 HUMAN ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Ali is behind the wagon. Ali pushes on the wagon. The wagon BLANK.	1	16.2500	1377.625	4051.331
The ball is on top of the box. Ali pushes the box. The ball BLANK.	1	10.4375	1445.155	3956.923
The box is floating in the water. Ali puts a bowling ball in the box. The box BLANK.	1	18.7500	1661.558	4645.192
The tennis ball is next to the bowling ball. Ali throws the tennis ball. The tennis ball is BLANK the bowling ball.	1	3.0000	2159.600	5780.386
The ball is in the box. Ali moves the BLANK. The ball moves.	1	10.1250	3624.236	3966.893

BOTTOM 5 HUMAN ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
The wine glass is on the BLANK. Ali bumps the wine glass. The wine spills on the floor.	0.2666667	-11.4375	4214.900	5121.275
The wine glass is on the table. The wine glass falls over. The wine spills on the BLANK.	0.4545455	-6.5000	2954.956	3674.290
The seesaw is balanced. Ali tilts the BLANK side of the seesaw. The left side goes up.	0.6190476	7.4375	1229.667	5361.017
Ali is BLANK the wagon. Ali pulls on the wagon. The wagon moves.	0.7142857	-3.2500	2633.475	3625.511
The book is on the table. Ali moves the table. The book BLANK.	0.7272727	8.6875	2142.057	3911.137

egocentric distant

TOP 5 MODEL ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Ali is facing away from Milwaukee. Ali turns 180 degrees. Ali is facing BLANK Milwaukee.	0.7142857	16.8750	1225.411	6244.925
Ali is facing away from Milwaukee. Ali turns around. Ali is facing BLANK Milwaukee.	1.0000000	16.7500	1217.460	3805.780
Ali is facing towards Milwaukee. Ali turns around. Ali is facing BLANK Milwaukee.	0.9375000	10.7500	1288.044	6259.240
Chicago is in front of Ali. Ali turns around. Chicago is BLANK Ali.	0.8750000	10.1875	1274.218	4958.550
Ali is facing away from Chicago. Ali turns around. Chicago is BLANK Ali.	0.6363636	8.5625	1399.200	3704.910

BOTTOM 5 MODEL ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Ali is facing BLANK Chicago. Ali turns around. Chicago is in front of Ali.	0.8666667	-4.4375	2564.983	6606.575
Chicago is BLANK Ali. Ali turns left. Chicago is right of Ali.	0.5625000	-3.6250	5018.243	3499.814
Ali is facing BLANK Chicago. Ali turns left. Chicago is left of Ali.	0.5000000	-3.1250	5681.267	6393.750
Ali is facing away from Chicago. Ali turns left. Chicago is BLANK Ali.	0.6000000	-2.1250	4086.390	4871.189
Ali is facing BLANK Chicago. Ali turns right. Chicago is left of Ali.	0.7333333	-1.7500	6373.540	4951.920

TOP 5 HUMAN ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Ali is facing away from Milwaukee. Ali turns around. Ali is facing BLANK Milwaukee.	1.0000000	16.7500	1217.460	3805.780
Ali is facing towards Milwaukee. Ali turns around. Ali is facing BLANK Milwaukee.	0.9375000	10.7500	1288.044	6259.240
Chicago is to Ali's left. Ali turns right. Ali is facing BLANK Chicago.	0.9375000	4.8125	2909.380	6113.936
Chicago is BLANK Ali. Ali turns around. Chicago is in front of Ali.	0.9333333	5.3125	3117.771	4751.854
Chicago is behind Ali. Ali turns around. Chicago is BLANK Ali.	0.9090909	7.1875	1420.427	3353.308

BOTTOM 5 HUMAN ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Ali is facing BLANK Chicago. Ali turns left. Chicago is left of Ali.	0.5000000	-3.1250	5681.267	6393.750
Chicago is BLANK Ali. Ali turns left. Chicago is left of Ali.	0.5238095	4.8750	3284.314	10713.000
Ali is facing away from Chicago. Ali turns right. Chicago is BLANK Ali.	0.5625000	3.1875	2399.343	3468.250
Chicago is BLANK Ali. Ali turns left. Chicago is right of Ali.	0.5625000	-3.6250	5018.243	3499.814
Ali is facing BLANK Chicago. Ali turns left. Chicago is right of Ali.	0.5714286	1.2500	1436.860	6938.340

state nonfollows

TOP 5 MODEL ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
The paper is in one piece. Ali BLANK the paper. The paper is in one piece.	0.9545455	16.0000	2988.367	3196.773
The full glass is on the table. Ali knocks the glass over. The table is BLANK.	0.9500000	14.3125	1458.350	4915.175
The ice cube is on the floor. Ali moves the ice cube to the counter. The ice cube gets BLANK.	0.8095238	13.5000	3375.136	6840.482
The glass is on the floor. Ali bumps the BLANK. The glass doesn't break.	0.4000000	9.8125	3493.180	2453.520
The BLANK glass is on the table. Ali knocks the glass over. The table is wet.	0.9333333	9.4375	2198.462	3705.215

BOTTOM 5 MODEL ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
The glass is on the floor. Ali bumps the table. The glass BLANK.	0.6363636	-17.125	2694.833	3767.855
The wine glass is on the floor. Ali touches the glass. The glass BLANK.	0.9375000	-16.625	3059.170	4372.120
The soup is on the table. Ali turns on the stove. The soup gets BLANK.	0.3125000	-16.250	1513.517	4282.686
The brick is on the table. Ali bumps the brick off the table. The brick BLANK.	0.5500000	-14.875	3364.625	3802.150
The empty glass is on the table. Ali knocks the glass over. The table is BLANK.	0.6190476	-14.250	2673.188	5764.233

TOP 5 HUMAN ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
The cup is empty. Ali pours sand in the cup. The cup is BLANK.	1.0000000	2.59375	3262.567	3818.525
The paper is in one piece. Ali BLANK the paper. The paper is in one piece.	0.9545455	16.00000	2988.367	3196.773
Ali is holding the book. Ali drops the book. The book BLANK.	0.9523810	-13.43750	1962.033	3254.892
The icecube is in the freezer. Ali puts the icecube in the snow. The icecube BLANK.	0.9523810	-7.62500	1610.555	4657.278
The soup is on the BLANK. Ali turns on the stove. The soup gets colder.	0.9523810	7.50000	4015.418	5338.190

BOTTOM 5 HUMAN ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
The soup is on the table. Ali turns on the stove. The soup gets BLANK.	0.3125000	-16.2500	1513.517	4282.686
The glass is on the floor. Ali bumps the BLANK. The glass doesn't break.	0.4000000	9.8125	3493.180	2453.520
The glass and the cup are on the table. Ali pushes the cup off the table. The glass BLANK.	0.4500000	7.7500	1444.225	4619.188
The window is behind Ali. Ali throws the baseball forward. The window BLANK.	0.4545455	-5.1250	863.900	4951.371
The glass and the cup are on the table. Ali pushes the BLANK off the table. The glass doesn't break.	0.5238095	3.6875	2181.033	4168.933

relative pos

TOP 5 MODEL ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
The ball is inside the box. Ali picks up the ball. The ball is BLANK the box.	0.6363636	14.000	1773.571	4069.400
The box is in front of the ball. Ali pushes the box BLANK the ball. The ball is not touching the box.	0.9047619	13.000	2017.490	5205.036
The glass is on the floor. Ali sets the glass on the BLANK. The glass is above the table.	0.9545455	12.125	5389.100	4850.736
The box is in front of the ball. Ali pushes the box away from the ball. The ball is BLANK the box.	0.8500000	11.875	1585.900	7536.482
The box is next to the desk. Ali stacks the box on the desk. The box is BLANK the desk.	1.0000000	11.750	3212.043	6862.100

BOTTOM 5 MODEL ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Ali is holding the tennis ball and the ping pong ball is on the floor. Ali BLANK the ping pong ball. The ping pong ball is closer to the tennis ball.	0.2000	-10.4375	6241.700	5296.833
The tennis ball is BLANK the bowling ball. Ali rolls the bowling ball. The bowling ball is further from the tennis ball.	0.4375	-9.0625	5629.300	4875.617
The ball and the box are on the table. The BLANK falls off the table. The box is lower than the ball.	0.8125	-7.3750	3845.687	4980.044
The box is next to the ball. Ali puts the ball inside the box. The ball is BLANK the box.	0.6875	-6.8125	3339.515	5150.000
The ball and the box are on the floor. Ali picks up the BLANK. The ball is higher than the box.	1.0000	-5.1250	3100.200	7249.460

TOP 5 HUMAN ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
The box is in front of the ball. Ali pushes the box into the ball. The ball is BLANK the box.	1	7.8125	2672.108	5027.685
The box is next to the desk. Ali stacks the box on the desk. The box is BLANK the desk.	1	11.7500	3212.043	6862.100
The glass is on the table. Ali bumps the glass off the table. The glass is BLANK the table.	1	10.6250	2893.482	4011.654
The ball and the box are on the floor. Ali picks up the BLANK. The ball is higher than the box.	1	-5.1250	3100.200	7249.460
The ball is next to the box. Ali puts the BLANK. The ball is smaller than the box.	1	4.7500	4580.800	10276.233

BOTTOM 5 HUMAN ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Ali is holding the tennis ball and the ping pong ball is on the floor. Ali BLANK the ping pong ball. The ping pong ball is closer to the tennis ball.	0.2000000	-10.4375	6241.700	5296.833
The tennis ball is BLANK the bowling ball. Ali rolls the bowling ball. The bowling ball is further from the tennis ball.	0.4375000	-9.0625	5629.300	4875.617
The box is next to the ball. Ali puts the BLANK. the ball is larger than the box.	0.4761905	-1.5000	3824.517	4695.683
The box is next to the desk. Ali stacks the BLANK. The box is below the desk.	0.5714286	3.2500	3256.371	5674.887
The box is next to the ball. Ali puts the box inside the ball. the ball is BLANK the box.	0.6000000	9.0625	3518.875	7945.186

state follows

TOP 5 MODEL ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
The paper is in two pieces. Ali BLANK the pieces. The paper is in one piece.	1.0000000	20.000	2492.393	4032.847
The room is dark. Ali turns on the light. The room gets BLANK.	0.9090909	19.250	1374.845	2736.892
The soup is on the stove. Ali turns on the stove. The soup gets BLANK.	0.8500000	17.125	1054.522	3576.118
The glass is on the table. Ali bumps the table. The glass BLANK.	0.7500000	17.000	1923.000	3601.210
The glass is on the table. Ali bumps the glass off the table. The glass BLANK.	1.0000000	16.750	2053.340	3575.110

BOTTOM 5 MODEL ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Ali is standing on carpet. Ali drops the glass. The glass BLANK.	0.8181818	-14.8750	1258.500	3485.278
The room is bright. Ali turns BLANK a light. The room gets brighter.	0.9523810	-2.1250	1019.390	2898.067
The room is bright. Ali BLANK the lights. The room is dark.	0.8666667	0.0000	2449.557	3447.738
The soup is on the BLANK. Ali turns on the stove. The soup gets warmer.	1.0000000	3.6875	2596.700	2943.580
Ali is standing on BLANK. Ali drops the glass. The glass breaks.	1.0000000	5.3125	3538.658	4456.550

TOP 5 HUMAN ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
The apple is in the bucket. Ali puts the apple in the refrigerator. The apple gets BLANK.	1	13.3125	1265.342	3741.707
The cup is empty. Ali pours water in the cup. The cup is BLANK.	1	13.5625	1570.590	2891.833
The glass is on the table. Ali bumps the glass off the table. The glass BLANK.	1	16.7500	2053.340	3575.110
The glass is on the table. Ali overfills the glass. The table is BLANK.	1	16.5000	1538.864	4169.207
The ice cream is on the table. Ali puts the ice cream in the freezer. The ice cream gets BLANK.	1	13.1875	1826.220	4351.436

BOTTOM 5 HUMAN ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
The glass is on the table. Ali bumps the table. The glass BLANK.	0.7500000	17.0000	1923.000	3601.210
The wall was painted BLANK ago. Ali touches the wall. Ali's hand is painted.	0.7500000	10.5000	2031.650	5129.637
The icecube is in the freezer. Ali puts the icecube in the BLANK. The icecube melts.	0.8095238	9.5000	1422.760	3187.573
Ali is standing on carpet. Ali drops the glass. The glass BLANK.	0.8181818	-14.8750	1258.500	3485.278
The icecube is in the freezer. Ali puts the icecube in the frying pan. The icecube BLANK.	0.8181818	13.5625	2108.625	4476.500

two refs geocentric

TOP 5 MODEL ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Mark is East of Budapest and Ali is West of Budapest. Ali walks directly BLANK Mark. Ali is heading East.	0.8750000	16.2500	2336.260	8195.336
Mark is East of Budapest and Ali is West of Budapest. Mark walks directly towards Ali. Mark is heading BLANK.	0.8571429	12.4375	1093.120	9971.664
Mark is East of Budapest and Ali is West of Budapest. Ali walks directly towards Mark. Ali is heading BLANK.	0.6500000	11.0625	1093.200	12220.812
Mark is facing East and Ali is facing West. Mark turns around. Mark is facing BLANK.	0.8000000	10.5625	3198.675	5910.137
Mark is facing BLANK and Ali is facing West. Mark turns around. Mark is facing West.	0.6666667	9.2500	2477.082	8286.470

BOTTOM 5 MODEL ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Mark is facing North and Ali is facing East. Mark turns 90 degrees to his right. Mark and Ali are facing BLANK.	0.6000000	-14.0625	1361.270	6746.560
Mark is facing South and Ali is facing East. Mark turns 90 degrees to his right. Mark and Ali are facing BLANK.	0.4761905	-13.7500	2567.733	11900.543
Mark is East of Ali facing East. Ali is West of Mark facing West. Mark is facing BLANK Ali.	0.7500000	-5.3750	2951.450	18572.371
Mark is BLANK of Ali. Mark turns right. Mark is East of Ali.	0.4000000	-5.0000	3194.460	9961.800
Mark is East of Ali. Mark turns right. Mark is BLANK of Ali.	0.9090909	-4.3750	2539.691	7708.164

TOP 5 HUMAN ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Mark is East of Ali. Mark turns right. Mark is BLANK of Ali.	0.9090909	-4.3750	2539.691	7708.164
Mark and Ali are both directly East of Budapest. Mark walks BLANK. Ali is South of Mark.	0.9090909	2.1250	6094.492	5593.955
Mark is East of Budapest and Ali is West of Budapest. Ali walks directly BLANK Mark. Ali is heading East.	0.8750000	16.2500	2336.260	8195.336
Mark and Ali are both directly West of Budapest. Mark walks BLANK. Ali is North of Mark.	0.8666667	4.6250	2403.367	9237.554
Mark is East of Budapest and Ali is West of Budapest. Mark walks directly towards Ali. Mark is heading BLANK.	0.8571429	12.4375	1093.120	9971.664

BOTTOM 5 HUMAN ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Milwaukee is South of Mark. Mark goes BLANK. Mark is Northeast of Milwaukee.	0.2727273	-2.6250	3848.825	1296.150
Milwaukee is North of Mark. Mark goes BLANK. Milwaukee is Northwest of Mark.	0.2857143	-2.3750	1845.460	3337.180
Milwaukee is South of Mark. Mark goes West. Mark is BLANK of Milwaukee.	0.3125000	0.0625	2402.525	7099.660
Milwaukee is South of Mark. Mark goes BLANK. Milwaukee is Southeast of Mark.	0.3125000	0.7500	7845.400	1762.525
Mark is BLANK of Ali. Mark turns right. Mark is East of Ali.	0.4000000	-5.0000	3194.460	9961.800

geocentric near

TOP 5 MODEL ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Mark is facing West. Mark turns 90 degrees to his right. Mark is facing BLANK.	0.8000000	9.6875	905.3125	8485.571
Mark is facing North. Mark turns 90 degrees to his right. Mark is facing BLANK.	0.6250000	9.1875	1664.2667	4931.650
Mark is facing BLANK. Mark turns 90 degrees to his right. Mark is facing East.	0.8636364	5.5625	3555.9800	5212.255
Mark is facing BLANK. Mark turns 90 degrees to his right. Mark is facing North.	0.7500000	3.7500	3973.2571	6091.729
Mark is facing BLANK. Mark turns 90 degrees to his right. Mark is facing South.	0.8000000	2.2500	5059.5923	8019.375

BOTTOM 5 MODEL ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
The box is West of Ali. Ali turns around. The box is BLANK Ali.	0.4761905	-10.8750	1869.775	5667.200
The painting is North of Ali. Ali turns right. The painting is BLANK Ali.	0.3750000	-10.8125	680.260	4979.017
The box is North of Ali. Ali turns around. The box is BLANK Ali.	0.6500000	-10.7500	1548.700	6636.337
The box is South of Ali. Ali turns around. The box is BLANK Ali.	0.3750000	-10.2500	5070.540	3533.025
The box is East of Ali. Ali turns around. The box is BLANK Ali.	0.3636364	-10.0000	1982.229	13070.675

TOP 5 HUMAN ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
Mark is facing BLANK. Mark turns 90 degrees to his right. Mark is facing East.	0.8636364	5.5625	3555.9800	5212.255
Mark is facing West. Mark turns 90 degrees to his right. Mark is facing BLANK.	0.8000000	9.6875	905.3125	8485.571
Mark is facing BLANK. Mark turns 90 degrees to his right. Mark is facing South.	0.8000000	2.2500	5059.5923	8019.375
Mark is facing BLANK. Mark turns 90 degrees to his right. Mark is facing North.	0.7500000	3.7500	3973.2571	6091.729
The statue is BLANK Ali. Ali walks East. The statue is North of Ali.	0.7333333	-6.8125	5661.8462	7832.646

BOTTOM 5 HUMAN ACCURACY

Prompt	Human Acc.	gemma-27b-it logit diff.	Choice RT	Space RT
The box is BLANK Ali. Ali turns around. The box is East of Ali.	0.1818182	-3.3750	1084.05	5355.900
The painting is BLANK Ali. Ali turns left. The painting is West of Ali.	0.1875000	0.3750	9288.80	1754.200
The painting is BLANK Ali. Ali turns left. The painting is South of Ali.	0.2666667	-3.6250	4177.08	3346.760
The painting is BLANK Ali. Ali turns right. The painting is South of Ali.	0.2666667	-5.6250	2243.60	10391.350
The painting is East of Ali. Ali turns around. The painting is BLANK Ali.	0.3125000	-8.6875	1698.55	5620.533